PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15

Sebastian Funk[1,2]*, **Anton Camacho**[1,2,3], **Adam J. Kucharski**[1,2], **Rachel Lowe**[1,2,4], **Rosalind M. Eggo**[1,2], **W. John Edmunds**[1,2]

1 Centre for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom, **2** Department of Infectious Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, United Kingdom, **3** Epicentre, Paris, France, **4** Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain

* sebastian.funk@lshtm.ac.uk

Citation: Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ (2019) Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area region of Sierra Leone, 2014-15. PLoS Comput Biol 15(2): e1006785. https://doi.org/10.1371/journal.pcbi.1006785

Editor: Christian Althaus, University of Bern, SWITZERLAND

Received: April 25, 2018

Accepted: January 14, 2019

Published: February 11, 2019

Data Availability Statement: All data and code are contained in an R package available at https://doi.org/10.5281/zenodo.2547701.

Funding: This work was funded by the Research for Health in Humanitarian Crises (R2HC) Programme, managed by Research for Humanitarian Assistance (Grant 13165). SF, AJK and AC were supported by fellowships from the UK Medical Research Council (SF: MR/K021680/1, AC: MR/J01432X/1, AJK: MR/K021524/1). SF was

Abstract

Real-time forecasts based on mathematical models can inform critical decision-making during infectious disease outbreaks. Yet, epidemic forecasts are rarely evaluated during or after the event, and there is little guidance on the best metrics for assessment. Here, we propose an evaluation approach that disentangles different components of forecasting ability using metrics that separately assess the calibration, sharpness and bias of forecasts. This makes it possible to assess not just how close a forecast was to reality but also how well uncertainty has been quantified. We used this approach to analyse the performance of weekly forecasts we generated in real time for Western Area, Sierra Leone, during the 2013–16 Ebola epidemic in West Africa. We investigated a range of forecast model variants based on the model fits generated at the time with a semi-mechanistic model, and found that good probabilistic calibration was achievable at short time horizons of one or two weeks ahead but model predictions were increasingly unreliable at longer forecasting horizons. This suggests that forecasts may have been of good enough quality to inform decision making based on predictions a few weeks ahead of time but not longer, reflecting the high level of uncertainty in the processes driving the trajectory of the epidemic. Comparing forecasts based on the semi-mechanistic model to simpler null models showed that the best semi-mechanistic model variant performed better than the null models with respect to probabilistic calibration, and that this would have been identified from the earliest stages of the outbreak. As forecasts become a routine part of the toolkit in public health, standards for evaluation of performance will be important for assessing quality and improving credibility of mathematical models, and for elucidating difficulties and trade-offs when aiming to make the most useful and reliable forecasts.

supported by a Wellcome Trust Senior Research Fellowship in Basic Biomedical Science (210758/Z/18/Z). AJK was supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (206250/Z/17/Z). RL was supported by a Royal Society Dorothy Hodgkin fellowship. RME acknowledges funding from an HDR UK Innovation Fellowship (grant MR/S003975/1). WJE and RME acknowledge funding from the Innovative Medicines Initiative 2 (IMI2) Joint Undertaking under grant agreement EBOVAC1 (grant 115854). The IMI2 is supported by the European Union Horizon 2020 Research and Innovation Programme and the European Federation of Pharmaceutical Industries and Associations. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Author summary

During epidemics, reliable forecasts can help allocate resources effectively to combat the disease. Various types of mathematical models can be used to make such forecasts. In order to assess how good the forecasts are, they need to be compared to what really happened. Here, we describe different approaches to assessing how good forecasts were that we made with mathematical models during the 2013–16 West African Ebola epidemic, focusing on one particularly affected area of Sierra Leone. We found that, using the type of models we used, it was possible to reliably predict the epidemic for a maximum of one or two weeks ahead, but no longer. Comparing different versions of our model to simpler models, we further found that it would have been possible to determine the model that was most reliable at making forecasts from early on in the epidemic. This suggests that there is value in assessing forecasts, and that it should be possible to improve forecasts by checking how good they are during an ongoing epidemic.

Introduction

Forecasting the future trajectory of cases during an infectious disease outbreak can make an important contribution to public health and intervention planning. Infectious disease modellers are now routinely asked for predictions in real time during emerging outbreaks [1]. Forecasting targets can revolve around expected epidemic duration, size, or peak timing and incidence [2–5], geographical distribution of risk [6], or short-term trends in incidence [7, 8]. However, forecasts made during an outbreak are rarely investigated during or after the event for their accuracy, and only recently have forecasters begun to make results, code, models and data available for retrospective analysis.

The growing importance of infectious disease forecasts is epitomised by the growing number of so-called forecasting challenges. In these, researchers compete in making predictions for a given disease and a given time horizon. Such initiatives are difficult to set up during unexpected outbreaks, and are therefore usually conducted on diseases known to occur seasonally, such as dengue [7, 9, 10] and influenza [11]. The *Ebola Forecasting Challenge* was a notable exception, triggered by the 2013–16 West African Ebola epidemic and set up in June 2015. Since the epidemic had ended in most places at that time, the challenge was based on simulated data designed to mimic the behaviour of the true epidemic instead of real outbreak data. The main lessons learned were that 1) ensemble estimates outperformed all individual models, 2) more accurate data improved the accuracy of forecasts and 3) considering contextual information such as individual-level data and situation reports improved predictions [12].

In theory, infectious disease dynamics should be predictable within the timescale of a single outbreak [13]. In practice, however, providing accurate forecasts during emerging epidemics comes with particular challenges such as data quality issues and limited knowledge about the processes driving growth and decline in cases. In particular, uncertainty about human behavioural changes and public health interventions can preclude reliable long-term predictions [14, 15]. Yet, short-term forecasts with an horizon of a few generations of transmission (e.g., a few weeks in the case of Ebola), can yield important information on current and anticipated outbreak behaviour and, consequently, guide immediate decision making.

The most recent example of large-scale outbreak forecasting efforts was during the 2013–16 Ebola epidemic, which vastly exceeded the burden of all previous outbreaks with almost 30,000 reported cases resulting in over 10,000 deaths in the three most affected countries: Guinea, Liberia and Sierra Leone. During the epidemic, several research groups provided

forecasts or projections at different time points, either by generating scenarios believed plausible, or by fitting models to the available time series and projecting them forward to predict the future trajectory of the outbreak [16–26]. One forecast that gained particular attention during the epidemic was published in the summer of 2014, projecting that by early 2015 there might be 1.4 million cases [27]. This number was based on unmitigated growth in the absence of further intervention and proved a gross overestimate, yet it was later highlighted as a "call to arms" that served to trigger the international response that helped avoid the worst-case scenario [28]. While that was a particularly drastic prediction, most forecasts made during the epidemic were later found to have overestimated the expected number of cases, which provided a case for models that can generate sub-exponential growth trajectories [29, 30].

Traditionally, epidemic forecasts are assessed using aggregate metrics such as the mean absolute error (MAE) [12, 31, 32]. This, however, only assesses how close the most likely or average predicted outcome is to the true outcome. The ability to correctly forecast uncertainty, and to quantify confidence in a predicted event, is not assessed by such metrics. Appropriate quantification of uncertainty, especially of the likelihood and magnitude of worst case scenarios, is crucial in assessing potential control measures. Methods to assess probabilistic forecasts are now being used in other fields, but are not commonly applied in infectious disease epidemiology [33, 34].

We produced weekly sub-national real-time forecasts during the Ebola epidemic, starting on 28 November 2014. Plots of the forecasts were published on a dedicated web site and updated every time a new set of data were available [35]. They were generated using a model that has, in variations, been used to forecast bed demand during the epidemic in Sierra Leone [21] and the feasibility of vaccine trials later in the epidemic [36, 37]. During the epidemic, we provided sub-national forecasts for the three most affected countries (at the level of counties in Liberia, districts in Sierra Leone and prefectures in Guinea).

Here, we apply assessment metrics that elucidate different properties of forecasts, in particular their probabilistic calibration, sharpness and bias. Using these methods, we retrospectively assess the forecasts we generated for Western Area in Sierra Leone, an area that saw one of the greatest number of cases in the region and where our model informed bed capacity planning.

Materials and methods

Ethics statement

This study has been approved by the London School of Hygiene & Tropical Medicine Research Ethics Committee (reference number 8627).

Data sources

Numbers of suspected, probable and confirmed Ebola cases at sub-national levels were initially compiled from daily *Situation Reports* (or *SitReps*) provided in PDF format by Ministries of Health of the three affected countries during the epidemic [21]. Data were automatically extracted from tables included in the reports wherever possible and otherwise manually converted by hand to machine-readable format and aggregated into weeks. From 20 November 2014, the World Health Organization (WHO) provided tabulated data on the weekly number of confirmed and probable cases. These were compiled from the patient database, which was continuously cleaned and took into account reclassification of cases avoiding potential double-counting. However, the patient database was updated with substantial delay so that the number of reported cases would typically be underestimated in the weeks leading up to the date at which the forecast was made. Because of this, we used the SitRep data for the most recent

weeks until the latest week in which the WHO case counts either equalled or exceeded the SitRep counts. For all earlier times, the WHO data were used.

Transmission model

We used a semi-mechanistic stochastic model of Ebola transmission described previously [21, 38]. Briefly, the model was based on a Susceptible–Exposed–Infectious–Recovered (SEIR) model with fixed incubation period of 9.4 days [39], following an Erlang distribution with shape 2. The country-specific infectious period was determined by adding the average delay to hospitalisation to the average time from hospitalisation to death or discharge, weighted by the case-fatality rate. Cases were assumed to be reported with a stochastic time-varying delay. On any given day, this was given by a gamma distribution with mean equal to the country-specific average delay from onset to hospitalisation and standard deviation of 0.1 day. We allowed transmission to vary over time in order to capture behavioural changes in the community, public health interventions or other factors affecting transmission for which information was not available at the time. The time-varying transmission rate was modelled using a daily Gaussian random walk with fixed volatility (or standard deviation of the step size) which was estimated as part of the inference procedure (see below). We log-transformed the transmission rate to ensure it remained positive. The behaviour in time can be written as

$$d\log\beta_t = \sigma dW_t \quad (1)$$

where β_t is the time-varying transmission rate, W_t is the Wiener process and σ the volatility of the transmission rate. The basic reproduction number $R_{0,t}$ at any time was obtained by multiplying β_t with the average infectious period. In fitting the model to the time series of cases we extracted posterior predictive samples of trajectories, which we used to generate forecasts.

Model fitting

Each week, we fitted the model to the available case data leading up to the date of the forecast. Observations were assumed to follow a negative binomial distribution. Since the *ssm* software used to fit the model only implemented a discretised normal observation model, we used a normal approximation of the negative binomial for observations, potentially introducing a bias at small counts. Four parameters were estimated in the process: the initial basic reproduction number R_0 (uniform prior within (1, 5)), initial number of infectious people (uniform prior within (1, 400)), overdispersion of the (negative binomial) observation process (uniform prior within (0, 0.5)) and volatility of the time-varying transmission rate (uniform prior within (0, 0.5)). We confirmed from the posterior distributions of the parameters that these priors did not set any problematic bounds. Samples of the posterior distribution of parameters and state trajectories were extracted using particle Markov chain Monte Carlo [40] as implemented in the *ssm* library [41]. For each forecast, 50,000 samples were extracted and thinned to 5000.

Predictive model variants

We used the samples of the posterior distribution generated using the Monte Carlo sampler to produce predictive trajectories, using the final values of estimated state trajectories as initial values for the forecasts and simulating the model forward for up to 10 weeks. While all model fits were generated using the same model described above, we tested a range of different predictive model variants to assess the quality of ensuing predictions. We tested variants where trajectories were stochastic (with demographic stochasticity and a noisy reporting process), as well as ones where these sources of noise were removed for predictions. We further tested predictive model variants where the transmission rate continued to follow a random walk

(unbounded, on a log-scale), as well as ones where the transmission rate stayed fixed during the forecasting period. When the transmission rate remained fixed for prediction, we tested variants where we used the final value of the transmission rate and ones where this value was averaged over a number of weeks leading up to the final fitted point, to reduce the potential influence of the last time point, at which the transmission rate may not have been well identified. We tested variants where the predictive trajectory was based on the final values and start at the last time point, and ones where it started at the penultimate time point, which could, again, be expected to be better informed by the data. For each model and forecast horizon, we generated point-wise medians and credible intervals from the sample trajectories.

Null models

To assess the performance of the semi-mechanistic transmission model we compared it to three simpler null models: two representing the constituent parts of the semi-mechanistic model, and a non-mechanistic time series model. For the first null model, we used a *deterministic* model that only contained the mechanistic core of the semi-mechanistic model, that is a deterministic SEIR model with fixed transmission rate and parameters otherwise the same as in the model described before [21]:

$$\frac{dS}{dt} = -\frac{R_0}{\Delta}\frac{I_c + I_h}{N}S \tag{2}$$

$$\frac{dE_1}{dt} = -\frac{R_0}{\Delta}\frac{I_c + I_h}{N}S - 2\nu E_1 \tag{3}$$

$$\frac{dE_2}{dt} = 2\nu E_1 - 2\nu E_2 \tag{4}$$

$$\frac{dI_c}{dt} = 2\nu E_2 - \tau I_c \tag{5}$$

$$\frac{dI_h}{dt} = \tau I_c - \gamma I_h \tag{6}$$

$$\frac{dR}{dt} = \gamma I_h \tag{7}$$

$$\frac{dA}{dt} = \tau I_c \tag{8}$$

$$Y_t \sim \mathrm{NB}(A_t - A_{t-1}, \phi) \tag{9}$$

where Y_t are observations at times t, S is the number susceptible, E the number infected but not yet infectious (split into two compartments for Erlang-distributed permanence times with shape 2), I_c is the number infectious and not yet notified, I_h is the number infectious and notified, R is the number recovered or dead, A is an accumulator for incidence, R_0 is the basic reproduction number, $\Delta = 1/\tau + 1/\nu$ is the mean time from onset to outcome, $1/\nu$ is the mean incubation period, $1/\tau + 1/\gamma$ is the mean duration of infectiousness, $1/\tau$ is the mean time from onset to hospitalisation $1/\gamma$ the mean duration from notification to outcome and $\mathrm{NB}(\mu, \phi)$ is a negative binomial distribution with mean μ and overdispersion ϕ. All these parameters were informed by individual patient observations [39] except the overdispersion in reporting ϕ, and

the basic reproduction number R_0, which were inferred using Markov-chain Monte Carlo with the same priors as in the semi-mechanistic model.

For the second null model, we used an *unfocused* model where the weekly incidence Z itself was modelled using a stochastic volatility model (without drift), that is a daily Gaussian random walk, and forecasts generated assuming the weekly number of new cases was not going to change:

$$\mathrm{d}\log Z = \sigma \mathrm{dW} \tag{10}$$

$$Y_t \sim \mathrm{NB}(Z_t, \phi) \tag{11}$$

where Y are observations, σ is the intensity of the random walk and ϕ the overdispersion of reporting (both estimated using Markov-chain Monte Carlo) and dW is the Wiener process.

Lastly, we used a null model based on a non-mechanistic Bayesian autoregressive AR(1) time series model:

$$\alpha_{t+1} \sim \mathcal{N}(\phi\alpha_t, \sigma_\alpha) \tag{12}$$

$$Y_t^* \sim \mathcal{N}(\alpha_t, \sigma_{Y^*}) \tag{13}$$

$$Y_t = \max(0, [Y_t^*]) \tag{14}$$

where ϕ, σ_α and σ_{Y^*} were estimated using Markov-chain Monte Carlo, and [. . .] indicates rounding to the nearest integer. An alternative model with Poisson distributed observations was discarded as it yielded poorer predictive performance.

The deterministic and unfocused models were implemented in *libbi* [42] via the *RBi* [43] and *RBi.helpers* [44] *R* packages [45]. The Bayesian autoregressive time series model was implemented using the *bsts* package [46].

Metrics

The paradigm for assessing probabilistic forecasts is that they should maximise the sharpness of predictive distributions subject to calibration [47]. We therefore first assessed model calibration at a given forecasting horizon, before assessing their sharpness and other properties.

Calibration or reliability [48] of forecasts is the ability of a model to correctly identify its own uncertainty in making predictions. In a model with perfect calibration, the observed data at each time point look as if they came from the predictive probability distribution at that time. Equivalently, one can inspect the probability integral transform of the predictive distribution at time t [49],

$$u_t = F_t(x_t) \tag{15}$$

where x_t is the observed data point at time $t \in t_1, \ldots, t_n$, n being the number of forecasts, and F_t is the (continuous) predictive cumulative probability distribution at time t. If the true probability distribution of outcomes at time t is G_t then the forecasts F_t are said to be *ideal* if $F_t = G_t$ at all times t. In that case, the probabilities u_t are distributed uniformly.

In the case of discrete outcomes such as the incidence counts that were forecast here, the PIT is no longer uniform even when forecasts are ideal. In that case a randomised PIT can be used instead:

$$u_t = P_t(k_t) + v(P_t(k_t) - P_t(k_t - 1)) \tag{16}$$

where k_t is the observed count, $P_t(x)$ is the predictive cumulative probability of observing

incidence k at time t, $P_t(-1) = 0$ by definition and v is standard uniform and independent of k. If P_t is the true cumulative probability distribution, then u_t is standard uniform [50]. To assess calibration, we applied the Anderson-Darling test of uniformity to the probabilities u_t. The resulting p-value was a reflection of how compatible the forecasts were with the null hypothesis of uniformity of the PIT, or of the data coming from the predictive probability distribution. We calculated the mean p-value of 10 samples from the randomised PIT and found the corresponding Monte-Carlo error to be negligible (maximum standard deviation: $s_p = 0.003$). We considered that there was no evidence to suggest a forecasting model was miscalibrated if the p-value found was greater than a threshold of $p \geq 0.1$, some evidence that it was miscalibrated if $0.01 < p < 0.1$, and good evidence that it was miscalibrated if $p \leq 0.01$. In this context it should be noted, though, that uniformity of the (randomised) PIT is a necessary but not sufficient condition of calibration [47]. The p-values calculated here merely quantify our ability to reject a hypothesis of good calibration, but cannot guarantee that a forecast is calibrated. Because of this, other indicators of forecast quality must be considered when choosing a model for forecasts.

All of the following metrics are evaluated at every single data point. In order to compare the forecast quality of models, they were averaged across the time series.

Sharpness is the ability of the model to generate predictions within a narrow range of possible outcomes. It is a data-independent measure, that is, it is purely a feature of the forecasts themselves. To evaluate sharpness at time t, we used the normalised median absolute deviation about the median (MADN) of y

$$S_t(P_t) = \frac{1}{0.675}\text{median}(|y - \text{median}(y)|) \tag{17}$$

where y is a variable with CDF P_t, and division by 0.675 ensures that if the predictive distribution is normal this yields a value equivalent to the standard deviation. The MAD (i.e., the MADN without the normalising factor) is related to the interquartile range (and in the limit of infinite sample size takes twice its value), a common measure of sharpness [33], but is more robust to outliers [51]. The sharpest model would focus all forecasts on one point and have $S = 0$, whereas a completely blurred forecast would have $S \to \infty$. Again, we used Monte-Carlo samples from P_t to estimate sharpness.

We further assessed the *bias* of forecasts to test whether a model systematically over- or underpredicted. We defined the forecast bias at time t as

$$B_t(P_t, x_t) = 1 - (P_t(x_t) + P_t(x_t - 1)) \tag{18}$$

The least biased model would have exactly half of predictive probability mass not concentrated on the data itself below the data at time t and $B_t = 0$, whereas a completely biased model would yield either all predictive probability mass above ($B_t = 1$) or below ($B_t = -1$) the data.

We further evaluated forecasts using two *proper scoring rules*, that is scores which are minimised if the predictive distribution is the same as the one generating the data. These scores combine the assessment of calibration and sharpness for comparison of overall forecasting skill. The *Ranked Probability Score* (RPS) [52, 53] for count data is defined as [50]

$$\text{RPS}(P_t, x_t) = \sum_{k=0}^{\infty}(P_t(k) - \mathbb{1}(k \geq x_t))^2. \tag{19}$$

It reduces to the mean absolute error (MAE) if the forecast is deterministic and can therefore be seen as its probabilistic generalisation for discrete forecasts. A convenient equivalent

formulation for predictions generated from Monte-Carlo samples is [47, 50]

$$\mathrm{RPS}(P_t, x_t) = \mathbb{E}_{P_t}|X - x_t| - \frac{1}{2}\mathbb{E}_{P_t}|X - X'|, \tag{20}$$

where X and X' are independent realisations of a random variable with cumulative distribution P_t.

The *Dawid-Sebastiani score* (DSS) only considers the first two moments of the predictive distribution and is defined as [50]

$$\mathrm{DSS}(P_t, x_t) = \left(\frac{x_t - \mu_{P_t}}{\sigma_{P_t}}\right)^2 + 2\log\sigma_{P_t} \tag{21}$$

where μ_{P_t} and σ_{P_t} are the mean and standard deviation of the predictive probability distribution, respectively, estimated here using Monte-Carlo samples.

For comparison, we also evaluated forecasts using the *absolute error* (AE) of the median forecast, that is

$$\mathrm{AE}(P_t, x_t) = |\mathrm{median}_{P_t}(X) - x_t| \tag{22}$$

where X is a random variable with cumulative distribution P_t.

All scoring metrics used are implemented in the *R* package accompanying the paper. The *goftest* package was used for the Anderson-Darling test [54] and the *scoringRules* package for the RPS and DSS [55].

Results

The semi-mechanistic model used to generate real-time forecasts during the epidemic was able to reproduce the trajectories up to the date of each forecast, following the data closely by means of the smoothly varying transmission rate (Fig 1). The overall behaviour of the reproduction number (ignoring depletion of susceptibles which did not play a role at the population level given the relatively small proportion of the population infected) was one of a near-monotonic decline, from a median estimate of 2.9 (interquartile range (IQR) 2.1–4, 90% credible interval (CI) 1.2–6.9) in the first fitted week (beginning 10 August, 2014) to a median estimate of 1.3 (IQR 0.9–1.9, 90% CI 0.4–3.7) in early November, 0.9 (IQR 0.6–1.3, 90% CI 0.2–2.2) in

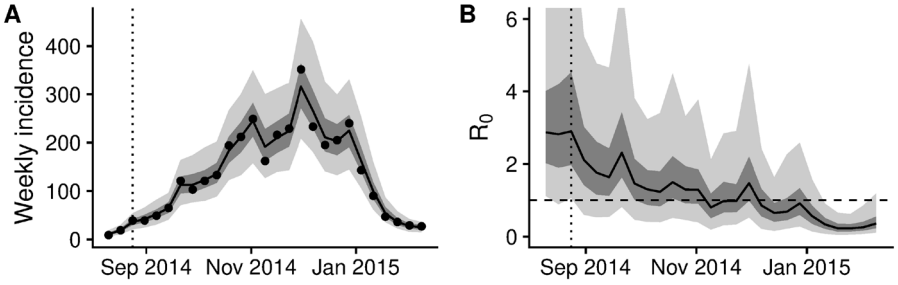

Fig 1. Final fit of the semi-mechanistic model to the Ebola outbreak in Western Area, Sierra Leone. (A) Final fit of the reported weekly incidence (black line and grey shading) to the data (black dots). (B) Corresponding dynamics of the reproduction number (ignoring depletion of susceptibles). Point-wise median state estimates are indicated by a solid line, interquartile ranges by dark shading, and 90% intervals by light shading. The threshold reproduction number ($R_0 = 1$), determining whether case numbers are expected to increase or decrease, is indicated by a dashed line. In both plots, a dotted vertical line indicates the date of the first forecast assessed in this manuscript (24 August 2014).

https://doi.org/10.1371/journal.pcbi.1006785.g001

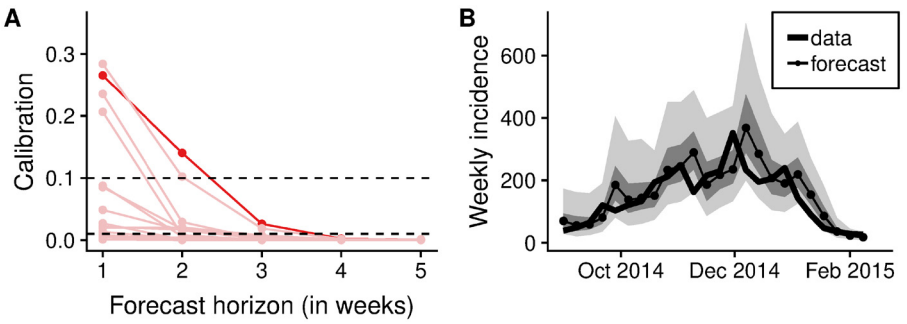

Fig 2. Calibration of forecasts from the semi-mechanistic model. (A) Calibration of predictive model variants (p-value of the Anderson-Darling test of uniformity) as a function of the forecast horizon. Shown in dark red is the best calibrated forecasting model variant (corresponding to the second row of Table 1). Other model variants are shown in light red. (B) Comparison of one-week forecasts of reported weekly incidence generated using the best semi-mechanistic model variant to the subsequently released data. The data are shown as a thick line, and forecasts as dots connected by a thin line. Dark shades of grey indicate the point-wise interquartile range, and lighter shades of grey the point-wise 90% credible interval.

https://doi.org/10.1371/journal.pcbi.1006785.g002

early December, 0.6 in early January (IQR 0.3–0.8, 90% CI 0.1–1.5) and 0.3 at the end of the epidemic in early February (IQR 0.2–0.4, 90% CI 0.1–0.9).

The epidemic lasted for a total of 27 weeks, with forecasts generated starting from week 3. For m-week ahead forecasts this yielded a sample size of $25 - m$ forecasts to assess calibration. Calibration of forecasts from the semi-mechanistic model were good for a maximum of one or two weeks, but deteriorated rapidly at longer forecasting horizons (Fig 2). The two semi-mechanistic forecast model variants with best calibration performance used deterministic dynamics starting at the last fitted data point (Table 1). Of these two, the forecast model that kept the

Table 1. Calibration of forecast model variants of the semi-mechanistic model. Calibration (p-value of the Anderson-Darling test of uniformity) of deterministic and stochastic predictive model variants starting either at the last data point or one week before, with varying (according to a Gaussian random walk) or fixed transmission rate either starting from the last value of the transmission rate or from an average over the last 2 or 3 weeks, at different forecast horizons up to 4 weeks. The p-values highlighted in bold reflect predictive models with no evidence of miscalibration. The second row corresponds to the highlighted model variant in Fig 2A.

Predictive model variant				Forecast horizon (weeks)			
Stochasticity	**Start**	**Transmission**	**Averaged**	**1**	**2**	**3**	**4**
deterministic	at last data point	varying	no	**0.28**	**0.1**	0.02	<0.01
deterministic	at last data point	fixed	no	**0.26**	**0.14**	0.03	<0.01
deterministic	at last data point	fixed	2 weeks	**0.24**	0.03	<0.01	<0.01
deterministic	at last data point	fixed	3 weeks	**0.21**	<0.01	<0.01	<0.01
deterministic	1 week before	varying	no	0.05	0.02	<0.01	<0.01
deterministic	1 week before	fixed	no	0.09	0.02	<0.01	<0.01
deterministic	1 week before	fixed	2 weeks	0.09	<0.01	<0.01	<0.01
deterministic	1 week before	fixed	3 weeks	0.03	<0.01	<0.01	<0.01
stochastic	at last data point	varying	no	0.02	0.02	<0.01	<0.01
stochastic	at last data point	fixed	no	0.02	0.02	<0.01	<0.01
stochastic	at last data point	fixed	2 weeks	0.01	<0.01	<0.01	<0.01
stochastic	at last data point	fixed	3 weeks	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	varying	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	fixed	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	fixed	2 weeks	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	fixed	3 weeks	<0.01	<0.01	<0.01	<0.01

https://doi.org/10.1371/journal.pcbi.1006785.t001

transmission rate constant from the value at the last data point performed slightly better across forecast horizons than one that continued to change the transmission rate following a random walk with volatility estimated from the time series. There was no evidence of miscalibration in both of the models with best calibration performance for two-week ahead forecasts, but increasing evidence of miscalibration for forecast horizons of three weeks or more. Calibration of all model variants was poor four weeks or more ahead, and all the stochastic model variants were miscalibrated for any forecast horizon, including the one we used to publish forecasts during the Ebola epidemic (stochastic, starting at the last data point, no averaging of the transmission rate, no projected volatility).

The calibration of the best semi-mechanistic forecast model variant (deterministic dynamics, transmission rate fixed and starting at the last data point) was better than that of any of the null models (Fig 3A and Table 2) for up to three weeks. While there was no evidence for miscalibration of the autoregressive null model for 1-week-ahead forecasts, there was good evidence of miscalibration for longer forecast horizons. There was some evidence of miscalibration of the unfocused null model, which assumes that the same number of cases will be reported in the weeks following the week during which the forecast was made, for 1 week ahead and good evidence of miscalibration beyond. Calibration of the deterministic null model was poor for all forecast horizons.

The semi-mechanistic and deterministic models showed a tendency to overestimate the predicted number of cases, while the autoregressive and null models tended to underestimate (Fig 3B and and Table 2). This bias increased with longer forecast horizons in all cases. The best calibrated semi-mechanistic model variant progressed from a 12% bias at 1 week ahead to 20% (2 weeks), 30% (3 weeks), 40% (4 weeks) and 44% (5 weeks) overestimation. At the same time, this model showed rapidly decreasing sharpness as the forecast horizon increased (Fig 3C and and Table 2). This is reflected in the proper scoring rules that combine calibration and sharpness, with smaller values indicating better forecasts (Fig 3D and 3E and and Table 2). At 1-week ahead, the mean RPS values of the autoregressive, unfocused and best semi-mechanistic

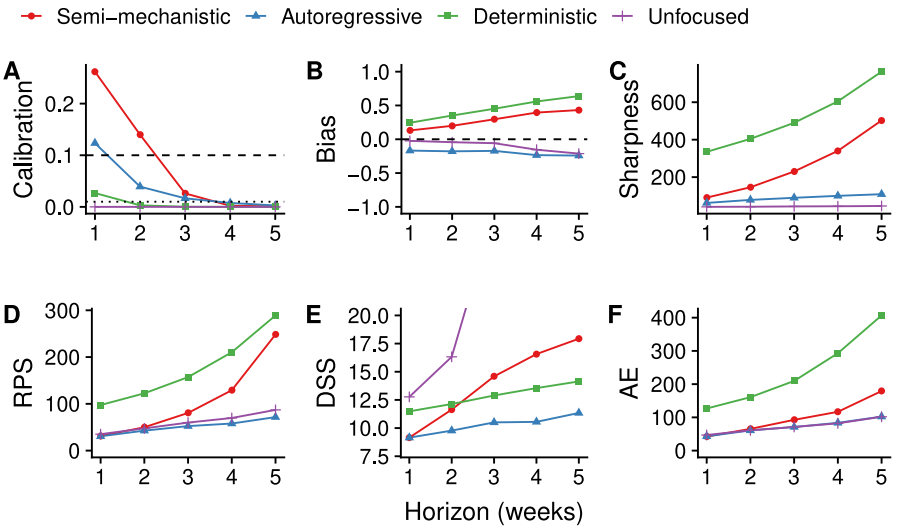

Fig 3. Forecasting metrics and scores of the best semi-mechanistic model variant compared to null models. Metrics shown are (A) calibration (p-value of Anderson-Darling test, greater values indicating better calibration, dashed lines at 0.1 and 0.01), (B) bias (less bias if closer to 0), (C) sharpness (MAD, sharper models having values closer to 0), (D) RPS (better values closer to 0), (E) DSS (better values closer to 0) and (F) AE (better values closer to 0), all as a function of the forecast horizon.

https://doi.org/10.1371/journal.pcbi.1006785.g003

Table 2. Forecasting metrics and scores of the best semi-mechanistic model variant compared to null models. The values shown are the same scores as in Fig 3, for forecasting horizons up to three weeks. The p-values for calibration highlighted in bold reflect predictive models with no evidence of miscalibration.

Model	Calibration	Sharpness	Bias	RPS	DSS	AE
1 week ahead						
Semi-mechanistic	**0.26**	91	0.13	31	9.2	42
Autoregressive	**0.1**	61	-0.17	31	9.1	43
Deterministic	0.03	340	0.24	97	11	130
Unfocused	<0.01	41	-0.024	35	13	47
2 weeks ahead						
Semi-mechanistic	**0.14**	150	0.2	50	12	65
Autoregressive	0.03	77	-0.18	43	9.9	60
Deterministic	<0.01	400	0.35	120	12	160
Unfocused	<0.01	42	-0.044	48	16	61
3 weeks ahead						
Semi-mechanistic	0.03	230	0.3	81	15	93
Autoregressive	0.02	90	-0.17	53	11	73
Deterministic	<0.01	490	0.45	160	13	210
Unfocused	<0.01	44	-0.058	60	29	71

https://doi.org/10.1371/journal.pcbi.1006785.t002

forecasting models were all around 30. At increasing forecasting horizon, the RPS of the semi-mechanistic model grew faster than the RPS of the autoregressive and unfocused null models. The DSS of the semi-mechanistic model, on the other hand, was very similar to the one of the autoregressive and better than that of the other null models at a forecast horizon of 1 week, with the autoregressive again performing best at increasing forecast horizons.

Focusing purely on the median forecast (and thus ignoring both calibration and sharpness), the absolute error (AE, Fig 3F and Table 2) was lowest (42) for the best semi-mechanistic model variant at 1-week ahead forecasts, although similar to the autoregressive and unfocused null models. With increasing forecasting horizon, the absolute error increased at a faster rate for the semi-mechanistic model than for the autoregressive and unfocused null models.

We lastly studied the calibration behaviour of the models over time; that is, using the data and forecasts available up to different time points during the epidemic (Fig 4). This shows that from very early on, not much changed in the ranking of the different semi-mechanistic model variants. Comparing the best semi-mechanistic forecasting model to the null models, again, for almost the whole duration of the epidemic calibration of the semi-mechanistic model was best for forecasts 1 or 2 weeks ahead.

Discussion

Probabilistic forecasts aim to quantify the inherent uncertainty in predicting the future. In the context of infectious disease outbreaks, they allow the forecaster to go beyond merely providing the most likely future scenario and quantify how likely that scenario is to occur compared to other possible scenarios. While correctly quantifying uncertainty in predicted trajectories has not commonly been the focus in infectious disease forecasting, it can have enormous practical implications for public health planning. Especially during acute outbreaks, decisions are often made based on so-called "worst-case scenarios" and their likelihood of occurring. The ability to adequately assess the magnitude as well as the probability of such scenarios requires accuracy at the tails of the predictive distribution, in other words good calibration of the forecasts.

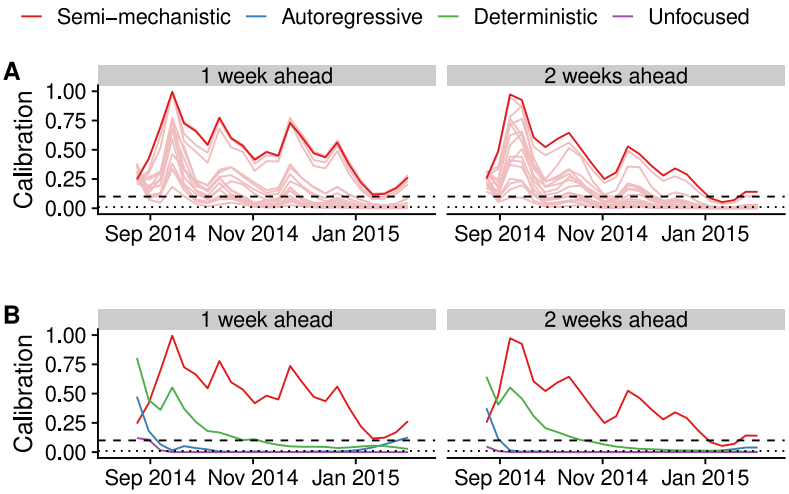

Fig 4. Calibration over time. Calibration scores of the forecast up to the time point shown on the x-axis. (A) Semi-mechanistic model variants, with the best model highlighted in dark red and other model variants are shown in light red. (B) Best semi-mechanistic model and null models. In both cases, 1-week (left) and 2-week (right) calibration (p-value of Anderson-Darling test) are shown.

https://doi.org/10.1371/journal.pcbi.1006785.g004

More generally, probabilistic forecasts need to be assessed using metrics that go beyond the simple difference between the central forecast and what really happened. Applying a suite of assessment methods to the forecasts we produced for Western Area, Sierra Leone, we found that probabilistic calibration of semi-mechanistic model variants varied, with the best ones showing good calibration for up to 2-3 weeks ahead, but performance deteriorated rapidly as the forecasting horizon increased. This reflects our lack of knowledge about the underlying processes shaping the epidemic at the time, from public health interventions by numerous national and international agencies to changes in individual and community behaviour. During the epidemic, we only published forecasts up to 3 weeks ahead, as longer forecasting horizons were not considered appropriate.

Our forecasts suffered from bias that worsened as the forecasting horizon expanded. Generally, the forecasts tended to overestimate the number of cases to be expected in the following weeks, as did most other forecasts generated during the outbreak [29]. This is in line with previous findings where our model was applied to predict simulated data of a hypothetical Ebola outbreak [38]. Log-transforming the transmission rate in order to ensure positivity skewed the underlying distribution and made very high values possible. Moreover, we did not model a trend in the transmission rate, whereas in reality transmission decreased over the course of the epidemic, probably due to a combination of factors ranging from better provision of isolation beds to increasing awareness of the outbreak and subsequent behavioural changes. While our model captured changes in the transmission rate in model fits, it did not forecast any trends such as the observed decrease over time. Capturing such trends in the attempt to identify underlying causes would be an important future improvement of real-time infectious disease models used for forecasting.

There are trade-offs between achieving good outcomes for the different forecast metrics we used. Deciding whether the best forecast is the best calibrated, the sharpest or the least biased, or some compromise between the three, is not a straightforward task. Our assessment of forecasts using separate metrics for probabilistic calibration, sharpness and bias highlights the underlying trade-offs. While the best calibrated semi-mechanistic model variant showed better calibration performance than the null models, this came at the expense of a decrease in the

sharpness of forecasts. Comparing the models using the RPS alone, the semi-mechanistic model of best calibration performance would not necessarily have been chosen. Following the paradigm of maximising sharpness subject to calibration, we therefore recommend to treat probabilistic calibration as a prerequisite to the use of forecasts, in line with what has recently been suggested for post-processing of forecasts [56]. Probabilistic calibration is essential for making meaningful probabilistic statements (such as the chances of seeing the number of cases exceed a set threshold in the upcoming weeks) that enable realistic assessments of resource demand, the possible future course of the epidemic including worst-case scenarios, as well as the potential impact of public health measures. Beyond the formal test for uniformity of the PIT applied here, alternative ways of assessing calibration can be used [47, 57]. Once a subset of models has been selected in an attempt to discard miscalibrated models, other criteria such as the RPS or DSS can be used to select the best model for forecasts, or to generate weights for ensemble forecasts combining several models. Such ensemble forecasts have become a standard in weather forecasting [58] and have more recently shown promise for infectious disease forecasts [12, 59, 60].

Other models may have performed better than the ones presented here. Because we did not have access to data that would have allowed us to assess the importance of different transmission routes (burials, hospitals and the community) we relied on a relatively simple, flexible model. The deterministic SEIR model we used as a null model performed poorly on all forecasting scores, and failed to capture the downturn of the epidemic in Western Area. On the other hand, a well-calibrated mechanistic model that accounts for all relevant dynamic factors and external influences could, in principle, have been used to predict the behaviour of the epidemic reliably and precisely. Yet, lack of detailed data on transmission routes and risk factors precluded the parameterisation of such a model and are likely to do so again in future epidemics in resource-poor settings. Future work in this area will need to determine the main sources of forecasting error, whether structural, observational or parametric, as well as strategies to reduce such errors [32].

In practice, there might be considerations beyond performance when choosing a model for forecasting. Our model combined a mechanistic core (the SEIR model) with non-mechanistic variable elements. By using a flexible non-parametric form of the time-varying transmission rate, the model provided a good fit to the case series despite a high levels of uncertainty about the underlying process. Having a model with a mechanistic core came with the advantage of enabling the assessment of interventions just as with a traditional mechanistic model. For example, the impact of a vaccine could be modelled by moving individuals from the susceptible into the recovered compartment [36, 37]. At the same time, the model was flexible enough to visually fit a wide variety of time series, and this flexibility might mask underlying misspecifications. Whenever possible, the guiding principle in assessing real-time models and predictions for public health should be the quality of the recommended decisions based on the model results [61].

Epidemic forecasts played a prominent role in the response to and public awareness of the Ebola epidemic [28]. Forecasts have been used for vaccine trial planning against Zika virus [62] and will be called upon again to inform the response to the next emerging epidemic or pandemic threat. Recent advances in computational and statistical methods now make it possible to fit models in near-real time, as demonstrated by our weekly forecasts [35]. Such repeated forecasts are a prerequisite for the use of metrics that assess not only how close the predictions were to reality, but also how well uncertainty in the predictions has been quantified. An agreement on standards of forecast assessment is urgently needed in infectious disease epidemiology, and retrospective or even real-time assessment should become standard for epidemic forecasts to prove accuracy and improve end-user trust. The metrics we have used here or

variations thereof could become measures of forecasting performance that are routinely used to evaluate and improve forecasts during epidemics.

For forecast assessment to happen in practice, evaluation strategies must be planned before the forecasts are generated. In order for such evaluation to be performed retrospectively, all forecasts as well as the data, code and models they were based on should be made public at the time, or at least preserved and decisions recorded for later analysis. We published weekly updated aggregate graphs and numbers during the Ebola epidemic, yet for full transparency it would have been preferable to allow individuals to download raw forecast data for further analysis.

If forecasts are not only produced but also evaluated in real time, this can give valuable insights into strengths, limitations, and reasonable time horizons. In our case, by tracking the performance of our forecasts, we would have noticed the poor calibration of the model variant chosen for the forecasts presented to the public, and instead selected better calibrated variants. At the same time, we did not store the predictive distribution samples for any area apart from Western Area in order to better use available storage space, and because we did not deem such storage valuable at the time. This has precluded a broader investigation of the performance of our forecasts.

Research into modelling and forecasting methodology and predictive performance at times during which there is no public health emergency should be part of pandemic preparedness activities. To facilitate this, outbreak data must be made available openly and rapidly. Where available, combination of multiple sources, such as epidemiological and genetic data, could increase predictive power. It is only on the basis of systematic and careful assessment of forecast performance during and after the event that predictive ability of computational models can be improved and lessons be learned to maximise their utility in future epidemics.

Author Contributions

Formal analysis: Sebastian Funk, Anton Camacho.

Methodology: Sebastian Funk, Anton Camacho, Adam J. Kucharski, Rachel Lowe, Rosalind M. Eggo, W. John Edmunds.

Writing – original draft: Sebastian Funk.

Writing – review & editing: Sebastian Funk, Anton Camacho, Adam J. Kucharski, Rachel Lowe, Rosalind M. Eggo, W. John Edmunds.

References

1. Heesterbeek H, Anderson RM, Andreasen V, Bansal S, Angelis DD, Dye C, et al. Modeling Infectious Disease Dynamics in the Complex Landscape of Global Health. Science. 2015; 347(6227):aaa4339–aaa4339. https://doi.org/10.1126/science.aaa4339 PMID: 25766240
2. Goldstein E, Cobey S, Takahashi S, Miller JC, Lipsitch M. Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: a statistical method. PLoS Med. 2011; 8(7):e1001051. https://doi.org/10.1371/journal.pmed.1001051 PMID: 21750666
3. Nsoesie E, Mararthe M, Brownstein J. Forecasting peaks of seasonal influenza epidemics. PLoS currents. 2013; 5. https://doi.org/10.1371/currents.outbreaks.bb1e879a23137022ea79a8c508b030bc PMID: 23873050
4. Yang W, Cowling BJ, Lau EH, Shaman J. Forecasting influenza epidemics in Hong Kong. PLoS Comput Biol. 2015; 11(7):e1004383. https://doi.org/10.1371/journal.pcbi.1004383 PMID: 26226185
5. Dawson PM, Werkman M, Brooks-Pollock E, Tildesley MJ. Epidemic predictions in an imperfect world: modelling disease spread with partial data. Proc R Soc B. 2015; 282(1808):20150205. https://doi.org/10.1098/rspb.2015.0205 PMID: 25948687

6. Lowe R, Barcellos C, Coelho CA, Bailey TC, Coelho GE, Graham R, et al. Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts. The Lancet infectious diseases. 2014; 14(7):619–626. https://doi.org/10.1016/S1473-3099(14)70781-9 PMID: 24841859
7. Johansson MA, Reich NG, Hota A, Brownstein JS, Santillana M. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. Sci Rep. 2016; 6:33707. https://doi.org/10.1038/srep33707 PMID: 27665707
8. Liu F, Porco TC, Amza A, Kadri B, Nassirou B, West SK, et al. Short-term forecasting of the prevalence of trachoma: expert opinion, statistical regression, versus transmission models. PLoS neglected tropical diseases. 2015; 9(8):e0004000. https://doi.org/10.1371/journal.pntd.0004000 PMID: 26302380
9. National Oceanic and Atmospheric Administration. Dengue Forecasting; 2017. Available from: http://dengueforecasting.noaa.gov/.
10. Centres for Disease Control and Prevention. Epidemic Prediction Initiative; 2017. Available from: https://predict.phiresearchlab.org/legacy/dengue/index.html.
11. Biggerstaff M, Alper D, Dredze M, Fox S, Fung ICH, Hickmann KS, et al. Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge. BMC Infect Diseases. 2016; 16(1):357. https://doi.org/10.1186/s12879-016-1669-x
12. Viboud C, Sun K, Gaffey R, Ajelli M, Fumanelli L, Merler S, et al. The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt. Epidemics. 2018; 22:13–21. https://doi.org/10.1016/j.epidem.2017.08.002 PMID: 28958414
13. Scarpino SV, Petri G. On the predictability of infectious disease outbreaks. arXiv;1703.07317v4.
14. Moran KR, Fairchild G, Generous N, Hickmann K, Osthus D, Priedhorsky R, et al. Epidemic forecasting is messier than weather forecasting: The role of human behavior and internet data streams in epidemic forecast. J Infect Dis. 2016; 214(suppl_4):S404–S408. https://doi.org/10.1093/infdis/jiw375 PMID: 28830111
15. Funk S, Ciglenecki I, Tiffany A, Gignoux E, Camacho A, Eggo RM, et al. The impact of control strategies and behavioural changes on the elimination of Ebola from Lofa County, Liberia. Phil Trans Roy Soc B. 2017; 372:20160302. https://doi.org/10.1098/rstb.2016.0302
16. Fisman D, Khoo E, Tuite A. Early epidemic dynamics of the west african 2014 ebola outbreak: estimates derived with a simple two-parameter model. PLOS Curr Outbreaks. 2014. https://doi.org/10.1371/currents.outbreaks.89c0d3783f36958d96ebbae97348d571
17. Lewnard JA, Ndeffo Mbah ML, Alfaro-Murillo JA, Altice FL, Bawo L, Nyenswah TG, et al. Dynamics and control of Ebola virus transmission in Montserrado, Liberia: a mathematical modelling analysis. Lancet Infect Dis. 2014; 14(12):1189–1195. https://doi.org/10.1016/S1473-3099(14)70995-8 PMID: 25455986
18. Nishiura H, Chowell G. Early transmission dynamics of Ebola virus disease (EVD), West Africa, March to August 2014. Euro Surveill. 2014; 19:20894. https://doi.org/10.2807/1560-7917.ES2014.19.36.20894 PMID: 25232919
19. Rivers CM, Lofgren ET, Marathe M, Eubank S, Lewis BL. Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia. PLOS Curr Outbreaks. 2014. https://doi.org/10.1371/currents.outbreaks.4d41fe5d6c05e9df30ddce33c66d084c
20. Towers S, Patterson-Lomba O, Castillo-Chavez C. Temporal variations in the effective reproduction number of the 2014 West Africa Ebola outbreak. PLOS Curr Outbreaks. 2014. https://doi.org/10.1371/currents.outbreaks.9e4c4294ec8ce1adad283172b16bc908
21. Camacho A, Kucharski A, Aki-Sawyerr Y, White MA, Flasche S, Baguelin M, et al. Temporal Changes in Ebola Transmission in Sierra Leone and Implications for Control Requirements: a Real-Time Modelling Study. PLOS Curr Outbreaks. 2015. https://doi.org/10.1371/currents.outbreaks.406ae55e83ec0b5193e30856b9235ed2
22. Dong F, Xu D, Wang Z, Dong M. Evaluation of ebola spreading in west africa and decision of optimal medicine delivery strategies based on mathematical models. Infect Genet Evol. 2015; 36:35–40. https://doi.org/10.1016/j.meegid.2015.09.003 PMID: 26343852
23. Drake JM, Kaul RB, Alexander LW, O'Regan SM, Kramer AM, Pulliam JT, et al. Ebola cases and health system demand in Liberia. PLoS Biol. 2015; 13(1):e1002056. https://doi.org/10.1371/journal.pbio.1002056 PMID: 25585384
24. Merler S, Ajelli M, Fumanelli L, Gomes MFC, y Piontti AP, Rossi L, et al. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis. Lancet Infect Dis. 2015; 15(2):204–211. https://doi.org/10.1016/S1473-3099(14)71074-6 PMID: 25575618
25. Siettos C, Anastassopoulou C, Russo L, Grigoras C, Mylonakis E. Modeling the 2014 ebola virus epidemic–agent-based simulations, temporal analysis and future predictions for liberia and

sierra leone. PLOS Curr Outbreaks. 2015. https://doi.org/10.1371/currents.outbreaks.8d5984114855fc425e699e1a18cdc6c9

26. White RA, MacDonald E, De Blasio BF, Nygård K, Vold L, Røttingen JA. Projected treatment capacity needs in Sierra Leone. PLOS Curr Outbreaks. 2015.
27. Meltzer MI, Atkins CY, Santibanez S, Knust B, Petersen BW, Ervin ED, et al. Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014-2015. MMWR Surveill Summ. 2014; 63 Suppl 3:1–14.
28. Frieden TR, Damon IK. Ebola in West Africa—CDC's role in epidemic detection, control, and prevention. Emerg Infect Dis. 2015; 21(11):1897. https://doi.org/10.3201/eid2111.150949 PMID: 26484940
29. Chretien JP, Riley S, George DB. Mathematical modeling of the West Africa Ebola epidemic. eLife. 2015; 4:e09186. https://doi.org/10.7554/eLife.09186 PMID: 26646185
30. Chowell G, Viboud C, Simonsen L, Merler S, Vespignani A. Perspectives on model forecasts of the 2014–2015 Ebola epidemic in West Africa: lessons and the way forward. BMC Med. 2017; 15(1):42. https://doi.org/10.1186/s12916-017-0811-y PMID: 28245814
31. Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling. 2017; 2(3):379–398. https://doi.org/10.1016/j.idm.2017.08.001 PMID: 29250607
32. Pei S, Shaman J. Counteracting structural errors in ensemble forecast of influenza outbreaks. Nature Communications. 2017; 8(1). https://doi.org/10.1038/s41467-017-01033-1
33. Gneiting T, Katzfuss M. Probabilistic Forecasting. Annu Rev Stat Appl. 2014; 1(1):125–151. https://doi.org/10.1146/annurev-statistics-062713-085831
34. Held L, Meyer S, Bracher J. Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture. Stat Med. 2017; 36(22):3443–3460. https://doi.org/10.1002/sim.7363 PMID: 28656694
35. Cente for the Mathematical Modelling of Infectious Diseases. Visualisation and projections of the Ebola outbreak in West Africa; 2015. Available from: http://ntncmch.github.io/ebola/.
36. Camacho A, Eggo RM, Funk S, Watson CH, Kucharski AJ, Edmunds WJ. Estimating the probability of demonstrating vaccine efficacy in the declining Ebola epidemic: a Bayesian modelling approach. BMJ Open. 2015; 5(12):e009346. https://doi.org/10.1136/bmjopen-2015-009346 PMID: 26671958
37. Camacho A, Eggo R, Goeyvaerts N, Vandebosch A, Mogg R, Funk S, et al. Real-time dynamic modelling for the design of a cluster-randomized phase 3 Ebola vaccine trial in Sierra Leone. Vaccine. 2017;in press. https://doi.org/10.1016/j.vaccine.2016.12.019
38. Funk S, Camacho A, Kucharski AJ, Eggo RM, Edmunds WJ. Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model. Epidemics. 2018; 22:56–61. https://doi.org/10.1016/j.epidem.2016.11.003 PMID: 28038870
39. WHO Ebola Response Team. Ebola Virus Disease in West Africa—The First 9 Months of the Epidemic and Forward Projections. N Engl J Med. 2014.
40. Andrieu C, Doucet A, Holenstein R. Particle Markov chain Monte Carlo methods. J R Stat Soc B. 2010; 72, pt. 3:269–342. https://doi.org/10.1111/j.1467-9868.2009.00736.x
41. Dureau J, Ballesteros S, Bogich T. SSM: Inference for time series analysis with State Space Models. arXiv. 2013;1307.5626.
42. Murray L. Bayesian State-Space Modelling on High-Performance Hardware Using LibBi. Journal of Statistical Software, Articles. 2015; 67(10):1–36.
43. Jacob PE, Funk S. RBi: R interface to LibBi; 2019. Available from: https://github.com/libbi/RBi.
44. Funk S. rbi.helpers: rbi helper functions; 2019. Available from: https://github.com/sbfnk/RBi.helpers.
45. R Core Team. R: A Language and Environment for Statistical Computing; 2018. Available from: https://www.R-project.org/.
46. Scott SL. bsts: Bayesian Structural Time Series; 2018. Available from: https://CRAN.R-project.org/package=bsts.
47. Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J R Stat Roc B. 2007; 69(2):243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x
48. Friederichs P, Thorarinsdottir TL. Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. Environmetrics. 2012; 23(7):579–594. https://doi.org/10.1002/env.2176
49. Dawid AP. Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach. J R Stat Soc A. 1984; 147(2):278. https://doi.org/10.2307/2981683
50. Czado C, Gneiting T, Held L. Predictive model assessment for count data. Biometrics. 2009; 65(4):1254–1261. https://doi.org/10.1111/j.1541-0420.2009.01191.x PMID: 19432783

51. Maronna RA, Martin RD, Yohai VJ, Salibián-Barrera M. Robust Statistics: Theory and Methods (with R). Wiley Series in Probability and Statistics. Wiley; 2018.

52. Epstein ES. A scoring system for probability forecasts of ranked categories. J Appl Meteorol. 1969; 8(6):985–987. https://doi.org/10.1175/1520-0450(1969)008%3C0985:ASSFPF%3E2.0.CO;2

53. Murphy AH. On the "ranked probability score". J Appl Meteorol. 1969; 8(6):988–989. https://doi.org/10.1175/1520-0450(1969)008%3C0988:OTPS%3E2.0.CO;2

54. Faraway J, Marsaglia G, Marsaglia J, Baddeley A. goftest: Classical Goodness-of-Fit Tests for Univariate Distributions; 2017. Available from: https://CRAN.R-project.org/package=goftest.

55. Jordan A, Krüger F, Lerch S. Evaluating probabilistic forecasts with scoringRules. arXiv;1709.04743v2.

56. Wilks DS. Enforcing calibration in ensemble postprocessing. Q J Roy Meteor Soc. 2018; 144(710):76–84. https://doi.org/10.1002/qj.3185

57. Held L, Rußbach K, Balabdaoui F. A Score Regression Approach to Assess Calibration of Continuous Probabilistic Predictions. Biometrics. 2010; 66(4):1295–1305. https://doi.org/10.1111/j.1541-0420.2010.01406.x PMID: 20353460

58. Gneiting T, Raftery AE. Weather forecasting with ensemble methods. Science. 2005; 310(5746):248–249. https://doi.org/10.1126/science.1115255 PMID: 16224011

59. Yamana TK, Kandula S, Shaman J. Superensemble forecasts of dengue outbreaks. J R Soc Interface. 2016; 13(123):20160410. https://doi.org/10.1098/rsif.2016.0410 PMID: 27733698

60. Yamana TK, Kandula S, Shaman J. Individual versus superensemble forecasts of seasonal influenza outbreaks in the United States. PLOS Comput Biol. 2017; 13(11):e1005801. https://doi.org/10.1371/journal.pcbi.1005801 PMID: 29107987

61. Probert WJM, Jewell CP, Werkman M, Fonnesbeck CJ, Goto Y, Runge MC, et al. Real-time decision-making during emergency disease outbreaks. PLOS Comput Biology. 2018; 14(7):e1006202. https://doi.org/10.1371/journal.pcbi.1006202

62. World Health Organization. Efficacy trials of ZIKV Vaccines: endpoints, trial design, site selection. WHO Workshop Meeting Report; 2017.