Assessing the performance of real-time epidemic forecasts

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

[1] Center for the Mathematical Modelling of Infectious Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom
[2] Epicentre, Paris, France
[3] Barcelona Institute for Global Health (ISGLOBAL), Barcelona, Spain
[*] Corresponding author. Email: sebastian.funk@lshtm.ac.uk

Abstract

Real-time forecasts based on mathematical models have become increasingly important to help guide critical decision-making during infectious disease outbreaks. Yet, epidemic forecasts are rarely evaluated during or after the event, and it has not been established what the best metrics for assessment are. Here, we disentangle different components of forecasting ability by defining three metrics that assess the calibration, sharpness and unbiasedness of forecasts. We use this approach to analyse the performance of weekly district-level forecasts generated in real time during the 2013–16 Ebola epidemic in West Africa, which informed a range of public health decisions during the outbreak. We found forecasting performance with respect to all three measures was good at short time horizons but deteriorated for long-term forecasts. This suggests that forecasts provided useful 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 two null models showed that the approach we chose performed best with respect to probabilistic calibration but sharpness decreased more rapidly for longer forecasting horizons than with simpler models. 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 [1]. Forecasting targets usually revolve around expected epidemic duration, size, or peak timing and incidence [2–6], geographical distribution of risk [7], or short-term trends in incidence [8, 9]. Despite their growing importance, however, the performance of forecasts made during an outbreak is rarely investigated during or after the event for their accuracy [8, 10].

Providing accurate forecasts during emerging epidemics comes with particular challenges as uncertainties about long-term circumstances, in particular human behavioural changes and public health interventions, can preclude reliable long-term predictions [11–13]. 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 a large scale outbreak forecasting effort 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, by fitting models to the available time series and running them forward to predict the future trajectory of the outbreak [14–24]. These were part of wider modelling efforts to evaluate the expected effect of interventions and assess the risk of international spread[25, 26]. 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 [27]. 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 particularly important "call to arms" that served to trigger the international response that helped avoid the worst-case scenario [28].

In contrast to one-off published forecasts, we produced weekly subnational 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 [29]. They were generated using a model that has, in variations, been used to forecast bed demand during the epidemic in Sierra Leone [19] and the feasibility of vaccine trials late in the epidemic [30, 31]. 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 develop assessment metrics that elucidate different properties of forecasts, in particular their probabilistic calibration, sharpness and unbiasedness. 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 was used to inform bed capacity planning. To investigate the accuracy of forecasts with different time horizons, we developed metrics to assess forecasts that elucidate different properties of forecasts, in particular their probabilistic calibration, sharpness and unbiasedness. In particular, we investigate the accuracy of forecasts with different time horizons. The development of a toolbox for assessment of real-time models is a critical step in preparedness planning for future epidemics. With this in mind, we focus on diagnostic tools for real-time and relatively short-term (i.e., a few weeks ahead) probabilistic forecasts in emergency situations with limited data to guide operational decisions.

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 [19]. 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 [19, 32]. Briefly, the model was based on a Susceptible-Exposed-Infectious-Recovered (SEIR) model with fixed incubation period of 9.4 days [33], 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. In the model, cases were 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. The time-varying transmission rate was modelled using a daily Gaussian random walk with fixed volatility (or standard deviation of step size). This imposed a relatively weak prior on the time course of the transmission rate. It was chosen to incorporate uncertainty in the behaviour of transmission intensity over time, reflecting behavioural changes in the community, public health interventions or other factors affecting transmission for which information was not available at the time. The volatility of the random walk was estimated as part of the inference procedure (see below). Its value determined degree to which the transmission rate could change each day. To ensure it remained positive, we log-transformed the transmission rate. 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 and forecasting

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 reporting (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 [34] as implemented in the *ssm* library [35]. For each forecast, 50,000 samples were extracted and thinned 5,000.

To produce forecasts, the value of the transmission rate was fixed to its last value and the model run forward with a fixed transmission rate. In other words, the forecasts should be interpreted as projections of what were to happen if no further changes occurred to the transmission rate. We used the samples of the posterior distribution generated using the Monte Carlo algorithm to produce a range of predictive trajectories. We used point-wise medians and confidence intervals across all these trajectories as forecasts with a given time horizon of up to 10 weeks.

To assess the performance of the semi-mechanistic transmission model we compared it to two simpler sub-models that represented its constituent parts and served as null models: one that only contained the mechanistic core of the semi-mechanistic model with a fixed transmission rate, and a non-mechanistic model where the number of cases followed a Wiener process, with forecasts generated assuming the weekly number of new cases was not going to change. The simpler models were implemented in *libbi* [36] via the *RBi* [37] and *RBi.helpers* [38] *R* packages [39].

Metrics

We assessed forecasts with respect to their *calibration* and *sharpness*, existing concepts in evaluating forecasts [40], as well as the ability of models to avoid bias. For these three concepts, we developed scoring metrics that take real values between 0 (worst possible performance) and 1 (best possible performance), and evaluated them using Monte-Carlo samples from the predictive posterior distributions.

Calibration or reliability [41] of forecasts refers to the ability of a model to correctly identify its own uncertainty in forecasting. It can be assessed using the probability integral transform [42],

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

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). If the true probability distribution of outcomes 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, u_t is distributed uniformly.

To define a single calibration score C, we assessed the degree to which the values u_t were uniformly distributed, by dividing the interval $[0, 1]$ into

$m \ll n$ sub-intervals of equal size and calculating

$$C(F_t, x_t) = 1 - \frac{m}{2(m-1)} \sum_j \left| p_j - \frac{1}{m} \right| \tag{3}$$

where p_j is the proportion of values of u_t in the j-th sub-interval. A perfectly calibrated model would yield $p_j = 1/m$ for all j, and $C = 1$. A very poorly calibrated model would yield $p_j = 1$ for some j, and $C = 0$.

Unbiasedness of forecasts is the ability of the model to not systematically over- or underestimate data. We defined unbiasedess at time t as

$$U_t(F_t, x_t) = 1 - 2 \left| \int_{-\infty}^{\infty} F_t(y) H(y - x_t) dy - 0.5 \right| \tag{4}$$

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

$$U_t(F_t, x_t) = 1 - 2 \left| E_{F_t} \left[H(X - x_t) \right] - 0.5 \right| \tag{5}$$

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 $U_t = 1$, whereas a completely biased model would yield either all or no forecasts above the median and therefore have $U_t = 0$. To get a single unbiasedness score U, we took the mean across forecast time

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

where T is the number of forecasting time points.

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. We defined sharpness at time t as

$$S_t(F_t) = 1 - \frac{\mathrm{MADM}(y)}{m(y)}, \tag{7}$$

where y is a variable distributed according to F_t, and MADM(y) is the normalised median absolute deviation about the median $m(y)$ of y,

$$\mathrm{MADM}(y) = m\left(|y - m(y)|\right) \tag{8}$$

The sharpest model would focus all forecasts on one point and have $S = 1$, whereas a completely blurred forecast would have $S \to 0$. Again, we used Monte-Carlo samples X from F_t to estimate sharpness. To get a single unbiasedness score S, we took the mean across forecast time

$$S(F_t, x_t) = \frac{1}{T}\sum_t S_t(F_t, x_t) \tag{9}$$

We also evaluated forecasts using an established metric, the *continuous ranked probability score* [CRPS, 43]. CRPS is a distance measure that measures forecasting performance at the scale of the predicted data, with 0 being the ideal score. It reduces to the mean absolute error if the forecast is deterministic and can therefore be seen as its probabilistic generalisation. It is defined as

$$\text{CRPS}(F_t, x_t) = -\int_{-\infty}^{\infty} \left(F_t(y) - H(y - x_t)\right)^2 dy, \tag{10}$$

A convenient equivalent formulation using independent samples from F_t was suggested by Gneiting et al. [40] and is given by

$$\text{CRPS}(F_t, x_t) = E_{F_t}\left|X - x_t\right| - \frac{1}{2}E_{F_t}\left|X - X'\right|, \tag{11}$$

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

Sample size

The measures of calibration and unbiasedness we defined are, strictly speaking, outcomes of finite random trials and therefore subject to limits even if forecasts are ideal. To assess behaviour of calibration and unbiasedness under ideal forecasts we conducted a simulation study by generating random samples of varying number T (corresponding to the number of forecasts n). For calibration C, we randomly sampled T proportions p_j from a multinomial distibution with equal probability $p_j = 1/j$ for each j. For unbiasedness U, we randomly generated T samples from integers $1 \ldots n$ (where n is the number of posterior samples generated in our Monte-Carlo samples, for comparison), calculating unbiasedness using the mean M of these samples and

$$U_{\text{ideal}} = 1 - 2\left|\frac{M}{n}\right| \tag{12}$$

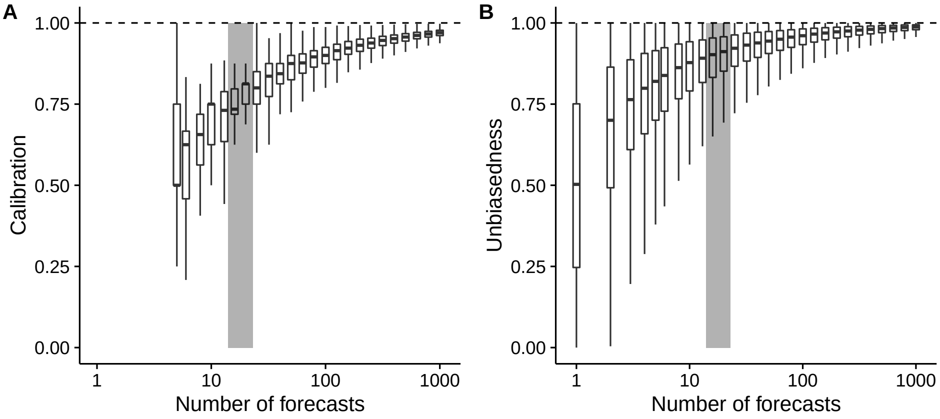

Figure 1. Performance of an ideal forecaster, determined in a simulation study as a function of the number of forecasts T (on a logarithmic scale), with respect to (A) calibration (with $m = 5$) and (B) unbiasedness (with $n = 5000$ samples). The number of forecasts assessed for Ebola in Western Area (15–24) is shaded in grey.

Results

We first assessed the forecasting scores achievable as a function of the number of forecasts considered. Performance as measured by the forecasting metrics of calibration and unbiasedness improves with the number of forecasts even when forecasts are perfect in a probabilistic sense (Fig. 1). For a single forecast, calibration (which relies on uniformity of a histogram) and unbiasedness cannot be defined. For the number of times we generated forecasts for the Western Area district of Sierra Leone (15–24, depending on the number of forecast weeks), achievable values ranged from 0.66–0.86 (IQR, calibration) and 0.83–0.96 (IQR, unbiasedness). Even if forecasts were perfect, it would take more than 100 forecasts to achieve scores greater than 0.8 for calibration and unbiasedness more than 95% of the time.

The forecasts generated during the Ebola epidemic were done by first fitting the model up to the current time point and then running the model forward with a the transmission rate fixed to its value at the last datapoint. The semi-mechanistic model used to generate real-time forecasts during the epidemic was able to reproduce the time up to the date of each forecast, following the data closely by means of the smoothly varying transmission rate (Fig. 2). The overall behaviour of the reproduction number 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,

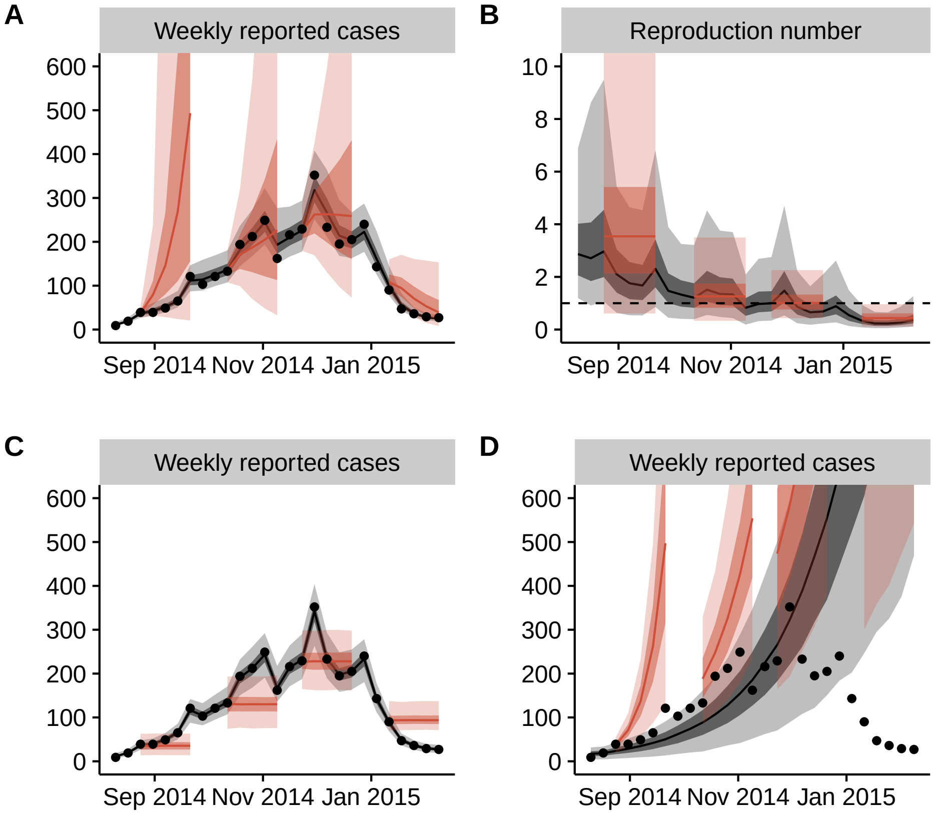

Figure 2. Example of model fits and forecasts produced during the Ebola outbreak in Western Area, Sierra Leone, using a semi-mechanistic model (A, B) and non-mechanistic (C) and deterministic mechanistic (D) null models. Shown in all panels is the final fit (black line and grey shading) to the data (black dots, panels A, C, D) and corresponding evolution of the basic reproduction number for the semi-mechanistic model (panel B, ignoring depletion of susceptibles because of small numbers). Four-week forecasts generated at four different time points are shown in red. Point-wise median state estimates are indicated by a solid curve, 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.

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, 1.0 (IQR 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).

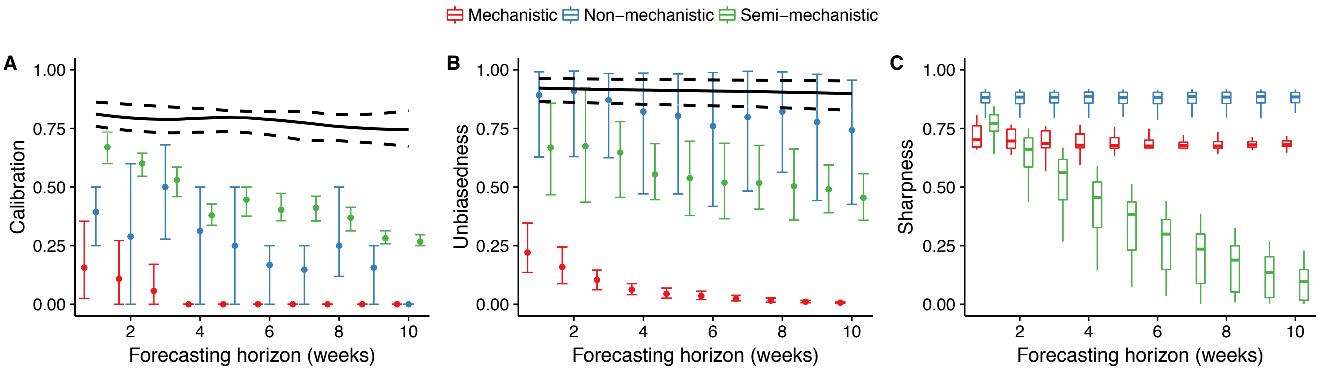

Figure 3. Performance of forecasts generated using the semi-mechanistic model (green) compared to two null models (red: deterministic SEIR, blue: constant incidence), as a function of the forecasting horizon: (A) calibration, (B) unbiasedness and (C) sharpness. In (A) and (B), the points indicate mean estimates and the error bars binomial 95% confidence intervals, estimated using a bootstrap. In (C), the horizontal bar indicates the median, top and bottom hinges the interquartile range (IQR), and the length of the whiskers data points within 1.5 times the IQR from the hinges. In (A) and (B), estimated performance of an ideal forecaster is indicated by dashed and solid black curves, corresponding to the median and interquartile range as in Fig. 1, respectively.

Applying the defined forecasting scores to our Ebola forecasts, we found that probabilistic calibration of forecasts was close to the achievable optimum for 1-week-ahead forecasts using the semi-mechanistic model (median: 0.67, IQR: 0.59–0.72), but rapidly deteriorated as the forecasting horizon increased (Fig. 3). At 4 weeks ahead, the median calibration score was 0.38 (IQR: 0.33–0.43), dropping to 0.27 (IQR: 0.25–0.30) at 10 weeks. Forecasts using the semi-mechanistic models were slightly biased at short forecast horizons, with bias increasing with lead time. The median estimate of unbiasedness was 0.67 at a 1-week horizon (IQR: 0.47–0.86), dropping to 0.55 (IQR: 0.45–0.69) at a 4-week horizon and 0.45 (IQR: 0.36–0.56) at 10 weeks. Sharpness of the forecasts generated by the semi-mechanistic model were initially high but deteriorated rapidly, from a median value of 0.77 for a forecasting horizon of 1 week (IQR: 0.74–0.81) to a median of 0.45 (IQR: 0.33–0.52) at 4 weeks and 0.10 (IQR: 0.02–0.15) at 10 weeks.

Comparing the semi-mechanistic model with a simpler mechanistic (i.e., pure deterministic SEIR) and a non-mechanistic (i.e., pure random walk) model revealed trade-offs in forecasting performance (Fig. 3). The semi-mechanistic model consistently performed best with respect to calibration, at the expense of unbiasedness and sharpness, both of which were lower than for the non-mechanistic model. The non-mechanistic model was not as well calibrated as the semi-mechanistic model, but produced largely unbiased forecasts of high sharpness (median: 0.88, IQR: 0.86–0.90). While forecasts generated by the mechanistic model were sharper than ones generated by the semi-mechanistic model, the model generally yielded poor fits to the data and forecasts, with low calibration (median: 0.16, IQR: 0.02–0.35) and unbiasedness (median: 0.22, IQR: 0.14–0.35) scores, even at 1 week ahead.

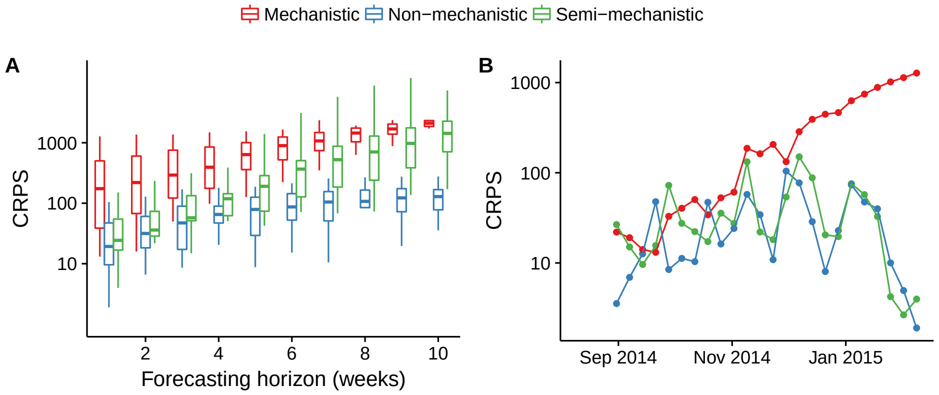

Figure 4. The Continuous Ranked Probability Score (CRPS) as a function of (A) the number of weeks ahead for which the forecast was generated and (B) the date of the forecast.

The CRPS is a probabilistic generalisation of the mean absolute error and can therefore be interpreted as the mean number of cases by which the forecast missed the true number. It increased as the forecasting horizon increased for all three models (Fig. 4A). At 1 week ahead, the median CRPS of the semi-mechanistic model was 24 (IQR: 17–55), rising to 190 (IQR: 70–290) at 4 weeks and 1400 (700–2200) at 10 weeks. The non-mechanistic model achieved similar but slightly lower CRPS at a 1 week horizon (median: 20, IQR: 10–47), with more significantly better scores at 4 weeks (median: 65, IQR: 47–89) and 10 weeks (median: 130, IQR: 70–170). The mechanistic model yielded consistently higher CRPS than the two other models.

The behaviour of the CRPS of 1-week forecasts over time yields differences in the ability of the different models to predict the epidemic trajectory at different time points (Fig. 4B). Generally, the CRPS of the semi-mechanistic model increased, indicating poor performance, at times when there was a change in direction of the epidemic trajectory, pointing to the inability of the model to correctly predict such a change. The CRPS of the non-mechanistic model, on the other hand, increased at times when the number of cases underwent a large change from one week to the next. The CRPS of the purely mechanistic model increased over time, indicating the increasingly poor performance of the model.

Discussion

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 [17, 44]. Defining optimum way to evaluated forecasts retrospectively, with respect to the specific aims of the forecasting framework in question, is an area of research of paramount importance. To our knowledge, we were the only research team making weekly forecasts available to the public 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 simple, flexible model. More generally, 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.

Applying a suite of assessment methods to our forecasts, we found they consistently provided good probabilistic calibration, sharpness and unbiasedness at short time horizons, but performance deteriorated as forecasting horizon increased. This reflects our lack of certainty about the underlying processes shaping the epidemic, 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. This is most likely because we assumed no future change in the transmission rate. In reality, transmission decreased significantly 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, but assumed that it remained constant. Capturing such trends and modelling the underlying causes would be an important future improvement of real-time infectious disease models, and help move them from scenario forecasts towards true prediction.

The aim of any forecast should be to maximise the sharpness of predictive distributions subject to calibration [40] while avoiding bias. In practice, there can be trade-offs between achieving good outcomes on these measures, 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 scores for probabilistic calibration, unbiasedness and sharpness highlights the underlying trade-offs. While the semi-mechanistic model we used during the Ebola epidemic was better calibrated than two simpler models, one purely stochastic and one purely mechanistic, this came at the expense of a decrease in the sharpness of forecasts. Comparing the models using the CRPS, a score combining assessment of calibration and sharpness, the simpler non-mechanistic model would be preferred to the semi-mechanistic model because the greater sharpness compensates for poorer calibration. In contrast, in the context of forecasts during epidemics, probabilistic calibration while avoiding bias should always be the main goal. This allows researchers to make 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, 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. However, 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.

There is a wide range of non-mechanistic methods for time-series forecasting [45], which we did not consider. 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. This had the advantage of allowing the assessment of intervention impact 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 [30, 31]. At the same time, the model is flexible enough to fit most time series, and this flexibility might mask underlying misspecifications. More generally, when choosing between model performance and the ability to explicitly account for the impact of interventions, a model that accounts for the latter might be preferable.

Epidemic forecasts played an important and prominent role in the response to and public awareness of the Ebola epidemic [28]. Forecasts are currently being used for vaccine trial planning against Zika virus [46] and will be called upon again to guide 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 [29]. Better standards of forecasting assessments are urgently needed, 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 developed 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 possible, a multi-source data, such as epidemiological and genetic data, could increase predictive power. However, it is the responsibility of researchers to not only generate and publish forecasts during an ensuing emergency, but to honestly and carefully assess forecast performance during and after the event and ensure lessons are learned for the next impending epidemic.

References

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

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

[3] E. Nsoesie, M. Mararthe, and J. Brownstein. “Forecasting peaks of seasonal influenza epidemics”. *PLoS currents* (2013).

[4] W. Yang et al. “Forecasting influenza epidemics in Hong Kong”. *PLoS computational biology* 7 (2015), e1004383.

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

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

[7] R. Lowe 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* 7 (2014), 619–626.

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

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

[10] R. Lowe et al. "Evaluating probabilistic dengue risk forecasts from a prototype early warning system for Brazil". *Elife* (2016), e11285.

[11] K. R. Moran et al. "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 (2016), S404–S408.

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

[13] S. V. Scarpino and G. Petri. "On the predictability of infectious disease outbreaks" (Mar. 21, 2017).

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

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

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

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

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

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

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

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

[22] S. Merler 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* 2 (Feb. 2015), 204–211.

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

[24] R. A. White et al. “Projected treatment capacity needs in Sierra Leone”. *PLOS Curr.: Outbreaks* (2015).

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

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

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

[28] T. R. Frieden and I. K. Damon. “Ebola in West Africa — CDC’s role in epidemic detection, control, and prevention”. *Emerging Infectious Diseases* 11 (2015), 1897.

[29] Center for the Mathematical Modelling of Infectious Diseases. *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. 2015.

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

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

[32] S. Funk et al. “Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model”. *Epidemics* (Dec. 2017).

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

[34] C. Andrieu, A. Doucet, and R. Holenstein. “Particle Markov chain Monte Carlo methods”. *J R Stat Soc Ser B* (2010), 269–342.

[35] J. Dureau, S. Ballesteros, and T. Bogich. “SSM: Inference for time series analysis with State Space Models” (July 2013).

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

[37] P. E. Jacob and S. Funk. *RBi: R interface to LibBi*. R package version 0.7.0. 2017.

[38] S. Funk. *rbi.helpers: rbi helper functions*. R package version 0.2. 2016.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017.

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

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

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

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

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

[45] C. Chatfield. *Time-series Forecasting*. Chapman and Hall/CRC Press, Boca Raton, United States, 2000.

[46] World Health Organization. *?Efficacy trials of ZIKV Vaccines: endpoints, trial design, site selection.? WHO Workshop Meeting Report.* 2017.