Refining Epidemiological Forecasts with Simple Scoring Rules

Robert E. Moore, Conor Rosato and Simon Maskell

Department of Electrical Engineering and Electronics, University of Liverpool, Brownlow Hill, Liverpool, L69 3GJ, UK

March 2022

Abstract

Estimates from infectious disease models have constituted a significant part of the scientific evidence used to inform the response to the COVID-19 pandemic in the UK. These estimates can vary strikingly in their bias and variability. Epidemiological forecasts should be consistent with the observations that eventually materialise. We use simple scoring rules to refine the forecasts of a novel statistical model for multisource COVID-19 surveillance data by tuning its smoothness hyperparameter.

1 Introduction

Several epidemiological modelling groups use statistical models of infectious disease to generate forecasts that contribute to a body of scientific evidence that informs the response to the COVID-19 pandemic in the UK. The models developed by the University of Cambridge MRC Biostatistics Unit and Public Health England (PHE) [1], the University of Warwick [2], and the London School of Hygiene & Tropical Medicine (LSHTM) [3] provide three notable examples of statistical models used to produce such estimates.

Although Cramer et al. [4] and Funk et al. [5] consider the assessment of quantile-format forecasts for COVID-19, it does not appear to be standard practice to assess full-distribution epidemiological forecasts by comparing them to the observations that eventually materialise. Maishman et al. [6] provide a set of anonymised estimates for the effective reproduction number R_t that highlights the striking differences in the bias and variability of estimates that different epidemiological models can produce. We contribute to the collective effort of modelling COVID-19 in this paper by introducing a statistical model for multisource COVID-19 surveillance data and by using simple scoring rules to refine its forecasts and improve its predictive performance. The statistical model is novel by its use of symptom report data from the NHS 111 telephone service, its compartments for convalescing and terminally ill individuals, and its implementation, which uses a bespoke numerical integrator to solve the system of ODEs for the transmission model.

We begin by describing the novel statistical model in Section 2 and then proceed to define a set of simple scoring rules in Section 3. In Section 4, we use the simple scoring rules to refine the forecasts of the statistical model. Finally, in Section 5, we bring the paper to an end by presenting our conclusions.

2 Statistical Model

The model we describe in this section builds on a previous version whose implementation we contributed to the CoDatMo (Covid Data Models) organisation on GitHub [7]. The purpose of CoDatMo is to provide a collection of COVID-19 models, all written in the statistical modelling language Stan [8]. In addition to models originally implemented in Stan, CoDatMo provides Stan

instantiations of COVID-19 models, initially written in other programming languages. By hosting a set of well-documented models, all implemented in a common language, CoDatMo hopes to improve understanding of the set of existing models and make it easier for newcomers and established researchers within the domain of epidemiology to make extensions and potential improvements.

We note that CoDatMo has already had some success against its objectives, with a group, primarily based at Universidade Nove de Julho in Brazil, creating a related but simpler model [9]. We also note that the UK Health Security Agency uses a slightly more sophisticated version of the model presented in this paper to generate weekly estimates of the effective reproduction number and growth rate for regions of the UK [10].

The statistical model consists of two parts: a transmission model that captures a simplified mechanism for the spread of coronavirus through the population; and an observation model that encapsulates the assumptions about the connection between the states of the transmission model and the observed surveillance data used to calibrate the model.

2.1 Transmission Model

The simple SIR compartmental model developed by Kermack and McKendrick [11] provides the theoretical basis for the transmission model. As with the SIR model, we assume a single geographical region with a large population of identical individuals who come into contact with one another uniformly at random but do not come into contact with individuals from other areas. In contrast to some other models, for example, those developed by Birrell et al. [1] and Keeling et al. [2], we treat the population as identical in terms of age and sex and only discriminate between individuals based on their disease states. We also assume that the population is closed, meaning that no births or deaths occur and no migration in or out of the population occurs.

The SIR model divides the population into three disease states: individuals who are (S) susceptible to infection; individuals who have been infected with the disease and are (I) infectious to other people; and individuals who have (R) "recovered" either by recuperating from the disease or dying. We augment these compartments in the transmission model by adding disease states for individuals who have been (E) exposed to the virus but are not yet infectious, individuals that are no longer infectious and whose final disease states are (P) pending, and individuals who have (D) died of the disease. The exposed and dead compartments are standard extensions to the original SIR model. In contrast, we believe the pending compartments to be a novel aspect of the model. These compartments contain individuals who are either convalescing or terminally ill as a result of infection. In addition to adding these compartments, we redefine the (R) recovered population to include the living only.

Inspired by the work of the University of Cambridge MRC Biostatistics Unit and Public Health England (PHE) [1], we partition each of the intermediate disease states (E, I, P) into two sub-states. Partitioning the sub-states in this way makes the model more realistic by implicitly constraining the times spent in each of these disease states to have Erlang rather than Exponential distributions.

We assume that there is at least one individual in each susceptible, exposed, and infectious compartment on the 17th of February 2020, which is the beginning of time in the model. At this stage, we assume that no population members are pending, have recovered, or have died. Two parameters, α_1 and α_2, determine the allocation of the rest of the population to the first five compartments of the transmission model. α_1 is the proportion of the remaining population initially in the susceptible compartment, and α_2 is the proportion of the infected population that is not yet infectious at time zero. For simplicity, we divide the exposed and infectious populations equally between the respective sub-states. We can express the exact relationship between the two parameters, α_1 and α_2, and the initial state of the transmission model as a set of equations:

$$S(0) = (N-5)\alpha_1 + 1, \tag{1}$$

$$E_1(0) = \frac{1}{2}(N-5)(1-\alpha_1)\alpha_2 + 1, \tag{2}$$

$$E_2(0) = \frac{1}{2}(N-5)(1-\alpha_1)\alpha_2 + 1, \tag{3}$$

$$I_1(0) = \frac{1}{2}(N-5)(1-\alpha_1)(1-\alpha_2) + 1, \tag{4}$$

$$I_2(0) = \frac{1}{2}(N-5)(1-\alpha_1)(1-\alpha_2) + 1, \tag{5}$$

$$P_1(0) = 0, \tag{6}$$

$$P_2(0) = 0, \tag{7}$$

$$R(0) = 0, \tag{8}$$

$$D(0) = 0. \tag{9}$$

Figure 1 provides a graphical illustration of the transmission model that captures the assumptions relating to the flow of individuals between disease states.
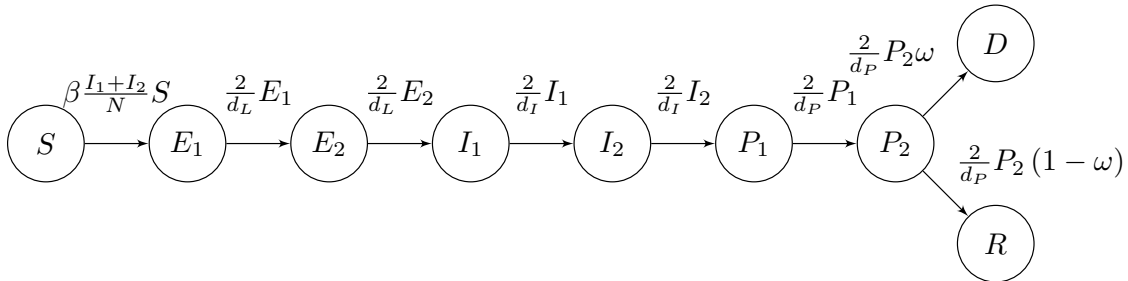

Figure 1: A graph of the transmission model. Individuals begin their journey in the susceptible (S) state. From here, they are infected and move into the exposed (E) state. After the virus has incubated for a while, they continue into the infectious (I) state. Next, they enter the pending (P) state, after which they either migrate into the recovered (R) state if convalescing or pass into the deceased (D) state if terminally ill.

We assume that the population randomly mixes as time elapses, with infectious and susceptible individuals coming into contact with one another, potentially transmitting the virus. Susceptible people who have become exposed through these contacts are not initially infectious. The virus replicates in their bodies for a time, known as the latent period, before they become infectious and have the potential to transmit the virus onto members of the remaining susceptible population. After being infectious for some time, we assume that individuals enter a state of pending before either recovering and becoming indefinitely immune to reinfection if they were convalescing or dying if they were terminally ill.

The number of individuals in each disease state varies with time according to a system of ordinary differential equations:

3

$$\frac{dS(t)}{dt} = -\beta(t)\frac{I_1(t) + I_2(t)}{N}S(t), \tag{10}$$

$$\frac{dE_1(t)}{dt} = \beta(t)\frac{I_1(t) + I_2(t)}{N}S(t) - \frac{2}{d_L}E_1(t), \tag{11}$$

$$\frac{dE_2(t)}{dt} = \frac{2}{d_L}\big[E_1(t) - E_2(t)\big], \tag{12}$$

$$\frac{dI_1(t)}{dt} = \frac{2}{d_L}E_2(t) - \frac{2}{d_I}I_1(t), \tag{13}$$

$$\frac{dI_2(t)}{dt} = \frac{2}{d_I}\big[I_1(t) - I_2(t)\big], \tag{14}$$

$$\frac{dP_1(t)}{dt} = \frac{2}{d_I}I_2(t) - \frac{2}{d_P}P_1(t), \tag{15}$$

$$\frac{dP_2(t)}{dt} = \frac{2}{d_P}\big[P_1(t) - P_2(t)\big], \tag{16}$$

$$\frac{dR(t)}{dt} = \frac{2}{d_P}P_2(t)\big[1 - \omega\big], \tag{17}$$

$$\frac{dD(t)}{dt} = \frac{2}{d_P}P_2(t)\omega, \tag{18}$$

where

- $S(t)$ is the number of susceptible individuals who have not yet been infected and are at risk of infection,

- $E_1(t) + E_2(t)$ is the number of exposed individuals who have been infected but are not yet infectious,

- $I_1(t) + I_2(t)$ is the number of infectious individuals,

- $P_1(t) + P_2(t)$ is the number of pending individuals who are either convalescing or are terminally ill,

- $R(t)$ is the number of recovered individuals,

- $D(t)$ is the number of dead individuals,

- $N = S(t) + E_1(t) + E_2(t) + I_1(t) + I_2(t) + P_1(t) + P_2(t) + R(t) + D(t)$ is the constant total number of individuals in the population,

- d_L is the mean time between infection and onset of infectiousness,

- d_I is the mean time for which individuals are infectious,

- d_P is the mean time for which individuals are pending,

- ω, the infection fatality ratio (IFR), is the proportion of infected individuals who will die,

- $\beta(t)$ is the mean rate of contacts between individuals per unit time that are sufficient to lead to transmission if one of the individuals is infectious and the other is susceptible. $\beta(t)$ is a continuous piecewise linear function of time:

$$\beta(t) = \sum_{j=1}^{J} \beta_j(t)\chi_{[t_{j-1}, t_j)}(t), \tag{19}$$

where the mean rate of effective contacts during the jth time interval, $\beta_j(t)$, is given by

$$\beta_j(t) = \frac{\beta_{j+1} - \beta_j}{t_j - t_{j-1}}(t - t_{j-1}) + \beta_j, \tag{20}$$

and

$$\chi_{[t_{i-1},t_i)}(t) = \begin{cases} 1 & \text{if } t \in [t_{i-1}, t_i), \\ 0 & \text{if } t \notin [t_{i-1}, t_i). \end{cases} \tag{21}$$

The effective contact rate parameters $\beta_1, \beta_2, ..., \beta_{J+1}$ in Equation 20 are the values $\beta(t)$ takes on a set of predefined dates $t_0, t_1, ..., t_J$. The first date is the 17th of February 2020, and each date that follows is seven days after the last, with the second date being the 24th of March 2020, the first day after the Prime Minister announced the first national lockdown.

2.2 Observation Model

Epidemiological modelling groups use many types of surveillance data to calibrate statistical models of infectious diseases. The observation model captures the assumptions about the relationship between the states of the transmission model and the surveillance data that we use for calibration.

We have designed the observation model to be extensible. Here, the model only has components for death, hospital admission and symptom report data to keep things simple. Nonetheless, the observation model can be extended to assimilate additional types of surveillance data, such as case data, by appending extra components similar in structure to those for ingesting the hospital admission and symptom report data.

2.2.1 Death Data

On their official website for coronavirus data [12], the UK government publishes a daily time series of the number of deaths of individuals whose death certificate mentioned COVID-19 as one of the causes. We assume that the number of deaths on day t, according to this definition, $d_{\text{obs}}(t)$, has a negative binomial distribution parameterised by a mean $d(t)$ and parameter ϕ_{deaths} which affects overdispersion:

$$d_{\text{obs}}(t) \sim \text{NegativeBinomial}(d(t), \phi_{\text{deaths}}), \tag{22}$$

where we use the alternative parameterisation of the negative binomial distribution as defined by the Stan Development Team [13]:

$$\text{NegativeBinomial}(\mu, \phi) = \binom{n + \phi - 1}{n} \left(\frac{\mu}{\mu + \phi}\right)^n \left(\frac{\phi}{\mu + \phi}\right)^\phi. \tag{23}$$

In Equation 22, $d(t)$ is the difference between the population of the D state of the transmission model between days $t - 1$ and t: $d(t) = D(t) - D(t - 1)$.

2.2.2 Hospital Admission Data

The UK government also publishes a daily time series of the number of COVID-19 patients admitted to hospital on their official website for coronavirus data [12]. We assume that, like the number of deaths, the number of hospital admissions on day t, $h_{\text{obs}}(t)$, has a negative binomial distribution parameterised by $h(t)$ and $\phi_{\text{admissions}}$:

$$h_{\text{obs}}(t) \sim \text{NegativeBinomial}(h(t), \phi_{\text{admissions}}), \tag{24}$$

where

$$h(t) = \rho_{\text{admissions}}(t) \times \frac{2}{d_I} I_2, \tag{25}$$

i.e. the mean number of hospital admissions on day t equals the ratio of hospital admissions to potential patients, $\rho_{\mathrm{admissions}}(t)$, multiplied by the number of new members of the pending state. $\rho_{\mathrm{admissions}}(t)$ is a continuous piecewise linear function of time:

$$\rho_{\mathrm{admissions}}(t) = \sum_{k=1}^{K} \rho_{\mathrm{admissions},k}(t)\chi_{[t_{k-1},t_k)}(t), \tag{26}$$

where the ratio of hospital admissions to potential patients during the kth time interval, $\rho_{\mathrm{admissions},k}(t)$, is given by

$$\rho_{\mathrm{admissions},k}(t) = \frac{\rho_{\mathrm{admissions},k+1} - \rho_{\mathrm{admissions},k}}{t_k - t_{k-1}}(t - t_{k-1}) + \rho_{\mathrm{admissions},k}, \tag{27}$$

and the indicator function for the kth time interval, $\chi_{[t_{k-1},t_k)}(t)$, is defined in Equation 21.

The parameters $\rho_{\mathrm{admissions},1}$, $\rho_{\mathrm{admissions},2}$, ..., $\rho_{\mathrm{admissions},K+1}$ in Equation 27 are the values $\rho_{\mathrm{admissions}}(t)$ takes on a set of predefined dates t_0, t_1, ..., t_K. The first date is the 24th of March 2020, and each date that follows is twelve weeks after the last, with the second date being the 16th of June 2020.

2.2.3 Symptom Report Data

Every weekday up to the previous calendar day, NHS Digital publishes a daily time series of the number of assessments completed through the NHS 111 telephone service where callers reported potential coronavirus (COVID-19) symptoms [14]. Leclerc et al. found a strong correlation between the volume of these symptom reports and the number of COVID-19 deaths reported 16 days later [15]. We assume that, like the other types of surveillance data, the number of assessment calls to NHS 111 on day t where callers reported potential COVID-19 symptoms, $c_{\mathrm{obs}}(t)$, has a negative binomial distribution parameterised by $c(t)$ and ϕ_{calls}:

$$c_{\mathrm{obs}}(t) \sim \mathrm{NegativeBinomial}\left(c(t), \phi_{\mathrm{calls}}\right), \tag{28}$$

where

$$c(t) = \rho_{\mathrm{calls}}(t) \times \left(\frac{2}{d_L}E_2 + \frac{2}{d_I}I_2\right), \tag{29}$$

i.e. the mean number of assessment calls to NHS 111 on day t where callers reported potential COVID-19 symptoms equals the ratio of symptom reports to potential symptom reporters, $\rho_{\mathrm{calls}}(t)$, multiplied by the sum of the number of new members of the infectious and pending states.

$\rho_{\mathrm{calls}}(t)$ is a continuous piecewise linear function of time almost identical to $\rho_{\mathrm{admissions}}(t)$, which is defined by Equations 26 and 27. The only difference is that the parameters $\rho_{\mathrm{calls},1}$, $\rho_{\mathrm{calls},2}$, ..., $\rho_{\mathrm{calls},L+1}$ are associated with a different set of predefined dates t_0, t_1, ..., t_L. The first of these dates is the 24th of March 2020, and each date that follows is four weeks after the last, with the second date being the 16th of March 2020.

3 Scoring Rules

Scoring rules produce real numbers, also called numerical scores, that summarise the quality of probabilistic forecasts. More concretely, consider a probabilistic forecast P of an uncertain future quantity X for which the observation x eventually materialises. In a scenario such as this, a scoring rule provides a numerical score $s(P, x)$ that quantifies the statistical consistency between the predictive distribution P and the observation x. Table 1 shows the simple scoring rules that we use in this paper.

The logarithmic, quadratic, spherical, ranked probability, Dawid-Sebastiani, and squared error scores in Table 1 are negatively oriented, with better forecasts typically resulting in lower scores.

Table 1: The set of simple scoring rules that feature in this paper. In the definitions, p_x is the probability mass of the predictive distribution for an observed count x, $\|p\|^2 = \sum_{k=0}^{\infty} p_k^2$, P_k is the value of cumulative predictive distribution for a count k, $\mathbf{1}(.)$ is the indicator function, and μ_P and σ_P^2 are the mean and variance of the predictive distribution.

Scoring Rule	Definition	Reference
Logarithmic score	$\mathrm{logs}(P, x) = -\log p_x$	Good [16]
Quadratic score	$\mathrm{qs}(P, x) = -2p_x + \|p\|^2$	Wecker [17]
Spherical score	$\mathrm{sphs}(P, x) = -\frac{p_x}{\|p\|}$	Czado, Gneiting and Held [18]
Ranked probability score	$\mathrm{rps}(P, x) = \sum_{k=0}^{\infty} \{P_k - \mathbf{1}(x \leq k)\}^2$	Epstein [19]
Dawid-Sebastiani score	$\mathrm{dss}(P, x) = \left(\frac{x - \mu_P}{\sigma_P}\right)^2 + 2\log \sigma_P$	Gneiting and Raftery [20]
Squared error score	$\mathrm{ses}(P, x) = (x - \mu_P)^2$	Czado, Gneiting and Held [18]
Normalised squared error score	$\mathrm{nses}(P, x) = \left(\frac{x - \mu_P}{\sigma_P}\right)^2$	Carroll and Cressie [21]

These scores are also said to be proper in the sense that a forecaster minimises them when quoting their true belief. Proper scoring rules are considered essential for incentivising honest forecasting, a position argued by Gneiting and Raftery [20].

In contrast, the normalised squared error score is improper, a quality that has resulted in it being discredited, for example, by Czado, Gneiting and Held [18], as a tool for evaluating probabilistic forecasts. We argue that viewing it through the lens of propriety leads to an under-appreciation of its unique properties, particularly its ability to distinguish between over-confidence and over-caution. Indeed, we see the normalised squared error score as a valuable diagnostic tool with advantages over proper scoring rules in certain situations.

Interestingly, the normalised squared error score is popular in the tracking and data fusion community, which has studied performance measures for evaluating estimation algorithms, such as Li and Zhao's absolute [22] and relative [23] error measures. Blasch, Rice, and Yang [24], and Chen et al. [25] describe a now-popular relative error measure called the Normalised Estimation Error Squared (NEES), which is a generalisation of the normalised squared error score to multiple dimensions. Researchers in the community have used the NEES extensively as an easily understood approach to arguing the merits of, for example, different extensions of the Kalman filter to specific non-linear settings where the Extended Kalman Filter is routinely over-confident [26].

Scores are typically reported as averages over multiple probabilistic forecasts, each for a distinct point in time. Following Czado, Gneiting, and Held [18], we use uppercase to denote the mean score over several forecasts. The tables in this paper use the mean scores LogS, QS, SphS, RPS, DSS, SES and NSES.

4 Computational Experiments

4.1 Setup

We implement the statistical model described in Section 2 in Stan [8]. Stan is a probabilistic programming language that allows users to articulate statistical models and calibrate them with data using a Markov chain Monte Carlo (MCMC) method called the No-U-Turn Sampler (NUTS), proposed by Hoffman and Gelman [27]. Stan also provides diagnostic information, such as warnings about divergent transitions, to help users check it has sampled the posterior faithfully.

In addition to encoding the statistical model, the Stan implementation includes prior distributions for the model parameters. Table 2 gives details of the prior distributions we have chosen to use in the computational experiments, along with the reasoning behind their selection. Four of the prior distributions, for the parameters d_L, d_I, d_P and ω, are based on estimates from the published literature on COVID-19. We base the prior distributions for d_L and d_I on estimates of the time variables rather than estimates of their mean values, for which we could not identify reliable estimates. This decision results in looser prior distributions and, in turn, wider posterior distributions for these two parameters.

The software implementation, which is publically available on GitHub[1], has two idiosyncrasies worthy of discussion. First, the current implementation does not use any of the integrators provided by Stan to solve the system of ordinary differential equations (ODEs) in Equation 10. Instead, a bespoke implementation of the explicit trapezoidal method [28] solves the system of ODEs for the transmission model. Anecdotally, the trapezoidal integrator significantly reduces runtime while producing acceptable numerical errors. Second, the current implementation only uses Stan's default initialisation strategy for $1/\phi_{\text{deaths}}$, $1/\phi_{\text{admissions}}$, $1/\phi_{\text{calls}}$, $\rho_{\text{admissions},1}$, ..., $\rho_{\text{admissions},K+1}$, $\rho_{\text{calls},1}$, ..., and $\rho_{\text{calls},L+1}$ by drawing values uniformly between -2 and 2 on the unconstrained parameter space. Rather than doing this for α_1, α_2, β_1, ..., β_J, d_L, d_I, d_P, and ω, the implementation draws uniformly from custom intervals to prevent initialisation failures caused by unrealistic parameter values.

We calibrate the model with data for England by using NUTS to draw six independent Markov chains for each of seven smoothness hyperparameter, σ_β, values. Each chain draws 512 samples and discards the first 256 drawn during warmup. We calibrate over the period from the 24th of March 2020 to the 31st of December 2020 with the death, hospital admission, and symptom report data described in Section 22.2. Each calibration job produces a posterior distribution for the statistical model's parameters, shown in Table 2. For each job, we generate two posterior predictive distributions for the daily number of deaths by simulating from the statistical model, first using the posterior samples for the parameters and second using only a point estimate for the parameters. We perceive that there are occasions where the use of point estimates can explain over-confident estimates. Specifically, we use the mean of the posterior samples as the point estimate for the parameters. The posterior predictive distributions span the 17th of February 2020 to the 21st of January 2021, with the last twenty-one elements being three-week forecasts for which we calculate the LogS, QS, SphS, RPS, DSS, SES and NSES.

4.2 Results

Mean scores for the forecasts that we generated with posterior samples for the parameters and point estimates for the parameters are presented in the top and bottom of Table 3, respectively. The columns contain results for different values of the smoothness hyperparameter σ_β defined in Table 2. Smaller values of σ_β make the random-walk prior on the effective contact rate $\beta(t)$ tighter, causing it to vary more slowly and, if low enough, to underfit the data. Conversely, larger values of the smoothness hyperparameter loosen the random-walk prior and allow overfitting of the data if σ_β is high enough.

Generally, the mean scores in Table 3 are lower, or closer to one in the case of NSES, for the forecasts generated with posterior samples for the parameters than for those generated with

---

Table 2: Prior distributions for the parameters of the statistical model with the rationale for their selection. We use the symbol $^+$ to indicate a distribution with its lower tail truncated at zero.

Parameter(s)	Prior Distribution	Comment
α_1	$\text{Beta}(5.0, 0.5)$	This reflects our belief that most of the population is initially susceptible.
α_2	$\text{Beta}(1.1, 1.1)$	This reflects our uninformed beliefs about the initial division of the infected population into those who are infectious and those who are not.
β_1	$\text{HalfNormal}(0.0, 0.5)$	This is a generic, weakly informative prior inspired by the work of Gelman [29].
$\beta_2, ..., \beta_J$	$\text{Normal}^+(\beta_{i-1}, \sigma_\beta)$	This is random-walk prior with smoothness hyperparameter σ_β that enforces correlation between contiguous β_i.
d_L	$\text{Normal}^+(4.0, 3.0)$	This is based on an estimate of the incubation period provided by Pellis et al. [30].
d_I	$\text{Normal}^+(5.0, 4.0)$	This is based on an estimate of the delay from onset of symptoms to hospitalisation provided by Pellis et al. [30].
d_P	$\text{Normal}^+(13.0, 4.0)$	This is based on an estimate of the mean delay from hospitalisation to death provided by Linton et al. [31].
ω	$\text{Beta}(5.7, 624.1)$	This is based on an estimate provided by Ward et al. [32].
$\frac{1}{\phi_{\text{deaths}}}, \frac{1}{\phi_{\text{admissions}}}, \frac{1}{\phi_{\text{calls}}}$	$\text{Exponential}(5.0)$	This is a containment prior for the overdispersion parameter, which is discussed by Simpson [33].
$\rho_{\text{admissions},k}, \rho_{\text{calls},l}$	$\text{Beta}(1.1, 1.1)$	This reflects our uninformed beliefs about these ratio parameters.

Table 3: Mean scores for the three-week forecasts generated with different σ_β values. Top: Simulating from the statistical model with a point estimate for the parameters. Bottom: Simulating from the statistical model with posterior samples for the parameters.

Scoring Rule	σ_β						
	0.0005	0.001	0.0025	0.005	0.01	0.025	0.05
				Point estimate			
LogS	9.595	7.403	9.212	9.711	9.752	9.260	10.000
QS	0.002	-0.001	0.002	0.004	0.005	0.006	0.990
SphS	-0.001	-0.027	-0.005	-0.001	0.000	-0.004	0.000
RPS	693.406	176.730	349.403	202.763	201.986	124.211	0.000
DSS	27.478	12.367	29.167	28.042	28.184	24.997	209338988
SES	665563	81671	268904	89048	89110	33435	1055387
NSES	16.908	1.689	19.666	19.743	19.895	17.410	209338994
				Posterior samples			
LogS	9.624	7.313	9.195	6.988	6.949	**6.076**	6.200
QS	0.002	0.000	0.002	0.000	0.000	-0.003	**-0.004**
SphS	-0.001	-0.024	-0.005	-0.025	-0.025	-0.052	**-0.058**
RPS	696.828	181.678	393.543	123.172	122.706	**50.085**	74.223
DSS	27.452	12.356	25.139	11.797	11.776	**9.992**	10.440
SES	669959	85439	261226	37660	37607	**4690**	44607
NSES	16.880	1.581	15.464	2.399	2.372	**0.340**	0.276

point estimates. This observation highlights the importance of allowing uncertainty propagation from statistical inference to forecasting. The differences between the values produced by the two forecasting methods are small for the lowest smoothness hyperparameter value of 0.0005 but increase with σ_β until they become large and pronounced.

We have emboldened the best value for each of the mean scores, corresponding to the lowest value for the mean proper scores, LogS, QS, SphS, RPS, DSS, and SES, and the value closest to and less than one for NSES. We select the best NSES value in this way because values of less than one correspond to over-cautious forecasts, which are arguably less damaging in terms of their impact on decision making than over-confident forecasts, which have NSES values of greater than one. The majority of the best mean scores are for the posterior samples for the parameters and a σ_β value of 0.025. We can see that this forecast, shown in Figure 2 (a), is over-cautious, which the NESE value correctly summarises. We can also see the forecast for the point estimate for the parameters in Figure 2 (a). This forecast is biased because the mean resides outside the posterior distribution's region of high probability mass, as it does for the Hybrid Rosenbrock distribution [34]. Although most of the best mean scores are for a σ_β value of 0.025, the best values for the QS and SphS are for a smoothness hyperparameter value of 0.05. The differences, however, between the mean scores, 0.001 for QS and 0.006 for SphS, are relatively small and indicate little difference between the quality of the two forecasts when assessed with QS and SphS.

The forecast generated with the posterior samples for the parameters and a σ_β value of 0.005 has a NSES value of 2.399, which indicates that it is over-confident. We can see in Figure 2 (b) that the forecast fails to predict most of the future observations and is indeed over-confident. The other scoring rules do not provide any information about the over-confidence of this forecast. We, therefore, believe that NSES's ability to distinguish between over-confidence and over-caution, given a single forecast, makes it a valuable diagnostic tool that should be used alongside proper scoring rules.

5 Conclusion

We have shown how to use simple scoring rules to develop a statistical model and improve its forecasting performance. The computational experiments presented in Section 4 demonstrate that the statistical model introduced in Section 2 provides the best forecasts when we use posterior samples for the parameters and a smoothness hyperparameter σ_β value of 0.025. We, therefore, advocate simple scoring rules for evaluating epidemiological forecasts and NSES specifically to establish if they are over-confident or over-cautious.

One of the significant limitations of simple scoring rules is that we can only use them to assess forecasts of observable variables. Epidemiological modellers cannot apply them to important latent quantities, such as the effective reproduction number R_t and the growth rate r, which they often forecast. Accordingly, the epidemiological community needs a method for assessing forecasts of quantities for which the truth is unknown. Simulation-Based Calibration (SBC) [35] is a candidate method for this task, worth investigating further.

There are four worthwhile directions in which we can extend the statistical model presented in Section 2. The first direction involves adding components to the observation model described in Section 2.2 to allow calibration with a greater quantity and diversity of surveillance data. The second direction involves modifying it to accommodate surveillance data from other countries. The third direction entails making the disease-specific, geography-independent parameters d_L, d_I, d_T, and ω global to facilitate information sharing between regions, as is done by Birrell et al. [1]. The fourth and final direction involves removing the assumption that recovered individuals are indefinitely immune to reinfection, allowing reinfection, which is more realistic.
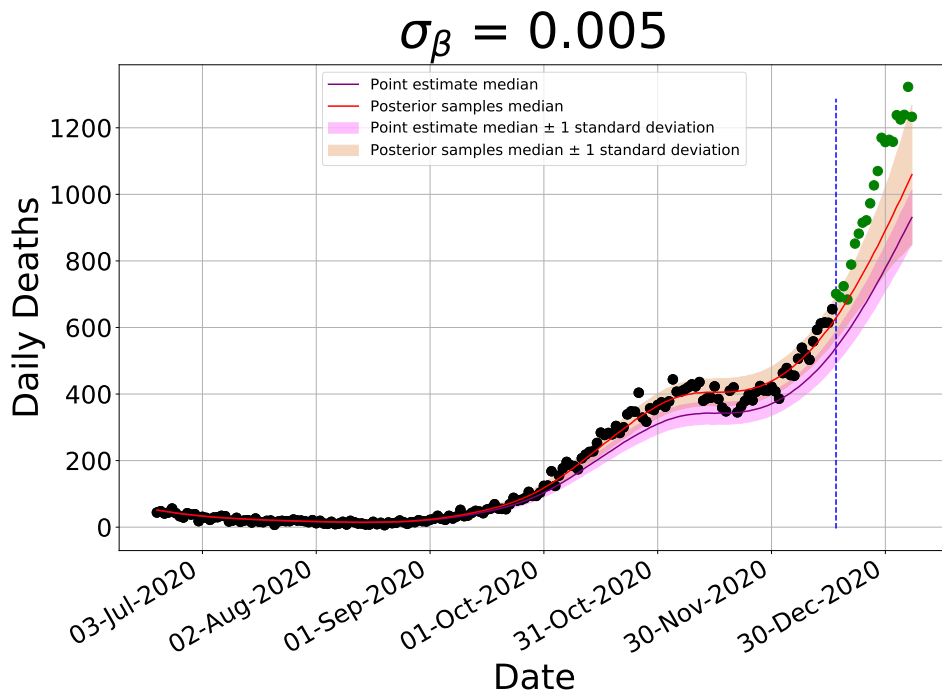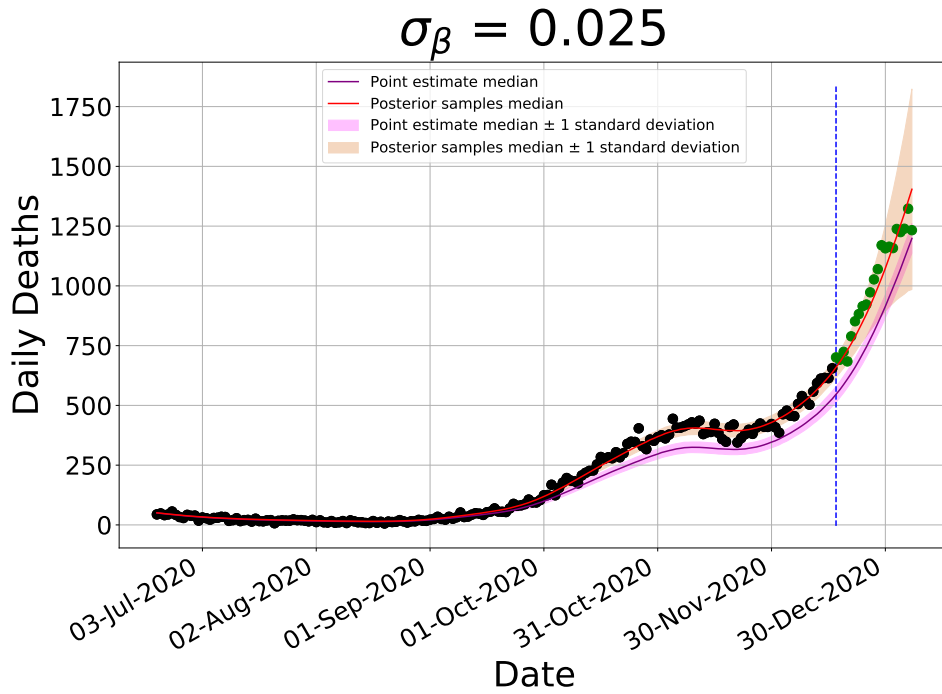
Figure 2: Forecasts generated with point estimates and posterior samples for smoothness hyper-parameter, σ_β, values of (a) 0.025 and (b) 0.005.

Data Accessibility

The code and data for the computational experiments are publically available on GitHub (https://github.com/codatmo/UniversityOfLiverpool_PaperSubmission).

Authors' Contributions

REM designed and implemented the statistical model, participated in the analyses and drafted the manuscript; CR carried out the computational experiments, participated in the analyses and helped draft the manuscript; SM conceived and supervised the work, participated in the analyses and critically revised the manuscript. All authors gave final approval for publication and agree to be held accountable for the work performed therein.

Competing Interests

We declare that we have no competing interests.

References

[1] Birrell P, Blake J, van Leeuwen E, Gent N, De Angelis D. 2021 Real-time nowcasting and forecasting of COVID-19 dynamics in England: the first wave. Phil. Trans. R. Soc. B 376: 20200279.

[2] Keeling MJ, Dyson L, Guyver-Fletcher G, Holmes A, Semple MG, ISARIC4C Investigators, Tildesley M, Hill EM. 2020 Fitting to the UK COVID-19 outbreak, short-term forecasts and estimating the reproductive number. medRxiv.

[3] Abbott S, Hellewell J, Thompson RN et al. 2020 Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts [version 2; peer review: 1 approved with reservations]. Wellcome Open Res, 5:112.

[4] Cramer EY, Lopez VK, Niemi J, George GE, Cegan JC, Dettwiller ID, England WP, Farthing MW, Hunter RH, Lafferty B, Linkov I. 2021 Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US. medRxiv.

[5] Funk S, Abbott S, Atkins BD, Baguelin M, Baillie JK, Birrell P, Blake J, Bosse NI, Burton J, Carruthers J, Davies NG. 2020 Short-term forecasts to inform the response to the Covid-19 epidemic in the UK. medRxiv preprint-BMJ Yale.

[6] Maishman T, Schaap S, Silk DS, Nevitt SJ, Woods DC, Bowman VE. 2021 Statistical methods used to combine the effective reproduction number, R (t), and other related measures of COVID-19 in the UK. arXiv preprint arXiv:2103.01742.

[7] CoDatMo. 2021 Welcome to the CoDatMo site. See https://codatmo.github.io (accessed 9 September 2021).

[8] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. 2017 Stan: A Probabilistic Programming Language. Journal of Statistical Software, 76(1), 1 - 32.

[9] Storopoli J, Santos ALMF, Pellini ACG, Baldwin B. 2021 Simulation-Driven COVID-19 Epidemiological Modeling with Social Media. arXiv, 2106.11686.

[10] UK Government. 2021 Reproduction number (R) and growth rate: methodology. See https://www.gov.uk/government/publications/reproduction-number-r-and-growth-rate-methodology/reproduction-number-r-and-growth-rate-methodology (accessed 21 September 2021).

[11] Kermack WO and McKendrick AG. 1927 A contribution to the mathematical theory of epidemics. Phil. Trans. R. Soc. A.

[12] UK Government. 2021 Coronavirus (COVID-19) in the UK. See https://coronavirus.data.gov.uk (accessed 13 September 2021).

[13] Stan Development Team. 2021 Negative binomial distribution (alternative parameterization). See https://mc-stan.org/docs/2_27/functions-reference/nbalt.html (accessed 13 September 2021).

[14] NHS Digital. 2021 Potential Coronavirus (COVID-19) symptoms reported through NHS Pathways and 111 online. See https://digital.nhs.uk/data-and-information/publications/statistical/mi-potential-covid-19-symptoms-reported-through-nhs-pathways-and-111-online/latest (accessed 14 September 2021).

[15] Leclerc QJ, Nightingale ES, Abbott S, CMMID COVID-19 Working Group, Jombart T. 2021 Analysis of temporal trends in potential COVID-19 cases reported through NHS Pathways England. Sci Rep 11, 7106.

[16] Good IJ. 1952 Rational decisions. Journal of the Royal Statistical Society, Series B, Methodological 14, 107-114.

[17] Wecker WE. 1989 Assessing the Accuracy of Time Series Model Forecasts of Count Observations. Journal of Business and Economic Statistics 7, 418-419.

[18] Czado C, Gneiting T and Held L. 2009 Predictive Model Assessment for Count Data. Biometrics, 65(4), pp.1254-1261.

[19] Epstein ES. 1969 A Scoring System for Probability Forecasts of Ranked Categories. Journal of Applied Meteorology 8, 985-987.

[20] Gneiting T and Raftery AE. 2007 Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association 102, 359-378.

[21] Carroll SS and Cressie N. 1997 Spatial modeling of snow water equivalent using covariances estimated from spatial and geomorphic attributes. Journal of Hydrology 190, 42-59.

[22] Li XR and Zhao Z. 2001 Measures of performance for evaluation of estimators and filters. Proc. SPIE 4473, Signal and Data Processing of Small Targets 2001.

[23] Li XR and Zhao Z. 2005 Relative error measures for evaluation of estimation algorithms. 7th International Conference on Information Fusion, 2005, pp. 8 pp.-.

[24] Blasch EP, Rice A, Yang C. 2006 Nonlinear tracking evaluation using absolute and relative metrics. Proc. SPIE 6236, Signal and Data Processing of Small Targets 2006, 62360L.

[25] Chen Z, Heckman C, Julier S, Ahmed N. 2018 Weak in the NEES?: Auto-Tuning Kalman Filters with Bayesian Optimization. 21st International Conference on Information Fusion (FUSION), 2018, pp. 1072-1079.

[26] Longbin M, Xiaoquan S, Yiyu Z, Kang SZ and Bar-Shalom Y. 1998 Unbiased converted measurements for tracking. IEEE Transactions on Aerospace and Electronic Systems, vol. 34, no. 3, pp. 1023-1027.

[27] Hoffman MD, Gelman A. 2014 The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. 15: 1593–1623.

[28] Ascher UM, Petzold LR. 1998 Computer methods for ordinary differential equations and differential-algebraic equations. Siam.

[29] Gelman A. 2020 Prior Choice Recommendations. See https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations (accessed 16 September 2021).

[30] Pellis L et al. 2021 Challenges in control of COVID-19: short doubling time and long delay to effect of interventions. Phil. Trans. R. Soc. B 376: 20200264.

[31] Linton NM, Kobayashi T, Yang Y, Hayashi K, Akhmetzhanov AR, Jung S-m, Yuan B, Kinoshita R, Nishiura H. 2020 Incubation Period and Other Epidemiological Characteristics of 2019 Novel Coronavirus Infections with Right Truncation: A Statistical Analysis of Publicly Available Case Data. Journal of Clinical Medicine. 2020; 9(2):538.

[32] Ward H, Atchison C, Whitaker M, Ainslie KE, Elliott J, Okell L, Redd R, Ashby D, Donnelly CA, Barclay W, Darzi A. 2021 SARS-CoV-2 antibody prevalence in England following the first peak of the pandemic. Nature communications. 2021 Feb 10; 12(1):1-8.

[33] Simpson D. 2018 Justify my love. See https://statmodeling.stat.columbia.edu/2018/04/03/justify-my-love/ (accessed 16 September 2021).

[34] Pagani F, Wiegand M, Nadarajah S. 2021 An n-dimensional Rosenbrock distribution for Markov chain Monte Carlo testing. Scandinavian Journal of Statistics.

[35] Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A. 2020 Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv preprint arXiv:1804.06788.