PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

**Cite this article:** Polonsky JA *et al*. 2019 Outbreak analytics: a developing data science for informing the response to emerging pathogens. *Phil. Trans. R. Soc. B* **374**: 20180276.
http://dx.doi.org/10.1098/rstb.2018.0276

Accepted: 4 December 2018

One contribution of 16 to a theme issue 'Modelling infectious disease outbreaks in humans, animals and plants: epidemic forecasting and control'.

**Subject Areas:**
health and disease and epidemiology

**Keywords:**
epidemics, infectious, methods, tools, pipeline, software

**Author for correspondence:**
Thibaut Jombart
e-mail: thibautjombart@gmail.com

†These authors contributed equally to the study.

THE ROYAL SOCIETY PUBLISHING

# Outbreak analytics: a developing data science for informing the response to emerging pathogens

Jonathan A. Polonsky[1,3,†], Amrish Baidjoe[4,†], Zhian N. Kamvar[4], Anne Cori[4], Kara Durski[2], W. John Edmunds[5,6], Rosalind M. Eggo[5,6], Sebastian Funk[5,6], Laurent Kaiser[3], Patrick Keating[5,8], Olivier le Polain de Waroux[5,8,9], Michael Marks[7], Paula Moraga[10], Oliver Morgan[1], Pierre Nouvellet[4,11], Ruwan Ratnayake[5,6], Chrissy H. Roberts[7], Jimmy Whitworth[5,8] and Thibaut Jombart[4,5,8]

[1]Department of Health Emergency Information and Risk Assessment, and [2]Department of Infectious Hazard Management, World Health Organization, Avenue Appia 20, 1211 Geneva, Switzerland
[3]Faculty of Medicine, University of Geneva, 1 rue Michel-Servet, 1211 Geneva, Switzerland
[4]Department of Infectious Disease Epidemiology, School of Public Health, MRC Centre for Global Infectious Disease Analysis, Imperial College London, Medical School Building, St Mary's Campus, Norfolk Place London W2 1PG, UK
[5]Department of Infectious Disease Epidemiology, [6]Centre for Mathematical Modelling of Infectious Diseases, and [7]Clinical Research Department, Faculty of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, UK
[8]UK Public Health Rapid Support Team, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, UK
[9]Public Health England, Wellington House, 133–155 Waterloo Road, London SE1 8UG, UK
[10]Centre for Health Informatics, Computing and Statistics (CHICAS), Lancaster Medical School, Lancaster University, Lancaster LA1 4YW, UK
[11]School of Life Sciences, University of Sussex, Sussex House, Brighton BN1 9RH, UK

JAP, 0000-0002-8634-4255; AB, 0000-0001-5295-5085; ZNK, 0000-0003-1458-7108; AC, 0000-0002-8443-9162; SF, 0000-0002-2842-3406; OM, 0000-0002-9543-3778; PN, 0000-0002-6094-5722; TJ, 0000-0003-2226-8692

Despite continued efforts to improve health systems worldwide, emerging pathogen epidemics remain a major public health concern. Effective response to such outbreaks relies on timely intervention, ideally informed by all available sources of data. The collection, visualization and analysis of outbreak data are becoming increasingly complex, owing to the diversity in types of data, questions and available methods to address them. Recent advances have led to the rise of *outbreak analytics*, an emerging data science focused on the technological and methodological aspects of the outbreak data pipeline, from collection to analysis, modelling and reporting to inform outbreak response. In this article, we assess the current state of the field. After laying out the context of outbreak response, we critically review the most common analytics components, their inter-dependencies, data requirements and the type of information they can provide to inform operations in real time. We discuss some challenges and opportunities and conclude on the potential role of outbreak analytics for improving our understanding of, and response to outbreaks of emerging pathogens.

This article is part of the theme issue 'Modelling infectious disease outbreaks in humans, animals and plants: epidemic forecasting and control'. This theme issue is linked with the earlier issue 'Modelling infectious disease outbreaks in humans, animals and plants: approaches and important themes'.

## 1. Introduction

Emerging infectious diseases are a constant threat to public health worldwide. In the past decade, several major outbreaks, such as the 2009 influenza pandemic [1],

© 2019 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

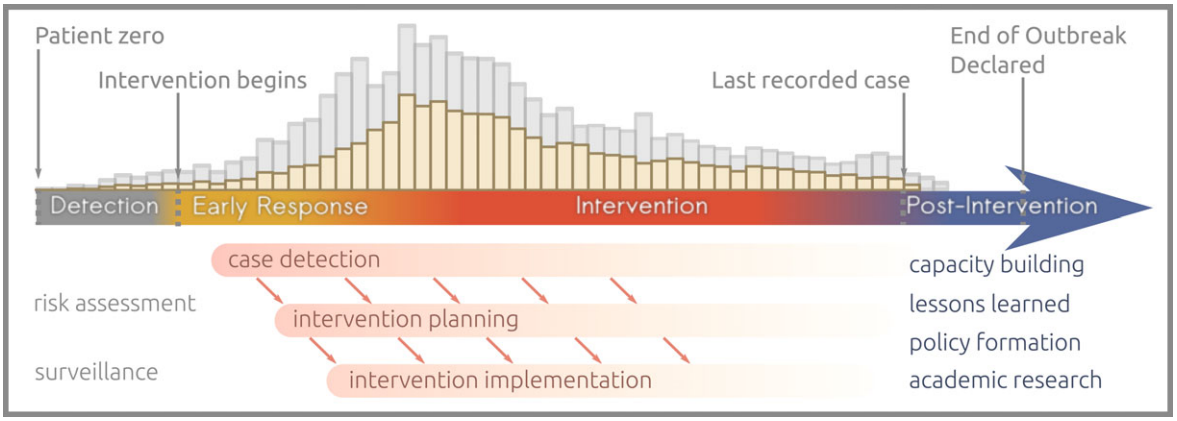



**Figure 1.** Successive phases of an outbreak response. The histogram along the top represents reported (yellow) and unreported (grey) incidence.

the Middle-East Respiratory Syndrome coronavirus (MERS-CoV) [2–4], the emergence of Zika [5,6] and the West African Ebola virus disease (EVD) outbreak [7,8], have been potent reminders of the need for robust surveillance systems and timely responses to nascent epidemics [9]. The West African EVD outbreak, by far the largest such epidemic in recorded history, in particular, had a strong impact on global health security and public health policy and practice [7,8,10]. It highlighted the difficulties of maintaining situational awareness in the absence of standards for surveillance, data collection and analysis, as well as the challenges of mounting and sustaining a large-scale international response [7,8,11,12]. Despite the lessons learnt [9,13,14], the recent (2018) EVD outbreaks in Democratic Republic of the Congo [15,16] are a stark reminder that a large number of these challenges remain.

An important feature of the modern response to epidemics is the increasing focus on exploiting all available data to inform the response in real time and allow evidence-based decision making [3,4,7,8,13,17]. Using data for improving situational awareness is complex, involving a range of inter-connected tasks and skills from point-of-care data collection to the generation of informative situational reports (sitreps). The science underpinning these data pipelines involves a wide range of approaches, including database design and mobile technology [18,19], frequentist statistics and maximum-likelihood estimation [7], interactive data visualization [20,21], geostatistics [22–24], graph theory [20,25,26], Bayesian statistics [8,27,28], mathematical modelling [29–31], genetic analyses [32–36] and evidence synthesis approaches [37]. This accretion of heterogeneous disciplines, which may be best summarized as 'outbreak analytics', forms an emerging domain of data science: an 'interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms' [38], dedicated to informing outbreak response. *Outbreak analytics* sits at the crossroads of public health planning, field epidemiology, methodological development and information technologies, opening up exciting opportunities for specialists in these fields to work together to meet the needs for an epidemic response.

In this article, we outline this developing research field and review the current state of outbreak analytics. In particular, we focus on how different analysis components interact within functional workflows, and how each component can be used to inform different stages of an outbreak response. We discuss key challenges and opportunities associated with the deployment of efficient, reliable and informative data analysis pipelines and their potential impact.

## 2. The outbreak response context

### (a) The different phases of an outbreak response

The focus of the public health response shifts during the course of an epidemic or outbreak, and so do the analytics. We identify four main stages (figure 1). The *detection* stage starts with the first case and ends with the first intervention activities (e.g. patient isolation, contact tracing, vaccination) and involves surveillance systems and mostly qualitative risk assessments. Next, the *early response* is the initial part of the intervention during which the first simple analytics can take place, essentially centred around estimating transmissibility. This blends into the *intervention* stage, where more complex analytics may be involved to inform planning (e.g. vaccination strategies), which ends once the last reported case has recovered or died. The *post-intervention* stage is for lessons to be learned, for improving preparedness for the next epidemic and for training and capacity building [39].

### (b) Questions during and after the intervention

During the early response, efforts are dedicated to estimating the likely impact of the outbreak and anticipating the nature, scale and timing of resources needed [7,13,15]. Theoretically, different factors including not only the total number of cases and fatalities but also the morbidity and overall impact on quality of life, as well as societal and economic impact, should ideally be taken into account when attempting to predict disease burden [40–43]. Generally, as the demographic and morbidity data needed by composite measures of health-adjusted life years [40] are lacking in outbreak response contexts, efforts tend to focus on other proxies of impact: assessing transmissibility, predicting future case incidence and associated mortality and investigating risk factors [1,3,7,15].

Analytical needs to diversify as the intervention progresses. While investigations of transmissibility, mortality and risk factors remain key throughout [8], new questions may arise to inform the implementation of control and mitigation measures. These may focus on predicting the impact of potential measures including testing (e.g. 'Could a rapid test help reduce incidence?' [29]), vaccine development (e.g. 'Could a candidate vaccine be evaluated in this outbreak?' [44,45]), vaccination campaigns (e.g. 'Which is the optimal vaccination strategy?' [46,47]) or travel restrictions (e.g. 'Should international travel be restricted?' [48]), or on estimating the impact of current measures such as improvements in access to care (e.g. 'Has the

delay between symptom onset and hospitalization been reduced?' [14,15]). As case incidence reduces, statistical modelling can also be useful for assessing or predicting the end of an outbreak [49–51]. At the field operational level, outbreak response analytics may be best focused on informing and monitoring core surveillance activities and performance indicators, such as contact tracing [11], through the use of tools for contact data visualization [52], mapping [53,54] and on analysis pipelines integrating mobile data collection tools [18,19,55,56] with automated reporting systems [57–59]. Finally, the post-intervention phase lends itself to retrospective studies, which can assess further the impact of interventions [60], tease apart finer processes driving the epidemic dynamics such as contact patterns [12,61], study risk factors [54,62], identify avenues for fortifying surveillance [13,36,63] and evaluate, improve and develop modelling techniques [28,64,65].

### (c) What are outbreak data?

The term '*outbreak data*' encompasses different types of information, of which we first distinguish '*case data*' from '*background data*'. *Case data* include the description of reported cases gathered in *linelists*, i.e. flat files where each row is a case and each column a recorded variable (e.g. dates of onset and admission, gender, age, location), thereby fulfilling the definition of 'tidy data' in the data science community [66]. Case data also include exposure and contact tracing data, either stored within a linelist or in separate files, pathogen whole genome sequencing (WGS) and data pertaining to outbreak investigations (e.g. case–control and cohort study data). *Background data* document the underlying characteristics of the affected populations. This includes demographic information (e.g. maps of population densities, age stratification, mixing patterns), movement data (e.g. borders, traveller flows, migration), health infrastructure (e.g. healthcare facilities, drug stockpiles) and epidemiological data themselves (e.g. levels of pre-existing immunity). A final type of data we consider here is '*intervention data*', which refers to information on decisions made and efforts deployed as part of the intervention, such as vaccination coverage, the extent of active case finding or potential changes in the epidemiological case definition. An in-depth discussion of data needs in outbreaks can be found in Cori *et al*. [13].

## 3. Outbreak analytics

### (a) An overview of the outbreak analytics toolbox

We use the term '*outbreak analytics*' to refer to the variety of tools and methods used to collect, curate, visualize, analyse, model and report on outbreak data. These tools and their inter-dependencies are summarized in an exemplary workflow represented in figure 2, derived from analyses pipelines used during recent epidemics of pandemic influenza [1], MERS-CoV [4] and EVD [7,8,17]. Note that workflows may vary substantially in other epidemic contexts. For instance, analyses of food-borne outbreaks may focus on traceback data [67–69], while vector-borne disease analysis may focus heavily on modelling the vector's ecological niche [70,71].

### (b) Tools for the collection of epidemiological data

Tools for data capture have become a focus of much discussion in recent years as those involved in outbreak response seek to make use of important technological advances including mobile data collection, cloud computing and built-in automated data analyses and reporting. In resource-limited settings, in particular, epidemiological data are still often collected with pen and paper, the advantages of which are familiarity, simplicity, low cost and reliability where access to Internet and power sources may be limited. However, there are some downsides to using paper as a data management tool, becoming increasingly important with larger outbreaks, as any system for the printing and distribution, collection and storage and digitization of forms becomes overwhelmed. Additionally, two-stage processes involving transcription of data from forms typically introduces additional data entry errors [72–75] and substantial delays from data capture to analysis [72].

Electronic data collection (EDC) is becoming increasingly popular [18,19,55,56]. These tools make use of widely available, low-cost hardware (e.g. smartphones and tablets) [76] that can, when appropriately configured, consume little power and collect data offline, making them suitable for use in resource-poor settings. Some of those may be part of existing surveillance systems or be deployed instead for specific enhanced surveillance and response activities during an outbreak. EDC platforms can also enhance data quality through the use of restriction rules and logical checks, and enforce reporting (even when there are zero cases) and entry of essential variables [72,76]. EDC can decrease the delay between data collection, centralization and analysis, which is critical for data-driven responses. Time can be saved through 'form logic' (e.g. automatically skipping sections of a survey not relevant to a participant), while real-time, automated centralization, data analysis and reporting can be directly built into the platform. In addition, mobile-based EDC enables the collection of other types of data including GPS coordinates, photographs, barcode (useful to link case data and clinical specimens) and even aiding diagnostics by directly interfacing with point-of-care diagnostic devices [77–79].

Maintaining confidentiality and privacy is a legitimate concern whenever data concerning human subjects are collected. While EDC systems provide opportunities for unauthorized interception and access to such information, many systems support end-to-end encryption during data transfer [80], although few provide additional security through encryption at the level of data entry.

### (c) Descriptive analyses

The first, and arguably one of the most important steps in data analysis is exploration, where visualization plays a central role, completed with informative summary statistics [81,82]. The first type of graphics needed for rapid assessment of ongoing dynamics is the epidemic curve (epicurve), which shows case incidence time series as a histogram of new onset dates for a given time interval [83–85]. Cumulative case counts, sometimes used in the absence of a raw linelist, are best avoided in epicurves, as they tend to obscure ongoing dynamics and create statistical dependencies in data points that will result in biases and lead to under-estimating uncertainty in downstream modelling [86].

Maps have been at the core of infectious disease epidemiology from a very early stage [87]. Nowadays, they are typically used to visualize the distribution of disease [88], for representing the 'ecological niche' of infectious diseases at large scales [23,24,89] and for assessing the spatial dynamics of an outbreak and strategizing interventions [7,8]. Providers of free and crowd-sourced [90] geographical data like the

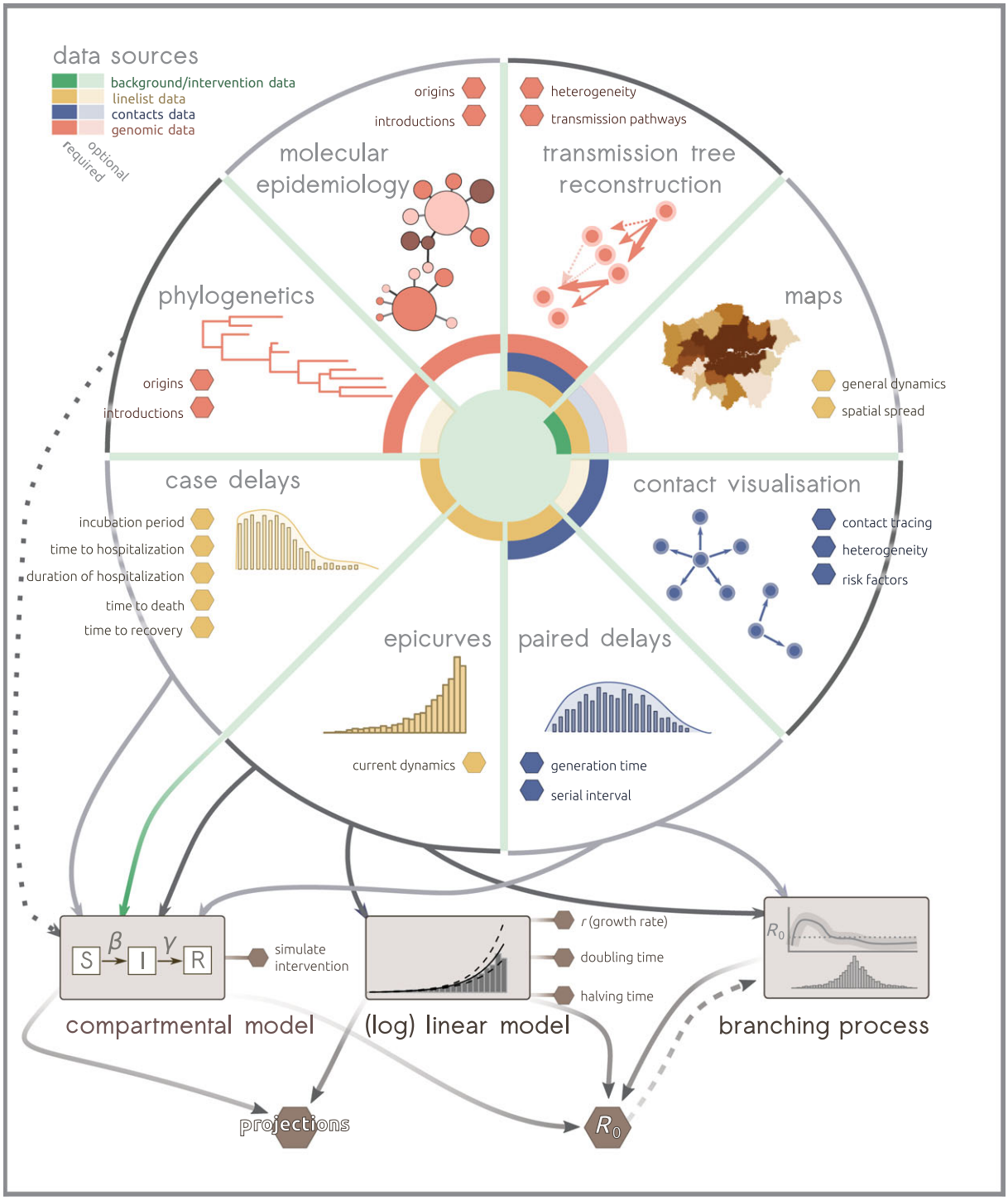



**Figure 2.** Example of outbreak analytics workflow. This schematic represents eight general analyses that can be performed from outbreak data. Outputs containing actionable information for the operations are represented as hexagons. Data needed for each analysis are represented as a different colour in the center, using plain and light shading for mandatory and optional data, respectively. (Online version in colour.)

Humanitarian Open Street Maps Team (Humanitarian OpenStreetMap Team Home; see https://www.hotosm.org/ (accessed 26 September 2018)), the Missing Maps project (MissingMaps; see https://www.missingmaps.org/ (accessed 26 September 2018)), healthsites.io (see https://healthsites.io/ (accessed 26 September 2018)) and the Radiant Earth Foundation (Radiant Earth Foundation – Earth imagery for impact; see https://www.radiant.earth (accessed 18 November 2018)) provide layers of spatial data that include information on the location of households and health facilities, among other determinants. Several tools including SaTScan and ClusterSeer are routinely applied to surveillance system data for automated outbreak detection and the evaluation of clustering of disease by time and space [91]. Other examples of freely available mapping tools that can help track the spread of infectious diseases include the Spatial Epidemiology of Viral Haemorrhagic Fevers (VHF) disease visualization (see http://www.healthdata.org/datavisualization/spatial-epidemiology-viralhemorrhagic-fevers; accessed 19 September 2018), which maps risks of emergence and spread of VHF diseases, Nextstrain [92] and Microreact [93], which focus on mapping pathogen evolution and epidemic spread, and HealthMap [94], which provides resources for the rapid detection of outbreaks. Geographical locations of reported cases can also be useful for informing more complex modelling approaches [95].

In epidemics driven by person-to-person transmission, a last essential source of data is contact data [20], which includes data on case *exposure* [12] as well as *contact tracing*, where appropriate [11,63]. Exposure data document transmission pairs, which can yield precious insights into 'paired delays' (figure 2) including the serial interval (time between onsets of a case and their infector) or the generation time (time between the dates of infections of a case and their infector) [7,8], which are in turn useful for estimating transmissibility [27,28,96,97]. Exposure data can also be used to investigate the occurrence and determinants of super-spreading events [12] and help identify introduction events in the case of zoonotic diseases [98]. Contact tracing, through the early detection of new cases and their subsequent isolation and treatment, plays a central role in reducing onward transmission and therefore containing outbreaks [11,63,99], while additionally providing potential information on risk factors [7,11].

Summary statistics are a useful complement to data visualization in the exploratory phase of data analysis. Some metrics, such as transmissibility, require the use of statistical or mathematical models in order to be estimated (see §3d below) and are therefore not readily available as descriptive tools. Other useful statistics can be readily computed from linelists, including different demographic indicators of the reported cases (e.g. gender, age, occupation [7,100,101]), case fatality ratios (the proportion of cases who died of the infection) or case delays such as the times to hospitalization, recovery or death, reported as a whole [1,7,8] or stratified by groups [100,101]. The incubation period (time from infection to symptom onset) is another important delay for informing the intervention (e.g. to define the duration of contact tracing or declare the end of an outbreak), but can be harder to derive as it requires data on case exposure as well. Note that in the case of delays, these are best analysed by characterising the full distribution (e.g. by fitting to an appropriate probability distribution such as discretized Gamma [7]) rather than reported as a single central value [7,8,102,103].

## (d) Quantifying transmissibility

The 'transmissibility' of a disease is here used to refer to the rate at which new cases arise in the population, resulting either in epidemic growth or decline [1,3,27,28]. Rather than an intrinsic property of a specific disease, transmissibility thus defined quantifies the propagation of a pathogen in a given epidemic setting and is impacted by multiple factors including population demographics, mixing and levels pre-existing immunity. Importantly, estimates of transmissibility reported in the literature will typically be biased towards higher values, as subcritical outbreaks are by definition less likely to be detected. Several metrics of transmissibility can be used depending on the type of data available and can be estimated using different approaches.

A first measure of transmissibility is the *growth rate* ($r$), which is estimated from a simple model where case incidence is either exponentially growing ($r > 0$) or declining ($r < 0$). Typically, $r$ is estimated directly from epicurves (figure 2) using a log-linear model, where $r$ is defined as the slope of a linear regression on log-transformed incidence [104,105]. Besides its simplicity and its computational efficiency, this approach has the benefits of being embedded in the linear modelling framework, thereby allowing one to measure the uncertainty associated with a given estimate of $r$, to test for differences in growth rates, e.g. between different locations, and to derive short-term incidence predictions. Moreover, the growth rate can also be used to estimate the doubling and halving times of the epidemic, i.e. the time during which incidence doubles (respectively is halved), as alternative metrics of transmissibility [103]. Unfortunately, the log-linear model can only fit exponentially growing or decaying outbreaks, which may not always be appropriate in the presence of complex spatial or age structure, or owing to changes in reporting, transmissibility or proportion of susceptible individuals over time. Besides, it cannot readily accommodate time periods with no cases, so that its applicability may in practice be restricted.

While $r$ quantifies the *speed* at which a disease spreads, it does not contain information on the *level* of the intervention that is necessary to control a disease [106]. This is better characterized by the *reproduction number* (here generically noted '$R$'), which measures the average number of secondary cases caused by each primary case. Researchers typically distinguish the basic reproduction number ($R_0$ [104]), which applies in a large, fully susceptible population, without any control measures, from the effective reproduction number ($R_{\text{eff}}$), which is the number of secondary cases after accounting for behavioural changes, interventions and declines in susceptibility [96]. The current reproduction number determines the dynamics of the epidemic in the near future, with values greater than 1 predicting an increase in cases, and values less than 1 predicting control [104]. The value of $R$ can also be used to calculate the fraction of the population that needs to be immunized (typically through vaccination) in order to contain an outbreak [104].

Different methodological approaches have been developed to estimate the reproduction number. $R$ can be approximated using estimates of the growth rate $r$ combined with knowledge of the generation time distribution [97]. $R$ can also be derived from compartmental models [104,107]. The formula will depend on the type of model used, but such estimation will usually require that different rates (e.g. rates of infection, recovery, death) are either known or estimated by fitting the model to data [104,107]. Real-world complexities can be incorporated into this approach; however, fitting such models can be challenging and may require computationally intensive algorithms such as data augmentation, approximate bayesian computation, or particle filters [108]. Compartmental models also require assumptions about the total population size and the proportion of the population at risk, which may be difficult to estimate in an outbreak. As an alternative, branching process models can be used to estimate $R$ directly from incidence data [27,28,96,109]. This requires a pre-specified distribution of the generation time, or of the serial interval, although recent developments suggest that in some cases, the generation time distribution itself can also be simultaneously estimated [4]. Branching process models are usually much simpler to fit to data than their compartmental counterparts, which facilitates their use in real time [27].

Beyond the mere estimation of transmissibility, it is often essential to forecast future incidence for advocacy and planning purposes, e.g. to compare different interventions and epidemic scenarios [7,8,15,30]. A variety of mathematical and statistical models, including those reviewed here for estimating transmissibility, can also be used for short-term incidence forecasting [65]. Despite the growing body of research focusing on predicting incidence during epidemics [65,110], there are currently no gold standards and the relative performances of forecasting methods largely remain to be assessed. Methods

that have been developed and applied in other fields to rigorously assess not just the accuracy of forecasts but also how well models quantify the inherent uncertainty in making predictions, are only rarely applied in infectious disease epidemiology [111,112]. Whether it is to estimate $R$ or predict future incidence, the most appropriate method ultimately depends on the particular epidemiological setting, existing knowledge of the transmission dynamics and data availability. Branching process models, for example, can be used for a quick estimate of the current value of $R$ from the recent trend in case numbers and, by extrapolating this forward, of expected case numbers in the near future [27,28,96]. Mechanistic or simulation models, on the other hand, aim to include a more explicit representation of the different factors that might influence transmission. They can be a more natural choice for assessing the expected impact of possible interventions, but they usually require careful parametrization and often intensive computation [29,30,45,113], both of which can be challenging early in an outbreak when data are scarce and rapid turnaround crucial.

### (e) Analytical epidemiological techniques

Analytical epidemiological studies use data to better describe outbreaks and populations at risk and inform real-time and subsequent response efforts. Typically, these are conducted during the intervention and post-intervention phases of an outbreak response (figure 1). They include observational designs such as retrospective cohort and case–control studies to identify risk factors and quantify associations between potential causes and their outcomes (typically, the disease in question), and experimental designs, such as randomized-control studies used to estimate the impact of interventions such as vaccination and treatments [114]. These studies reside outside of the normal scope of outbreak response activities, being inserted *ad hoc* as functions that are not necessarily routine response activities such as strengthening surveillance. In the case of observational epidemiological studies, data on exposures and outcomes are required, permitting estimations of the increased risk of disease among people exposed to risk factors of interest [54,62,115,116]. In the case of experimental epidemiology, data on outcomes of interest are collected to permit estimations of heterogeneity among groups (e.g. in the presence/absence of intervention).

The usefulness of such studies in informing outbreak response is highly context-dependent. Observational studies may be undertaken early on in the intervention phase to help identify ongoing infection sources of environmental, food-borne or water-borne nature [117] and to stop the outbreak at its source. In longer-running outbreaks, they can provide insights into opportunities for control [53,115,118] and inform global policy decisions that relate to outbreak response [119]. However, the time and expertise needed to prepare and implement these studies may preclude their application in the midst of an ongoing outbreak, so that the cost and benefits of such an undertaking need to be carefully weighed in emergency settings.

### (f) Genetic analyses

Whole genome sequencing of pathogens is increasingly affordable and reliable, and therefore more frequent in outbreak investigations [1,120–126]. As technology is making real-time sequencing in the field a developing standard in the coming years [127,128], genetic analysis will likely carve out its own space in the outbreak analytics toolkit.

Different methods can be used to extract information from pathogen WGS. In bacterial genomics, molecular epidemiology methods have been used extensively for defining strains of related isolates [32,129], which can be used to infer various features of the pathogens sampled such as their origins, antimicrobial resistance profiles, virulence or antigenic characteristics [130–132]. These methods usually exploit only a fraction of the information contained within pathogens' genomes, as they rely on genetic variation in a limited number of housekeeping genes [32,129]. While these methods will likely remain useful in years to come, substantially more information can be extracted by using WGS to reconstruct phylogenetic trees, which represent the evolutionary history of the sampled isolates, assuming the absence of selection or horizontal gene transfers [133]. Different types of phylogenetic reconstruction methods can be used, including fast, scalable distance-based methods [134] or more computer-intensive approaches using a maximum-likelihood [135,136] or the Bayesian framework [33,137]. Phylogenies can be used to assess the origins of a set of pathogens [138], patterns of geographical spread [125], host species jumps [139,140], past fluctuations in the pathogen population sizes [141] and even, in some cases, the reproduction number [1]. Importantly, there is a growing tendency to analyse phylogenetic trees in the broader context of other epidemiological data (mainly geographical locations until now), which is facilitated by user-friendly Web applications [92,93].

A further step towards integrating WGS alongside epidemiological data is the reconstruction of transmission trees (who infects whom) using evidence synthesis approaches. This methodological field has been growing fast over the past decade [25,142–148], but most applications of these methods remain within academia and their usefulness in the field in an outbreak response context needs to be critically assessed. A potential benefit of accurately reconstructing transmission trees lies in the identification of multiple introductions, the quantification of the proportion of unreported cases and the detection of heterogeneities in individual transmissibility [145]. Unfortunately, the reconstruction of transmission trees is a difficult and computationally intensive problem. First, most diseases do not accumulate sufficient genetic diversity during the course of an outbreak to allow the accurate reconstruction of transmission chains, so that multiple data sources need to be used [35], making these methods more data-demanding than most other approaches in outbreak analytics (figure 2). In addition, the complex nature of the problem requires the use of Bayesian methods for model fitting, making these approaches difficult to interpret by non-experts [145,146,148].

## 4. Discussion

In this article, we reviewed methodological and technological resources forming the basis of outbreak analytics, an emerging data science for informing outbreak response. Outbreak analytics is embedded within a broader public health information context that starts with disease surveillance systems, followed by risk assessment and management, the epidemiological response itself, and finishes with the production of actionable information for decision making. Part of the challenge that this new field will face in the coming years

pertains to the seamless integration of data analytics pipelines within existing workflows. As responders can allocate only limited time to data analysis, analytics resources should produce simple, interpretable results, highlighting the most pressing issues that need addressing and monitoring all relevant indicators to inform the response.

Outbreak analytics and resulting outputs are central to the surveillance pillar of any outbreak response, yet resources and capacities to ensure data availability and quality are often limited owing to operational constraints [16]. Priorities in terms of data needs should be defined by what actionable information it may give access to through the available analytics pipelines [13]. In this respect, we foresee that typical linelist data such as dates of events (e.g. onset, reporting, hospitalization, discharge), age, gender, disease outcome, geographical locations and exposure data will fulfil most needs, while other data such as WGS may only be useful for specific diseases and contexts [34,35]. Intervention data are rarely collected but should be given more consideration, as they are key to assessing the impact and effectiveness of control measures, both during and after the operations. Similarly, data on the fraction of cases reported (and its variations through time), as well as behavioural changes (e.g. care-seeking behaviour) in the affected populations, can be very important sources of information for modelling [149].

Fortunately, what we called 'background data' in this article can be gathered and shared outside of the epidemic context. Besides maps, population census, sero-surveys or genetic databanks, data on the natural histories of diseases derived from past epidemics, such as key delay distributions and transmissibility, can form a useful substitute to real-time estimates, especially in the early stages of outbreaks when such data may be lacking. While crowd-sourced initiatives are promising and have been used successfully in low resource settings [90], more efforts are needed to collate and curate open data sources, assess their quality and make them widely available to the community. We argue that international public health agencies and non-governmental agencies should play a central role in orchestrating such background data preparedness.

Outbreak analytics is a developing field, and as such, there remain many gaps in terms of data collection, analysis and reporting tools. Some methodological challenges persist, such as better characterising forecasting methods [28,64,65], including spatial information and population flows into existing transmission models [95], and improving the integration of different types of data for reconstructing transmission trees [35]. In order to ensure transparent methods and availability to analysts in any setting, the implementation must be as freely available, open-source software. Among other popular programming languages, such as Python, Java, or Julia, the R software [150] arguably offers the largest collection of free tools for data analysis and reporting, and an increasing number of packages for infectious disease epidemiology [20,21,27,84,145] may form a solid starting point for the development of a comprehensive, robust and transparent toolkit for the analysis of epidemic data [151]. Importantly, the use of a common platform for the development and use of outbreak analytics tools will also likely contribute to standardizing data practices, including collection, sharing and analysis.

A final point relates to the use and dissemination of these new resources: how can outbreak analytics best help improve public health? As noted by Bausch & Clougherty [39], *health science should not be an entity unto itself, but a means to an end*. Insofar as it can help field epidemiologists collect, visualize and analyse data, and subsequently provide decision-makers with actionable information, outbreak analytics will likely occupy an increasing space in field epidemiology over the years to come. We foresee that the dissemination of free training resources [152], the modernization of field epidemiology training programmes and the deployment of applied data scientists to the field with a sustained capacity building in resource-poor and vulnerable countries will be instrumental in shaping the future of this emerging field of health science.

Data accessibility. This article has no additional data.

Authors' contributions. T.J. drafted the outline of the review and revised the manuscript. J.A.P., A.B., T.J. wrote the first draft of the manuscript. Z.N.K. produced the figures. A.C., W.J.E., R.M.E., S.F., L.K., P.K., M.M., P.M., P.N., O.P.W., R.R., J.W. contributed the content.

Competing interests. We declare we have no competing interests.

Funding. This paper was supported with funding from the Global Challenges Research Fund (GCRF) for the project 'RECAP—research capacity building and knowledge generation to support preparedness and response to humanitarian crises and epidemics' managed through RCUK and ESRC (ES/P010873/1). P.K., O.L.P., J.W. and T.J. receive support from the UK Public Health Rapid Support Team, which is funded by the United Kingdom Department of Health and Social Care. We acknowledge the National Institute for Health Research—Health Protection Research Unit for Modelling Methodology (T.J.) for funding. M.M., C.H.R., receive funding through the National Institute for Health Research (PR-OD-1017-20001). R.M.E. acknowledges funding from an HDR UK Innovation Fellowship (grant no. MR/S003975/1). A.C. thanks the Medical Research Council for funding. S.F. was supported by the Wellcome Trust (210758/Z/18/Z). The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated.

Acknowledgements. We would like to thank Annick Lenglet and Isidro Carrion-Martin, Epidemiologists at Medecins Sans Frontieres (MSF, Operational Centre Amsterdam) for their additional reflections. The views expressed in this publication are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research or the Department of Health and Social Care. The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated.

## References

1. Fraser C *et al.* 2009 Pandemic potential of a strain of influenza A (H1N1): early findings. *Science* **324**, 1557–1561. (doi:10.1126/science.1176062)
2. Assiri A *et al.* 2013 Hospital outbreak of Middle East respiratory syndrome coronavirus. *N. Engl. J. Med.* 369, 407–416. (doi:10.1056/NEJMoa1306742)
3. Cauchemez S, Fraser C, Van Kerkhove MD, Donnelly CA, Riley S, Rambaut A, Enouf V, van der Werf S, Ferguson NM. 2014 Middle East respiratory syndrome coronavirus: quantification of the extent of the epidemic, surveillance biases, and transmissibility. *Lancet Infect. Dis.* **14**, 50–56. (doi:10.1016/S1473-3099(13)70304-9)
4. Cauchemez S *et al.* 2016 Unraveling the drivers of MERS-CoV transmission. *Proc. Natl Acad. Sci. USA* **113**, 9081–9086. (doi:10.1073/pnas.1519235113)
5. Campos GS, Bandeira AC, Sardi SI. 2015 Zika virus outbreak, Bahia, Brazil. *Emerg. Infect. Dis.* **21**, 1885–1886. (doi:10.3201/eid2110.150847)

6. European Centre for Disease Prevention and Control. 2015 Zika virus epidemic in the Americas: potential association with microcephaly and Guillain-Barré syndrome (first update), 21 January 2016. (See https://ecdc.europa.eu/sites/portal/files/media/en/publications/Publications/zika-virus-americas-association-with-microcephaly-rapid-risk-assessment.pdf).
7. WHO Ebola Response Team. 2014 Ebola virus disease in West Africa – the first 9 months of the epidemic and forward projections. *N. Engl. J. Med.* **371**, 1481–1495. (doi:10.1056/NEJMoa1411100)
8. WHO Ebola Response Team *et al.* 2015 West African Ebola epidemic after one year – slowing but not yet under control. *N. Engl. J. Med.* **372**, 584–587. (doi:10.1056/NEJMc1414992)
9. Moon S *et al*. 2015 Will Ebola change the game? Ten essential reforms before the next pandemic. The report of the Harvard-LSHTM independent panel on the global response to Ebola. *Lancet* **386**, 2204–2221. (doi:10.1016/S0140-6736(15)00946-0)
10. Van Kerkhove MD, Bento AI, Mills HL, Ferguson NM, Donnelly CA. 2015 A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. *Sci Data* **2**, 150019. (doi:10.1038/sdata.2015.19)
11. Senga M *et al*. 2017 Contact tracing performance during the Ebola virus disease outbreak in Kenema district, Sierra Leone. *Phil. Trans. R. Soc. B* **372**, 20160300. (doi:10.1098/rstb.2016.0300)
12. International Ebola Response Team. 2016 Exposure patterns driving Ebola transmission in West Africa: a retrospective observational study. *PLoS Med.* **13**, e1002170. (doi:10.1371/journal.pmed.1002170)
13. Cori A *et al*. 2017 Key data for outbreak evaluation: building on the Ebola experience. *Phil. Trans. R. Soc. B* **372**, 20160371. (doi:10.1098/rstb.2016.0371)
14. Lewnard JA. 2018 Ebola virus disease: 11 323 deaths later, how far have we come? *Lancet* **392**, 189–190. (doi:10.1016/S0140-6736(18)31443-0)
15. Ebola Outbreak Epidemiology Team. 2018 Outbreak of Ebola virus disease in the Democratic Republic of the Congo, April–May, 2018: an epidemiological study. *Lancet* **392**, 213–221. (doi:10.1016/S0140-6736(18)31387-4)
16. Polonsky J *et al*. 2019 Lessons learnt from Ebola virus disease surveillance in Équateur Province, May–July 2018. *Weekly Epidemiological Record* **94**, 23–27.
17. 2017 WHO | Ebola outbreak Democratic Republic of the Congo 2017. See https://www.who.int/emergencies/ebola-DRC-2017/en/.
18. Hartung C, Lerer A, Anokwa Y, Tseng C, Brunette W, Borriello G. 2010 Open Data Kit: Tools to build information services for developing regions. In *Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development*, pp. 18:1–18:12. New York, NY: ACM.
19. Brunette W, Sundt M, Dell N, Chaudhri R, Breit N, Borriello G. 2013 Open Data Kit 2.0: Expanding and refining information services for developing regions. In *Proceedings of the 14th Workshop on Mobile Computing Systems and Applications*, pp. 10:1–10:6. New York, NY: ACM.
20. Nagraj VP, Randhawa N, Campbell F, Crellen T, Sudre B, Jombart T. 2018 epicontacts: Handling, visualisation and analysis of epidemiological contacts. *F1000Res.* **7**, 566. (doi:10.12688/f1000research.14492.1)
21. Moraga P, Dorigatti I, Kamvar ZN, Piatkowski P, Toikkanen SE, Nagraj VP, Donnelly CA, Jombart T. 2018 epiflows: an R package for risk assessment of travel-related spread of disease. *F1000Res.* **7**, 1374. (doi:10.12688/f1000research.16032.1)
22. Pigott DM *et al*. 2017 Local, national, and regional viral haemorrhagic fever pandemic potential in Africa: a multistage analysis. *Lancet* **390**, 2662–2672. (doi:10.1016/S0140-6736(17)32092-5)
23. Messina JP *et al*. 2016 Mapping global environmental suitability for Zika virus. *Elife* **5**, e15272. (doi:10.7554/eLife.15272)
24. Pigott DM *et al*. 2014 Mapping the zoonotic niche of Ebola virus disease in Africa. *Elife* **3**, e04395. (doi:10.7554/eLife.04395)
25. Jombart T, Eggo RM, Dodd PJ, Balloux F. 2011 Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity* **106**, 383–390. (doi:10.1038/hdy.2010.78)
26. Famulare M, Hu H. 2015 Extracting transmission networks from phylogeographic data for epidemic and endemic diseases: Ebola virus in Sierra Leone, 2009 H1N1 pandemic influenza and polio in Nigeria. *Int. Health* **7**, 130–138. (doi:10.1093/inthealth/ihv012)
27. Cori A, Ferguson NM, Fraser C, Cauchemez S. 2013 A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.* **178**, 1505–1512. (doi:10.1093/aje/kwt133)
28. Nouvellet P *et al*. 2017 A simple approach to measure transmissibility and forecast incidence. *Epidemics* **22**, 29–35. (doi:10.1016/j.epidem.2017.02.012)
29. Nouvellet P *et al*. 2015 The role of rapid diagnostics in managing Ebola epidemics. *Nature* **528**, S109–S116. (doi:10.1038/nature16041)
30. Finger F, Funk S, White K, Siddiqui R, John Edmunds W, Kucharski AJ. 2018 Real-time analysis of the diphtheria outbreak in forcibly displaced Myanmar nationals in Bangladesh. *bioRxiv*. 388645. (doi:10.1101/388645)
31. Bausch DG, Edmunds J. 2018 Real-time modeling should be routinely integrated into outbreak response. *Am. J. Trop. Med. Hyg.* **98**, 1214–1215. (doi:10.4269/ajtmh.18-0150)
32. Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. 2004 eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J. Bacteriol.* **186**, 1518–1530. (doi:10.1128/JB.186.5.1518-1530.2004)
33. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014 BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537. (doi:10.1371/journal.pcbi.1003537)
34. Holmes EC, Rambaut A, Andersen KG. 2018 Pandemics: spend on surveillance, not prediction. *Nature* **558**, 180–182. (doi:10.1038/d41586-018-05373-w)
35. Campbell F, Strang C, Ferguson N, Cori A, Jombart T. 2018 When are pathogen genome sequences informative of transmission events? *PLoS Pathog.* **14**, e1006885. (doi:10.1371/journal.ppat.1006885)
36. Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, Pablos-Méndez A, Tomori O, Mazet JAK. 2018 The Global Virome Project. *Science* **359**, 872–874. (doi:10.1126/science.aap7463)
37. Birrell PJ, De Angelis D, Presanis AM. 2018 Evidence synthesis for stochastic epidemic models. *Stat. Sci.* **33**, 34–43. (doi:10.1214/17-STS631)
38. Wikipedia contributors. 2018 Data science. *Wikipedia, The Free Encyclopedia*. See https://en.wikipedia.org/w/index.php?title=Data_science&oldid=868658447 (accessed on 16 November 2018).
39. Bausch DG, Clougherty MM. 2015 Ebola virus: sensationalism, science, and human rights. *J. Infect. Dis.* **212**(Suppl. 2), S79–S83. (doi:10.1093/infdis/jiv359)
40. Kwong JC *et al*. 2012 The impact of infection on population health: results of the Ontario burden of infectious diseases study. *PLoS ONE* **7**, e44103. (doi:10.1371/journal.pone.0044103)
41. Vos T *et al*. 2012 Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* **380**, 2163–2196. (doi:10.1016/S0140-6736(12)61729-2)
42. Global Burden of Disease Study 2013 Collaborators. 2015 Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **386**, 743–800. (doi:10.1016/S0140-6736(15)60692-4)
43. Prüss-Üstün A *et al*. 2003 *Introduction and methods: assessing the environmental burden of disease at national and local levels*. WHO Environmental Burden of Disease Series, No. 1. Geneva, Switzerland: World Health Organization.
44. Camacho A, Eggo RM, Funk S, Watson CH, Kucharski AJ, Edmunds WJ. 2015 Estimating the probability of demonstrating vaccine efficacy in the declining Ebola epidemic: a Bayesian modelling approach. *BMJ Open* **5**, e009346. (doi:10.1136/bmjopen-2015-009346)
45. Camacho A *et al*. 2017 Real-time dynamic modelling for the design of a cluster-randomized phase 3 Ebola vaccine trial in Sierra Leone. *Vaccine* **35**, 544–551. (doi:10.1016/j.vaccine.2016.12.019)
46. Garske T, Van Kerkhove MD, Yactayo S, Ronveaux O, Lewis RF, Staples JE, Perea W, Ferguson NM, Yellow Fever Expert Committee. 2014 Yellow fever in Africa: estimating the burden of disease and impact of mass vaccination from outbreak and serological

data. *PLoS Med.* **11**, e1001638. (doi:10.1371/journal.pmed.1001638)

47. Kraemer MUG *et al*. 2017 Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015–16: a modelling study. *Lancet Infect. Dis.* **17**, 330–338. (doi:10.1016/S1473-3099(16)30513-8)
48. Dorigatti I, Hamlet A, Aguas R, Cattarino L, Cori A, Donnelly CA, Garske T, Imai N, Ferguson NM. 2017 International risk of yellow fever spread from the ongoing outbreak in Brazil, December 2016 to May 2017. *Euro Surveill.* **22**, 30572. (doi:10.2807/1560-7917.ES.2017.22.28.30572)
49. Brookmeyer R, You X. 2006 A hypothesis test for the end of a common source outbreak. *Biometrics* **62**, 61–65. (doi:10.1111/j.1541-0420.2005.00421.x)
50. Nishiura H, Miyamatsu Y, Chowell G, Saitoh M. 2015 Assessing the risk of observing multiple generations of Middle East respiratory syndrome (MERS) cases given an imported case. *Euro Surveill.* **20**, 21181. (doi:10.2807/1560-7917.ES2015.20.27.21181)
51. Nishiura H, Miyamatsu Y, Mizumoto K. 2016 Objective determination of end of MERS outbreak, South Korea, 2015. *Emerg. Infect. Dis.* **22**, 146–148. (doi:10.3201/eid2201.151383)
52. Fähnrich C *et al*. 2015 Surveillance and outbreak response management system (SORMAS) to support the control of the Ebola virus disease outbreak in West Africa. *Euro Surveill.* **20**, 21071. (doi:10.2807/1560-7917.ES2015.20.12.21071)
53. Polonsky JA, Martínez-Pino I, Nackers F, Chonzi P, Manangazira P, Van Herp M, Maes P, Porten K, Luquero FJ. 2014 Descriptive epidemiology of typhoid fever during an epidemic in Harare, Zimbabwe, 2012. *PLoS One* **9**, e114702. (doi:10.1371/journal.pone.0114702)
54. Page A-L *et al*. 2015 Geographic distribution and mortality risk factors during the cholera outbreak in a rural region of Haiti, 2010–2011. *PLoS Neglect. Trop. Dis.* **9**, e0003605. (doi:10.1371/journal.pntd.0003605)
55. Aanensen DM, Huntley DM, Feil EJ, al-Own F, Spratt BG. 2009 EpiCollect: linking smartphones to web applications for epidemiology, ecology and community data collection. *PLoS One* **4**, e6968. (doi:10.1371/journal.pone.0006968)
56. Tom-Aba D *et al*. 2015 Innovative technological approach to Ebola virus disease outbreak response in Nigeria using the open data kit and form hub technology. *PLoS One* **10**, e0131000. (doi:10.1371/journal.pone.0131000)
57. Xie Y, Allaire JJ, Grolemund G. 2018 *R markdown: The definitive guide*. Boca Raton, FL: Chapman and Hall/CRC. https://bookdown.org/yihui/rmarkdown/.
58. Xie Y. 2016 *Bookdown: authoring books and technical documents with R markdown*. Boca Raton, FL: CRC Press.
59. Karo B, Haskew C, Khan AS, Polonsky JA, Mazhar MKA, Buddha N. 2018 World Health Organization early warning, alert, and response system in the Rohingya Crisis, Bangladesh, 2017–2018. *Emerg. Infect. Dis.* **24**, 2074–2076. (doi:10.3201/eid2411.181237)
60. Cauchemez S, Valleron A-J, Boëlle P-Y, Flahault A, Ferguson NM. 2008 Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature* **452**, 750–754. (doi:10.1038/nature06732)
61. Cauchemez S, Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, Ferguson NM, Swerdlow D, Pennsylvania H1N1 working group. 2011 Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proc. Natl Acad. Sci. USA* **108**, 2825–2830. (doi:10.1073/pnas.1008895108)
62. Gignoux E, Polonsky J, Ciglenecki I, Bichet M, Coldiron M, Thuambe Lwiyo E, Akonda I, Serafini M, Porten K. 2018 Risk factors for measles mortality and the importance of decentralized case management during an unusually large measles epidemic in eastern Democratic Republic of Congo in 2013. *PLoS One* **13**, e0194276. (doi:10.1371/journal.pone.0194276)
63. Saurabh S, Prateek S. 2017 Role of contact tracing in containing the 2014 Ebola outbreak: a review. *Afr. Health Sci.* **17**, 225–236. (doi:10.4314/ahs.v17i1.28)
64. Funk S, Camacho A, Kucharski AJ, Eggo RM, Edmunds WJ. 2018 Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. *Epidemics* **22**, 56–61. (doi:10.1016/j.epidem.2016.11.003)
65. Viboud C *et al*. 2018 The RAPIDD Ebola forecasting challenge: synthesis and lessons learnt. *Epidemics* **22**, 13–21. (doi:10.1016/j.epidem.2017.08.002)
66. Wickham H. 2014 Tidy data. *J. Stat. Softw.* **59**, 1–23. (doi:10.18637/jss.v059.i10)
67. Dallman T *et al*. 2016 Phylogenetic structure of European *Salmonella enteritidis* outbreak correlates with national and international egg distribution network. *Microb Genom* **2**, e000070. (doi:10.1099/mgen.0.000070)
68. Jenkins C *et al*. 2015 Public health investigation of two outbreaks of shiga toxin-producing *Escherichia coli* O157 associated with consumption of watercress. *Appl. Environ. Microbiol.* **81**, 3946–3952. (doi:10.1128/AEM.04188-14)
69. Inns T *et al*. 2015 A multi-country *Salmonella enteritidis* phage type 14b outbreak associated with eggs from a German producer: 'near real-time' application of whole genome sequencing and food chain investigations, United Kingdom, May to September 2014. *Eurosurveillance* **20**, 21098. (doi:10.2807/1560-7917.ES2015.20.16.21098)
70. Bousema T *et al*. 2016 The impact of hotspot-targeted interventions on malaria transmission in Rachuonyo South District in the Western Kenyan highlands: a cluster-randomized controlled trial. *PLoS Med.* **13**, e1001993. (doi:10.1371/journal.pmed.1001993)
71. Baidjoe AY *et al*. 2016 Factors associated with high heterogeneity of malaria at fine spatial scale in the Western Kenyan highlands. *Malar. J.* **15**, 307. (doi:10.1186/s12936-016-1362-y)
72. Ahmed R, Robinson R, Elsony A, Thomson R, Squire SB, Malmborg R, Burney P, Mortimer K. 2018 A comparison of smartphone and paper data-collection tools in the Burden of Obstructive Lung Disease (BOLD) study in Gezira state, Sudan. *PLoS One* **13**, e0193917. (doi:10.1371/journal.pone.0193917)
73. Solomon AW *et al*. 2018 Quality assurance and quality control in the global trachoma mapping project. *Am. J. Trop. Med. Hyg.* **99**, 858–863. (doi:10.4269/ajtmh.18-0082)
74. King JD *et al*. 2013 A novel electronic data collection system for large-scale surveys of neglected tropical diseases. *PLoS One* **8**, e74570. (doi:10.1371/journal.pone.0074570)
75. Njuguna HN *et al*. 2014 A comparison of smartphones to paper-based questionnaires for routine influenza sentinel surveillance, Kenya, 2011–2012. *BMC Med. Inform. Decis. Mak.* **14**, 107. (doi:10.1186/s12911-014-0107-5)
76. Poushter J. 2016 Smartphone ownership and internet usage continues to climb in emerging economies. *Pew Res. Center* **22**, 1–44.
77. Bogoch II, Koydemir HC, Tseng D, Ephraim RKD, Duah E, Tee J, Andrews JR, Ozcan A. 2017 Evaluation of a mobile phone-based microscope for screening of *Schistosoma haematobium* infection in rural Ghana. *Am. J. Trop. Med. Hyg.* **96**, 1468–1471. (doi:10.4269/ajtmh.16-0912)
78. Kühnemund M *et al*. 2017 Targeted DNA sequencing and *in situ* mutation analysis using mobile phone microscopy. *Nat. Commun.* **8**, 13913. (doi:10.1038/ncomms13913)
79. Quesada-González D, Merkoçi A. 2017 Mobile phone-based biosensing: an emerging 'diagnostic and communication' technology. *Biosens. Bioelectron.* **92**, 549–562. (doi:10.1016/j.bios.2016.10.062)
80. Macharia P, Dunbar MD, Sambai B, Abuna F, Betz B, Njoroge A, Bukusi D, Cherutich P, Farquhar C. 2015 Enhancing data security in open data kit as an mHealth application. In *2015 International Conference on Computing, Communication and Security (ICCCS)*, Pamplemousses, Mauritius, 4–5 December 2015. (doi:10.1109/cccs.2015.7374205)
81. Crawley MJ. 2012 *The R book*. Hoboken, NJ: John Wiley & Sons.
82. Wickham H. 2016 *Ggplot2: elegant graphics for data analysis*. Berlin, Germany: Springer.
83. Höhle M. 2007 surveillance: An R package for the monitoring of infectious diseases. *Comput. Stat.* **22**, 571–582. (doi:10.1007/s00180-007-0074-8)
84. Jombart T *et al*. 2014 OutbreakTools: a new platform for disease outbreak analysis using the R software. *Epidemics* **7**, 28–34. (doi:10.1016/j.epidem.2014.04.003)
85. Jombart T, Kamvar ZN, FitzJohn R. 2018 Incidence: compute, handle, plot and model incidence of dated events. R package version 1.5.4. https://CRAN.R-project.org/package=incidence.
86. King AA, Domenech de Cellès M, Magpantay FMG, Rohani P. 2015 Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc. R. Soc. B* **282**, 20150347. (doi:10.1098/rspb.2015.0347)

87. Snow J. 1855 *On the mode of communication of cholera*. London, UK: John Churchill.
88. Wertheim HFL, Horby P, Woodall JP. 2012 *Atlas of human infectious diseases*. Hoboken, NJ: John Wiley & Sons.
89. Nunes MRT *et al*. 2015 Emergence and potential for spread of Chikungunya virus in Brazil. *BMC Med.* **13**, 102. (doi:10.1186/s12916-015-0348-x)
90. In press. Radiant Earth Foundation – Earth imagery for impact. See https://www.radiant.earth (accessed on 18 November 2018).
91. In press. Spatial epidemiology of Viral Hemorrhagic Fevers. See http://www.healthdata.org/data-visualization/spatial-epidemiology-viral-hemorrhagic-fevers (accessed on 19 September 2018).
92. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018 Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123. (doi:10.1093/bioinformatics/bty407)
93. Argimón S *et al*. 2016 Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* **2**, e000093. (doi:10.1099/mgen.0.000093)
94. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. 2008 HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J. Am. Med. Inform. Assoc.* **15**, 150–157. (doi:10.1197/jamia.M2544)
95. Backer JA, Wallinga J. 2016 Spatiotemporal analysis of the 2014 Ebola epidemic in West Africa. *PLoS Comput. Biol.* **12**, e1005210. (doi:10.1371/journal.pcbi.1005210)
96. Wallinga J, Teunis P. 2004 Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidemiol.* **160**, 509–516. (doi:10.1093/aje/kwh255)
97. Wallinga J, Lipsitch M. 2007 How generation intervals shape the relationship between growth rates and reproductive numbers. *Proc. R. Soc. B* **274**, 599–604. (doi:10.1098/rspb.2006.3754)
98. Cauchemez S, Van Kerkhove MD, Riley S, Donnelly CA, Fraser C, Ferguson NM. 2013 Transmission scenarios for Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and how to tell them apart. *Euro Surveill.* **18**, 20503.
99. Shrivastava SR, Shrivastava PS, Ramasamy J. 2014 Utility of contact tracing in reducing the magnitude of Ebola disease. *Germs* **4**, 97–99. (doi:10.11599/germs.2014.1063)
100. WHO Ebola Response Team *et al*. 2015 Ebola virus disease among children in West Africa. *N. Engl. J. Med.* **372**, 1274–1277. (doi:10.1056/NEJMc1415318)
101. WHO Ebola Response Team. 2016 Ebola virus disease among male and female persons in West Africa. *N. Engl. J. Med.* **374**, 96–98. (doi:10.1056/NEJMc1510305)
102. Donnelly CA *et al*. 2003 Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet* **361**, 1761–1766. (doi:10.1016/S0140-6736(03)13410-1)
103. Anderson RM, Fraser C, Ghani AC, Donnelly CA, Riley S, Ferguson NM, Leung GM, Lam TH, Hedley AJ. 2004 Epidemiology, transmission dynamics and control of SARS: the 2002–2003 epidemic. *Phil. Trans. R. Soc. Lond. B* **359**, 1091–1105. (doi:10.1098/rstb.2004.1490)
104. Anderson RM, May RM. 1991 *Infectious diseases of humans*, vol. 1. Oxford, UK: Oxford University Press.
105. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. 1996 A statistical algorithm for the early detection of outbreaks of infectious disease. *J. R. Stat. Soc. Ser. A Stat. Soc.* **159**, 547–563. (doi:10.2307/2983331)
106. Park SW, Champredon D, Weitz J, Dushoff J. 2018 A practical generation interval-based approach to inferring the strength of epidemics from their speed. *bioRxiv*. 312397. (doi:10.1101/312397)
107. Keeling M, Rohani P. 2008 Modeling infectious diseases in humans and animals. *Clin. Infect. Dis.* **47**, 864–866. (doi:10.1086/591197)
108. McKinley T, Cook AR, Deardon R. 2009 Inference in epidemic models without likelihoods. *Int. J. Biostat.* **5**(1): Article 24. (doi:10.2202/1557-4679.1171)
109. Obadia T, Haneef R, Boëlle P-Y. 2012 The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Med. Inform. Decis. Mak.* **12**, 147. (doi:10.1186/1472-6947-12-147)
110. Chowell G, Viboud C, Simonsen L, Merler S, Vespignani A. 2017 Perspectives on model forecasts of the 2014–2015 Ebola epidemic in West Africa: lessons and the way forward. *BMC Med.* **15**, 42. (doi:10.1186/s12916-017-0811-y)
111. Held L, Meyer S, Bracher J. 2017 Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. *Stat. Med.* **36**, 3443–3460. (doi:10.1002/sim.7363)
112. Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. 2017 Assessing the performance of real-time epidemic forecasts. *bioRxiv*. 177451. (doi:10.1101/177451)
113. Kucharski AJ, Camacho A, Flasche S, Glover RE, Edmunds WJ, Funk S. 2015 Measuring the impact of Ebola control measures in Sierra Leone. *Proc. Natl Acad. Sci. USA* **112**, 14 366–14 371. (doi:10.1073/pnas.1508814112)
114. Buring JE. 1987 *Epidemiology in medicine*. Philadelphia, PA: Lippincott Williams & Wilkins.
115. Grandesso F *et al*. 2014 Risk factors for cholera transmission in Haiti during inter-peak periods: insights to improve current control strategies from two case-control studies. *Epidemiol. Infect.* **142**, 1625–1635. (doi:10.1017/S0950268813002562)
116. Gross M. 1976 Oswego County revisited. *Public Health Rep.* **91**, 168–170.
117. Buchholz U *et al*. 2011 German outbreak of *Escherichia coli* O104:H4 associated with sprouts. *N. Engl. J. Med.* **365**, 1763–1770. (doi:10.1056/NEJMoa1106482)
118. Ebola ça Suffit Ring Vaccination Trial Consortium. 2015 The ring vaccination trial: a novel cluster randomised controlled trial design to evaluate vaccine efficacy and effectiveness during outbreaks, with special reference to Ebola. *BMJ* **351**, h3740. (doi:10.1136/bmj.h3740)
119. Grais RF, Conlan AJK, Ferrari MJ, Djibo A, Le Menach A, Bjørnstad ON, Grenfell BT. 2008 Time is of the essence: exploring a measles outbreak response vaccination in Niamey, Niger. *J. R. Soc. Interface* **5**, 67–74. (doi:10.1098/rsif.2007.1038)
120. Harris SR *et al*. 2013 Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect. Dis.* **13**, 130–136. (doi:10.1016/S1473-3099(12)70268-2)
121. Gire SK *et al*. 2014 Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372. (doi:10.1126/science.1259657)
122. Cotten M *et al*. 2013 Transmission and evolution of the Middle East respiratory syndrome coronavirus in Saudi Arabia: a descriptive genomic study. *Lancet* **382**, 1993–2002. (doi:10.1016/S0140-6736(13)61887-5)
123. Robinson ER, Walker TM, Pallen MJ. 2013 Genomics and outbreak investigation: from sequence to consequence. *Genome Med.* **5**, 36. (doi:10.1186/gm440)
124. Hatherell H-A, Didelot X, Pollock SL, Tang P, Crisan A, Johnston JC, Colijn C, Gardy JL. 2016 Declaring a tuberculosis outbreak over with genomic epidemiology. *Microb. Genomics* **2**, e000060. (doi:10.1099/mgen.0.000060)
125. Dudas G *et al*. 2017 Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature* **544**, 309–315. (doi:10.1038/nature22040)
126. Faria NR *et al*. 2017 Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* **546**, 406–410. (doi:10.1038/nature22401)
127. Quick J *et al*. 2016 Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232. (doi:10.1038/nature16996)
128. Pallen MJ, Loman NJ, Penn CW. 2010 High-throughput sequencing and clinical microbiology: progress, opportunities and challenges. *Curr. Opin. Microbiol.* **13**, 625–631. (doi:10.1016/j.mib.2010.08.003)
129. Spratt BG, Hanage WP, Li B, Aanensen DM, Feil EJ. 2004 Displaying the relatedness among isolates of bacterial species—the eBURST approach. *FEMS Microbiol. Lett.* **241**, 129–134. (doi:10.1016/j.femsle.2004.11.015)
130. Enright MC, Robinson DA, Randle G, Feil EJ, Grundmann H, Spratt BG. 2002 The evolutionary history of methicillin-resistant *Staphylococcus aureus* (MRSA). *Proc. Natl Acad. Sci. USA* **99**, 7687–7692. (doi:10.1073/pnas.122108599)
131. King SJ, Leigh JA, Heath PJ, Luque I, Tarradas C, Dowson CG, Whatmore AM. 2002 Development of a multilocus sequence typing scheme for the pig pathogen *Streptococcus suis*: identification of virulent clones and potential capsular serotype

exchange. *J. Clin. Microbiol.* **40**, 3671–3680. (doi:10.1128/JCM.40.10.3671-3680.2002)

132. Urwin R, Maiden MCJ. 2003 Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol.* **11**, 479–487. (doi:10.1016/j.tim.2003.08.006)
133. Felsenstein J. 2004 *Inferring phylogenies*. Sunderland, MA: Sinauer Associates Sunderland.
134. Popescu A-A, Huber KT, Paradis E. 2012 ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536–1537. (doi:10.1093/bioinformatics/bts184)
135. Felsenstein J. 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376. (doi:10.1007/BF01734359)
136. Schliep KP. 2011 phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593. (doi:10.1093/bioinformatics/btq706)
137. Ronquist F, Huelsenbeck JP. 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574. (doi:10.1093/bioinformatics/btg180)
138. Grubaugh ND, Faria NR, Andersen KG, Pybus OG. 2018 Genomic insights into zika virus emergence and spread. *Cell* **172**, 1160–1162. (doi:10.1016/j.cell.2018.02.027)
139. Smith GJD *et al*. 2009 Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature* **459**, 1122–1125. (doi:10.1038/nature08182)
140. Siddle KJ *et al*. 2018 Genomic analysis of lassa virus during an increase in cases in Nigeria in 2018. *N. Engl. J. Med.* **379**, 1745–1753. (doi:10.1056/NEJMoa1804498)
141. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, Holmes EC. 2004 Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332. (doi:10.1126/science.1090727)
142. Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, King DP, Haydon DT. 2008 Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. R. Soc. B* **275**, 887–895. (doi:10.1098/rspb.2007.1442)
143. Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. 2012 Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proc. R. Soc. B* **279**, 444–450. (doi:10.1098/rspb.2011.0913)
144. Ypma RJF, van Ballegooijen WM, Wallinga J. 2013 Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* **195**, 1055–1062. (doi:10.1534/genetics.113.154856)
145. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. 2014 Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **10**, e1003457. (doi:10.1371/journal.pcbi.1003457)
146. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. 2017 Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput. Biol.* **13**, e1005495. (doi:10.1371/journal.pcbi.1005495)
147. Didelot X, Gardy J, Colijn C. 2014 Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol. Biol. Evol.* **31**, 1869–1879. (doi:10.1093/molbev/msu121)
148. De Maio N, Wu C-H, Wilson DJ. 2016 SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput. Biol.* **12**, e1005130. (doi:10.1371/journal.pcbi.1005130)
149. Springborn M, Chowell G, MacLachlan M, Fenichel EP. 2015 Accounting for behavioral responses during a flu epidemic using home television viewing. *BMC Infect. Dis.* **15**, 21. (doi:10.1186/s12879-014-0691-0)
150. R Core Team. 2018 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
151. RECON-R Epidemics Consortium. 2018 R epidemics consortium. See https://www.repidemicsconsortium.org/ (accessed on 26 September 2018).
152. 2018 RECON learn. See https://www.reconlearn.org (accessed on 26 September 2018).