Assessing the performance of real-time epidemic forecasts: A case study of the 2013–16 Ebola epidemic

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] Center for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom
[2] 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
[*] Corresponding author. Email: sebastian.funk@lshtm.ac.uk

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 what the best metrics for assessment are. Here, we propose to disentangle different components of forecasting ability by using metrics that separately assess the calibration, sharpness and unbiasedness of forecasts. We used this approach to analyse the performance of weekly forecasts generated in real time in Western Area, Sierra Leone, during the 2013–16 Ebola epidemic in West Africa. We found that probabilistic calibration was good at short time horizons but deteriorated for long-term forecasts. This suggests that forecasts provided usable performance only a few weeks ahead of time, reflecting the high level of uncertainty in the processes driving the trajectory of the epidemic. Comparing the semi-mechanistic model we used during the epidemic to simpler null models showed that the our model performed better 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.

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 (Heesterbeek et al., 2015). Forecasting targets usually revolve around expected epidemic duration, size, or peak timing and incidence (Goldstein et al., 2011; Nsoesie et al., 2013; Yang et al., 2015; Dawson et al., 2015), geographical distribution of risk (Lowe et al., 2014), or short-term trends in incidence (Johansson et al., 2016; Liu et al., 2015). Despite the increase in activity, however, forecasts made during an outbreak is rarely investigated during or after the event for their accuracy.

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 (Johansson et al., 2016; National Oceanic and Atmospheric Administration, 2017; Centres for Disease Prevention and Control, 2017) and influenza (Biggerstaff et al., 2016). 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 (Viboud et al., 2017).

Providing accurate forecasts during emerging epidemics comes with particular challenges as uncertainties about the processes driving growth and decline in cases, in particular human behavioural changes and public health interventions, can preclude reliable long-term predictions (Moran et al., 2016; Funk et al., 2017b). Short-term forecasts with an horizon of a few generations of transmission (e.g., a few weeks in the case of Ebola), on the other hand, 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 of the disease, 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 out-

break (Fisman et al., 2014; Lewnard et al., 2014; Nishiura and Chowell, 2014; Rivers et al., 2014; Towers et al., 2014; Camacho et al., 2015b; Dong et al., 2015; Drake et al., 2015; Merler et al., 2015; Siettos et al., 2015; White et al., 2015). (Chretien et al., 2015; Chowell et al., 2017). One forecast that gained attention during the epidemic was published in the summer of 2014, projecting that by early 2015 there might be 1.4 million cases (Meltzer et al., 2014). While this number was based on unmitigated growth in the absence of further intervention and proved a gross overestimate, it was later highlighted as a "call to arms" that served to trigger the international response that helped avoid the worst-case scenario (Frieden and Damon, 2015).

Traditionally, epidemic forecasts are assessed using aggregate metrics such as the mean absolute error (MAE, Chowell, 2017; Pei and Shaman, 2017; Viboud et al., 2017). These, however, often only assess 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 (Gneiting and Katzfuss, 2014; Held et al., 2017). It is worth noting that good predictive ability need not coincide with good fit, as statistical evidence may not translate into forecast capability because of model uncertainty and noisy, incomplete data.

We produced weekly sub-national real-time forecasts during the Ebola epidemic, starting on 28 November 2014. These were published on a dedicated web site and updated every time a new set of data were available (Center for the Mathematical Modelling of Infectious Diseases, 2015). They were generated using a model that has, in variations, been used to forecast bed demand during the epidemic in Sierra Leone (Camacho et al., 2015b) and the feasibility of vaccine trials later in the epidemic (Camacho et al., 2015a; Camacho et al., 2017). During the epidemic, we provided sub-national forecasts for 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

Data sources

Numbers of suspected, probable and confirmed 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 (Camacho et al., 2015b). 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 of the forecast. 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 (Camacho et al., 2015b; Funk et al., 2017a). Briefly, the model was based on a Susceptible-Exposed-Infectious-Recovered (SEIR) model with fixed incubation period of 9.4 days (WHO Ebola Response Team, 2014), 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 be able 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). To ensure the transmission rate remained positive, we log-transformed it, so that its behaviour in time

can be written as

$$d \log \beta_t = \sigma dW_t \tag{1}$$

where β_t is the time-varying transmission rate, W_t is the Wiener process and σ the volatility of the transmission rate. 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, approximated as a discretised normal distribution for numerical convenience. Four parameters were estimated in the process: the 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 (Andrieu et al., 2010) as implemented in the *ssm* library (Dureau et al., 2013). 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 a range of 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. Where 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 would be averaged over

a number of weeks leading up to the final fitted point, to reduce the potential influence of the last time point, where the transmission rate may not have been well identified. We tested variants where the predictive trajectory would be based on the final values and start at the last time point, and ones where they would start 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 simpler null models: two representing the constituent parts of the semi-mechanistic model, and a non-mechanistic time series model. As first null model, we used a *deterministic* model that only contained the mechanistic core of the semi-mechanistic model with a fixed transmission rate. As second null model, we used an *unfocused* model where the number of cases 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. Lastly, we used a null model based on a non-mechanistic Bayesian *autoregressive* linear model. The deterministic and models were implemented in *libbi* (Murray, 2015) via the *RBi* (Jacob and Funk, 2017) and *RBi.helpers* (Funk, 2016) *R* packages (R Core Team, 2017). The autoregressive model was implemented using the *bsts* package (Scott, 2017).

Metrics

The paradigm for assessing probabilistic forecasts is that they should maximise the sharpness of predictive distributions subject to calibration (Gneiting et al., 2007). We therefore first assessed whether models were calibrated at a given forecasting horizon, before assessing their sharpness and other properties.

Calibration or reliability (Friederichs and Thorarinsdottir, 2012) of forecasts is the ability of a model to correctly identify its own uncertainty in making predictions. In a perfectly calibrated model, the 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 (Dawid, 1984),

$$u_t = F_t(x_t) \tag{2}$$

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 (CDF) 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.

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 considered a model to be calibrated if the p-value found was greater than a threshold of $p \geq 0.1$, possibly calibrated if $0.01 < p < 0.1$, and uncalibrated if $p \leq 0.01$.

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 median absolute deviation about the median (MADM) of y

$$S_t(F_t) = m\left(|y - m(y)|\right) \tag{3}$$

where y is a variable distributed according to F_t, and $m(y)$ is the median of y. 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 X from F_t to estimate sharpness.

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

$$B_t(F_t, x_t) = 2\left(\int_{-\infty}^{\infty} F_t(y) H(y - x_t) dy - 0.5\right) \tag{4}$$

where $H(x)$ is the Heaviside step function with the half-maximum convention $H(0) = 1/2$. This metric is equivalent to

$$B_t(F_t, x_t) = 2\left(E_{F_t}\left[H(X - x_t)\right] - 0.5\right) \tag{5}$$

which can be estimated using a finite number of samples, such as the Monte-Carlo samples generated in our inference procedure. Here, x_t are the observed data points, E_{F_t} is the expectation with respect to the predictive CDF F_t and X are independent realisations of a variable with distribution F_t. The most unbiased model would have exactly half of forecasts above or equal to the data at time t and $B_t = 0$, whereas a completely biased model would yield either all forecasts above ($B_t = 1$) or below ($B_t = -1$) the data. To get a single bias score U, we took the mean across forecast time

$$B(F_t, x_t) = \frac{1}{T} \sum_t B_t(F_t, x_t), \tag{6}$$

where T is the number of forecasting time points.

Lastly, we evaluated forecasts using the *Continuous Ranked Probability Score* (CRPS, Hersbach, 2000). CRPS is a distance measure that measures forecasting performance at the scale of the predicted data, combining an assessment calibration and sharpness. It is a *strictly proper forecasting score*, that is one which is optimised if the predictive distribution is the same as the one generating the data, with 0 being the ideal score. CRPS reduces to the mean absolute error (MAE) if the forecast is deterministic and can therefore be seen as its probabilistic generalisation. It is defined as

$$\mathrm{CRPS}(F_t, x_t) = -\int_{-\infty}^{\infty} \left(F_t(y) - H(y - x_t)\right)^2 dy, \tag{7}$$

A convenient equivalent formulation using independent samples from F_t was suggested by Gneiting et al. (2007) and is given by

$$\mathrm{CRPS}(F_t, x_t) = E_{F_t}\left|X - x_t\right| - \frac{1}{2} E_{F_t}\left|X - X'\right|, \tag{8}$$

where X and X' are independent realisations of a random variable with CDF F_t.

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) 2.2–3.8, 95% credible interval (CI) 1.1–7.8) in the first fitted week (beginning 10 August, 2014) to a median estimate of 1.3 (IQR 0.9–1.9, 95% CI 0.3–3.9) in early October, 1.4 (IQR 1.0–2.0, 95% CI 0.4–4.6) in early November, 1IQR 0.7–1.4, 95% CI 0.2–3.0) in early December, 0.6 in early January (IQR 0.4–0.9, 95% CI 0.1–1.9) and 0.3 at the end of the epidemic in early Feburary (IQR 0.2–0.5, 95% CI 0.1–1.3).

Forecasts from the semi-mechanistic model were calibrated for one or two weeks, but deteriorated rapidly at longer forecasting horizons (Table 1 and Fig. 2). The two best calibrated models used deterministic forecasts starting at the last fitted data point. Of these two, forecasts that kept the transmission rate constant from the value at the last data point performed

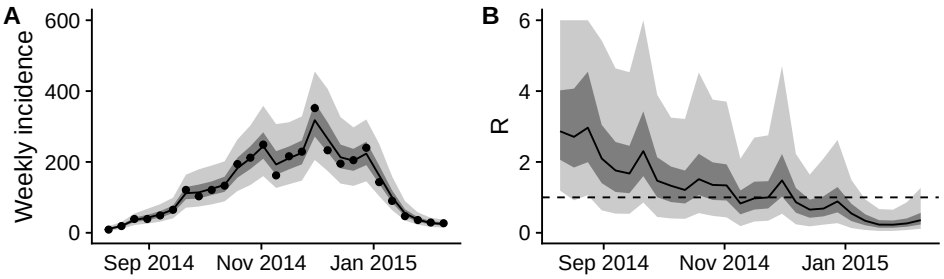

Figure 1. Final fit of the semi-mechanistic model to the Ebola outbreak in Western Area, Sierra Leone. (A) Final fit of the reported weekly incidence (black line and grey shading) to the data (black dots). (B) Corresponding dynamics of the reproduction number (ignoring depletion of susceptibles). Point-wise 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.

slightly better than one that continued to change the transmission rate following a random walk with volatility estimated from the time series. Both of the best calibrated models were calibrated for two-week ahead forecasts, and possibly calibrated for three weeks. All of the model variants were uncalibrated four weeks or more ahead, and none of the stochastic models was calibrated for any forecast horizon.

The best-calibrated of our semi-mechanistic forecasts was better calibrated than any of the null models (Fig. 3A) for up to three weeks. While the autoregressive null model was calibrated for 1-week-ahead forecasts, it was not calibrated for longer forecast horizons. 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, was only possibly calibrated for 1-week ahead and uncalibrated beyond. The deterministic null model was uncalibrated for all forecast horizons.

Our model as well as all null models except the unfocused model showed a tendency to overestimate the predicted number of cases (Fig. 3B). This bias increased with the forecast horizon. The best-calibrated semi-mechanistic model 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). This is reflected in the mean CRPS values (Fig. 3D), which combine calibration and sharpness and reflect a probabilistic analogue

Model				Forecast horizon (weeks)			
stochasticity	start	averaged	volatility	1	2	3	4
deterministic	at last data point	no	yes	**0.24**	**0.1**	0.01	<0.01
deterministic	at last data point	no	no	**0.3**	**0.13**	0.02	<0.01
deterministic	at last data point	2 weeks	no	**0.26**	0.03	<0.01	<0.01
deterministic	at last data point	3 weeks	no	**0.24**	<0.01	<0.01	<0.01
deterministic	1 week before	no	yes	**0.05**	0.01	<0.01	<0.01
deterministic	1 week before	no	no	**0.07**	0.02	<0.01	<0.01
deterministic	1 week before	2 weeks	no	**0.08**	<0.01	<0.01	<0.01
deterministic	1 week before	3 weeks	no	0.03	<0.01	<0.01	<0.01
stochastic	at last data point	no	yes	0.02	0.02	<0.01	<0.01
stochastic	at last data point	no	no	0.02	0.02	<0.01	<0.01
stochastic	at last data point	2 weeks	no	0.01	<0.01	<0.01	<0.01
stochastic	at last data point	3 weeks	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	no	yes	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	no	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	2 weeks	no	<0.01	<0.01	<0.01	<0.01
stochastic	1 week before	3 weeks	no	<0.01	<0.01	<0.01	<0.01

Table 1. Calibration of forecast model variants of our semi-mechanistic model. Shown is the calibration (p-value of the Anderson-Darling test of uniformity) for deterministic and stochastic forecasts starting either at the last data point or one week before, either starting from the last value of the transmission rate or from an average over the last 2 or 3 weeks, and including volatility (in a Gaussian random walk) in the transmission rate or not, at different forecast horizons up to 4 weeks. The p-values highlighted in bold reflect predictive models we consider likely to be calibrated.

to the MAE. At 1-week ahead, the mean CRPS values of the autoregressive, unfocused and best semi-mechanistic forecasting models were all around 30 (i.e., on average the prediction was out by approximately 30 cases). At increasing forecasting horizon, the CRPS of the semi-mechanistic model grew faster than the CRPS of the autoregressive and unfocused null models, but since these were no longer calibrated at horizons loner than one week, the semi-mechanistic model would still be preferred for forecast horizons up to three weeks.

We 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 the semi-mechanistic model would

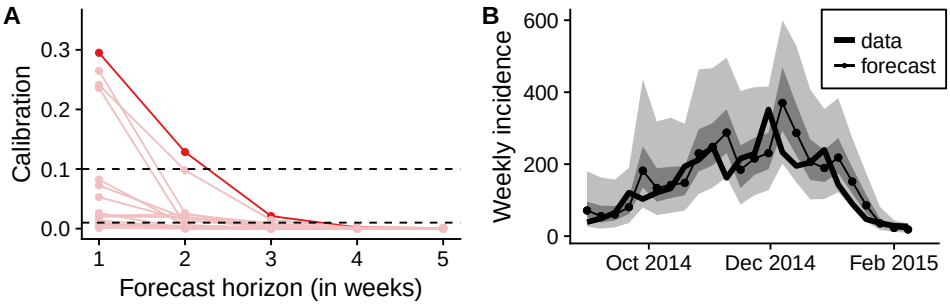

Figure 2. Calibration of forecasts from the semi-mechanistic model. (A) Calibration of model variants (p-value of Anderson-Darling test) as a function of the forecast horizon. Shown in dark red is the best calibrated forecasting model variant. Other model variants are shown in light red. (B) Comparison 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.

have been determined to be the best calibrated for forecasts 1 or 2 weeks ahead.

Discussion

Outbreaks of emerging infectious diseases in resource-poor settings are often characterised by limited data and a need for short-term forecasts to inform bed demands and allocation of other human and financial resources. Several groups produced and published forecasts over the course of the Ebola epidemic, and the alleged failure of some to predict the correct number of cases by several orders of magnitude generated some controversy around the usefulness of mathematical models (Butler, 2014; Rivers et al., 2014). To our knowledge, we were the only research team making weekly forecasts available in real time, distributing them to a wide range of international public health practitioners via a dedicated email list, as well as on a publicly accessible web page. 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.

Applying a suite of assessment methods to our forecasting model, we found that the used semi-mechanistic model variants were probabilistically

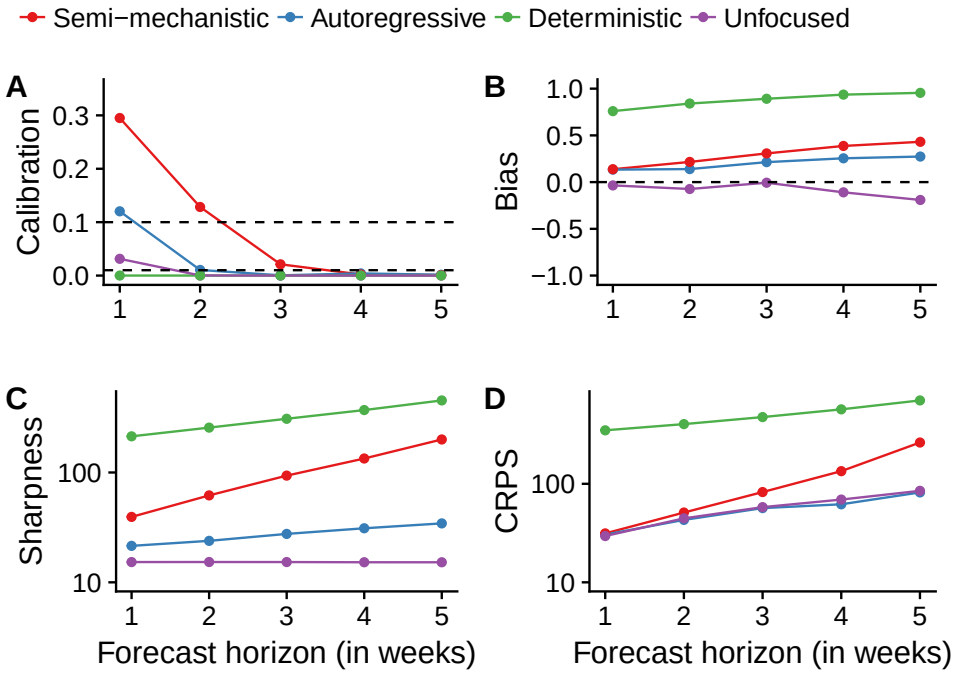

Figure 3. Forecasting metrics of the best semi-mechanistic model variant compared to null models. Metrics shown are (A) calibration (p-value of Anderson-Darling test, (B) bias, (C) sharpness (MADM) and (D) CRPS, all as a function of the forecast horizon.

calibrated to varying degree with the best ones calibrated for up to 2-3 weeks ahead, but performance deteriorated rapidly as the forecasting horizon increased. Since the model variants were similar enough to produce the same mean future trajectories, differences in calibration reflected differences in the quantification of uncertainty. The best performing forecasts were the once generated the least variance in the trajectories, indicating that, in general, our models overestimated the possible diversity in future trajectories. A possible future improvement could be to post-process predictions by tuning their variance to improve performance (Liu et al., 2015).

The rapid deterioration of probabilistic calibration even of our best performing model variants 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.

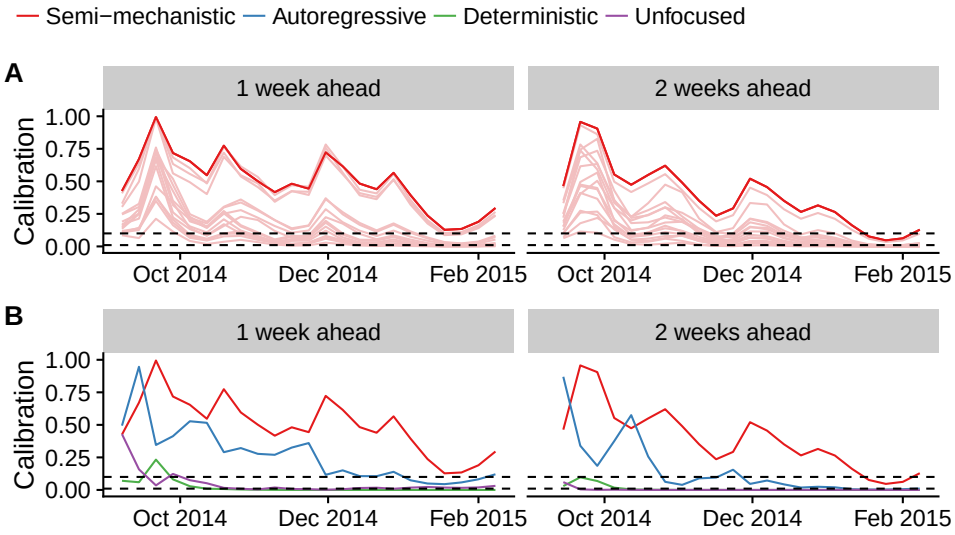

Figure 4. Calibration over time. Shown are calibration scores of the forecast up to the time point shown on the x-axis. (A) Semi-mechanistic model variants, with the best model highlighted in dark red and other model variants are shown in light red. (B) Best semi-mechanistic model and null models. In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

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. 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 a the observed decrease over time. Capturing such trends and modelling the underlying causes would be an important future improvement of real-time infectious disease models used for forecasting.

There can be trade-offs between achieving good outcomes on the different forecast metrics we used, so that 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 semi-mechanistic model we

used during the Ebola epidemic was better calibrated than the null models, this came at the expense of a decrease in the sharpness of forecasts. Comparing the models using the CRPS alone, the best calibrated semi-mechanistic model 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 ensemble forecasts (Wilks, 2018). 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.

Other models may have performed better than the ones presented here. 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 (Pei and Shaman, 2017).

In practice, there might be considerations beyond performance when choosing a model for forecasting. Our model combined a mixture of 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. At the same time, 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 (Camacho et al., 2015a; Camacho et al., 2017). 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. More generally, when choosing between forecast performance and the ability to explicitly account for the impact of interventions, a model that accounts for the latter might, in some cases, be preferable.

Epidemic forecasts played an important and prominent role in the response to and public awareness of the Ebola epidemic (Frieden and Damon,

2015). Forecasts have been used for vaccine trial planning against Zika virus (World Health Organization, 2017) 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 (Center for the Mathematical Modelling of Infectious Diseases, 2015). An agreement on standards of forecasting assessment is urgently needed in infectious disease epidemiology, and retrospective or even real-time assessment of forecasts 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. 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.

References

Andrieu, C., A. Doucet, and R. Holenstein (2010). "Particle Markov chain Monte Carlo methods". *J R Stat Soc B*, 269–342.

Biggerstaff, M. et al. (2016). "Results from the centers for disease control and prevention's predict the 2013–2014 Influenza Season Challenge". *BMC Infectious Diseases* 1, 357.

Butler, D. (Nov. 2014). "Models overestimate Ebola cases." *Nature* (7525), 18. ISSN: 1476-4687.

Camacho, A. et al. (Dec. 2015a). "Estimating the probability of demonstrating vaccine efficacy in the declining Ebola epidemic: a Bayesian modelling approach". *BMJ Open* 12, e009346.

Camacho, A. et al. (2015b). "Temporal Changes in Ebola Transmission in Sierra Leone and Implications for Control Requirements: a Real-Time Modelling Study". *PLOS Curr.: Outbreaks*.

Camacho, A. et al. (Dec. 2017). "Real-time dynamic modelling for the design of a cluster-randomized phase 3 Ebola vaccine trial in Sierra Leone". *Vaccine*.

Center for the Mathematical Modelling of Infectious Diseases (2015). *Visualisation and projections of the Ebola outbreak in West Africa. http://ntncmch.github.io/ebola/.* Archived at *http://www.webcitation.org/6oYEBYoeD* on Feb 24, 2017.

Centres for Disease Prevention and Control (Oct. 2017). *Epidemic Prediction Initiative.* URL:

https://predict.phiresearchlab.org/legacy/dengue/index.html, Archived at *http://www.webcitation.org/6rsS5QDar* on 11 July, 2017.

Chowell, G. (2017). "Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts". *Infectious Disease Modelling* 3, 379–398.

Chowell, G. et al. (2017). "Perspectives on model forecasts of the 2014–2015 Ebola epidemic in West Africa: lessons and the way forward". *BMC Med* 1, 42.

Chretien, J.-P., S. Riley, and D. B. George (Dec. 2015). "Mathematical modeling of the West Africa Ebola epidemic". *eLife*, e09186.

Dawid, A. P. (1984). "Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach". *J R Stat Soc [Ser A]* 2, 278.

Dawson, P. M. et al. (2015). "Epidemic predictions in an imperfect world: modelling disease spread with partial data". In: *Proc. R. Soc. B*. 1808. The Royal Society, 20150205.

Dong, F. et al. (Dec. 2015). "Evaluation of ebola spreading in west africa and decision of optimal medicine delivery strategies based on mathematical models". *Infection, Genetics and Evolution*, 35–40.

Drake, J. M. et al. (2015). "Ebola cases and health system demand in Liberia". *PLoS Biol* 1, e1002056.

Dureau, J., S. Ballesteros, and T. Bogich (2013). "SSM: Inference for time series analysis with State Space Models".

Fisman, D., E. Khoo, and A. Tuite (2014). "Early epidemic dynamics of the west african 2014 ebola outbreak: estimates derived with a simple two-parameter model." *PLOS Curr.: Outbreaks*.

Frieden, T. R. and I. K. Damon (2015). "Ebola in West Africa — CDC's role in epidemic detection, control, and prevention". *Emerging Infectious Diseases* 11, 1897.

Friederichs, P. and T. L. Thorarinsdottir (2012). "Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction". *Environmetrics* 7, 579–594.

Funk, S. (2016). *rbi.helpers: rbi helper functions*. R package version 0.2.

Funk, S. et al. (Dec. 2017a). "Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model". *Epidemics*.

Funk, S. et al. (2017b). "The impact of control strategies and behavioural changes on the elimination of Ebola from Lofa County, Liberia". *Phil Trans Roy Soc B* (1721), 20160302.

Gneiting, T., F. Balabdaoui, and A. E. Raftery (Apr. 2007). "Probabilistic forecasts, calibration and sharpness". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2, 243–268.

Gneiting, T. and M. Katzfuss (Jan. 2014). "Probabilistic Forecasting". *Annual Review of Statistics and Its Application* 1, 125–151.

Goldstein, E. et al. (2011). "Predicting the epidemic sizes of influenza A/H1N1, A/H3N2, and B: a statistical method". *PLoS Med* 7, e1001051.

Heesterbeek, H. et al. (2015). "Modeling Infectious Disease Dynamics in the Complex Landscape of Global Health". *Science* 6227, aaa4339–aaa4339.

Held, L., S. Meyer, and J. Bracher (June 2017). "Probabilistic forecasting in infectious disease epidemiology: the 13th Armitage lecture". *Statistics in Medicine* 22, 3443–3460.

Hersbach, H. (2000). "Decomposition of the continuous ranked probability score for ensemble prediction systems". *Weather and Forecasting* 5, 559–570.

Jacob, P. E. and S. Funk (2017). *RBi: R interface to LibBi*. R package version 0.7.0.

Johansson, M. A. et al. (2016). "Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico". *Scientific reports*.

Lewnard, J. A. et al. (Dec. 2014). "Dynamics and control of Ebola virus transmission in Montserrado, Liberia: a mathematical modelling analysis." *Lancet Infect Dis* 12, 1189–1195.

Liu, F. et al. (2015). "Short-term forecasting of the prevalence of trachoma: expert opinion, statistical regression, versus transmission models". *PLoS neglected tropical diseases* 8, e0004000.

Lowe, R. et al. (2014). "Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts". *The Lancet infectious diseases* 7, 619–626.

Meltzer, M. I. et al. (Sept. 2014). "Estimating the future number of cases in the Ebola epidemic–Liberia and Sierra Leone, 2014-2015." *MMWR Surveill Summ*, 1–14.

Merler, S. et al. (Feb. 2015). "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* 2, 204–211.

Moran, K. R. et al. (2016). "Epidemic forecasting is messier than weather forecasting: The role of human behavior and internet data streams in epidemic forecast". *The Journal of Infectious Diseases* suppl_4, S404–S408.

Murray, L. (2015). "Bayesian State-Space Modelling on High-Performance Hardware Using LibBi". *Journal of Statistical Software, Articles* 10, 1–36. ISSN: 1548-7660.

National Oceanic and Atmospheric Administration (Oct. 2017). *Dengue Forecasting*. URL: *http://dengueforecasting.noaa.gov/*, Archived at *http://www.webcitation.org/6oWfUBKnC* on Feb 24, 2017.

Nishiura, H. and G. Chowell (2014). "Early transmission dynamics of Ebola virus disease (EVD), West Africa, March to August 2014". *Euro Surveill* (36), 20894.

Nsoesie, E., M. Mararthe, and J. Brownstein (2013). "Forecasting peaks of seasonal influenza epidemics". *PLoS currents*.

Pei, S. and J. Shaman (Oct. 2017). "Counteracting structural errors in ensemble forecast of influenza outbreaks". *Nature Communications* 1.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rivers, C. M. et al. (2014). "Modeling the impact of interventions on an epidemic of Ebola in Sierra Leone and Liberia". *PLOS Curr.: Outbreaks*.

Scott, S. L. (2017). *bsts: Bayesian Structural Time Series*. R package version 0.7.1.

Siettos, C. et al. (2015). "Modeling the 2014 ebola virus epidemic–agent-based simulations, temporal analysis and future predictions for liberia and sierra leone". *PLOS Curr.: Outbreaks*.

Towers, S., O. Patterson-Lomba, and C. Castillo-Chavez (2014). "Temporal variations in the effective reproduction number of the 2014 West Africa Ebola outbreak". *PLOS Curr.: Outbreaks*.

Viboud, C. et al. (Aug. 2017). "The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt". *Epidemics*.

White, R. A. et al. (2015). "Projected treatment capacity needs in Sierra Leone". *PLOS Curr.: Outbreaks*.

WHO Ebola Response Team (Sept. 2014). "Ebola Virus Disease in West Africa - The First 9 Months of the Epidemic and Forward Projections." *N Engl J Med*.

Wilks, D. S. (2018). "Enforcing calibration in ensemble postprocessing". *Quarterly Journal of the Royal Meteorological Society* 710, 76–84.

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

Yang, W. et al. (2015). "Forecasting influenza epidemics in Hong Kong". *PLoS computational biology* 7, e1004383.

Acknowledgements

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). RL was supported by a Royal Society Dorothy Hodgkins fellowship. RME was supported by 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.

Author contribution

SF, AC and WJE conceived the study; SF and AC analysed the data; SF wrote a first draft of the paper; AC led the generation of real-time forecasts during the Ebola epidemic; all authors contributed to the text of the final version.

Competing interests

There are no competing interests.

Data and materials availability

The authors declare that all data supporting the findings of this study will be available within the paper and its supplementary information files.