Manuscript File

Click here to view linked References

7. Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling

(IDP Challenges Series)

Authors: Ben Swallow[1,2], Paul Birrell[3], Joshua Blake[3], Mark Burgman[4], Peter Challenor[5], Luc E. Coffeng[6], Philip Dawid[7], Daniela De Angelis[3], Michael Goldstein[8], Victoria Hemming[9], Glenn Marion[10,2], Trevelyan J. McKinley[11], Christopher Overton[12,13,14], Jasmina Panovska-Griffiths[15,16], Lorenzo Pellis[12,14,17], Will Probert[15], Katriona Shea[18], Daniel Villela[19], Ian Vernon[8]

Author affiliations:

[1]School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

[2]Scottish COVID-19 Response Consortium, UK, www.gla.ac.uk/scrc/

[3]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[4]Centre for Environmental Policy, Imperial College London, London, UK

[5]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

[6]Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands

[7]Statistical Laboratory, University of Cambridge, Cambridge, UK

[8]Department of Mathematical Sciences, Durham University, Stockton Road, Durham, UK.

[9]Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada

[10]Biomathematics and Statistics Scotland, Edinburgh, UK

[11]College of Medicine and Health, University of Exeter, Exeter, UK

[12]Department of Mathematics, University of Manchester, Manchester, UK

[13]Clinical Data Science Unit, Manchester University NHS Foundation Trust, Manchester, UK

[14]Joint UNIversities Pandemic and Epidemiological Research, UK

[15] The Big Data Institute, University of Oxford, Oxford, UK

[16]The Queen's College, University of Oxford, Oxford, UK

[17]The Alan Turing Institute, London, UK

[18]Department of Biology and Centre for Infectious Disease Dynamics, The Pennsylvania State University, PA 16802, USA

[19]Program of Scientific Computing, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

Abstract: The estimation of parameters and model structure for informing infectious disease response has become a focal point of the recent pandemic. However, it has also highlighted a plethora of challenges remaining in the fast and robust extraction of information using data and models to help inform policy. In this paper, we identify and discuss four broad challenges in the estimation paradigm relating to infectious disease modelling, namely the Uncertainty Quantification framework, data challenges in estimation, model-based inference and prediction, and expert

judgement. We also postulate priorities in estimation methodology to facilitate preparation for future pandemics.

Key words: statistical estimation, uncertainty quantification, expert elicitation, pandemic modelling.

1. Introduction

Efficient and timely estimation in parametric models of epidemiological processes for real-world systems is highly challenging, but fundamental to scientific understanding, forecasting and decision-making under uncertainty (Shea et al. (2020)). There are different dimensions to the estimation paradigm that can be conducted independently, including parameter estimation, quantification of uncertainty and sensitivity and model structure uncertainty, but ideally should be united in a single coherent framework due to their dependence on each other. Estimation approaches should incorporate all major sources of uncertainty, otherwise estimates may be biased and/or overly precise. Key sources of uncertainty include inherent variation in natural systems and our lack of knowledge about these systems, typically broken down into: observation error or bias (where the process of data collection is imperfect); stochastic uncertainty (where inherent randomness in the transmission process impacts outcomes of interest); parameter uncertainty (where data are insufficient to fully identify model inputs); structural uncertainty (where the choice of model structure is unknown); and model discrepancy (reflecting differences between the reality and the mathematical approximation to it that the model provides). Adequate treatment of uncertainty increases robustness of forecasts, predictions and decisions, facilitating a robust description and understanding of the processes involved. The uncertainty estimates can either be a natural by-product of statistical inference procedures, or a quantity of statistical interest in their own right.

Statistical inference for mechanistic infectious disease models is challenging for many reasons, which has been discussed extensively in Lloyd-Smith et al. (2015) and accompanying papers. Chief amongst these is the fact that the transition processes (e.g., transmission, recovery etc.) depend on the numbers of individuals in each epidemiological state at any given time. In practice, these are only partially observed. For example, infection times must be inferred from events such as onset-of-symptoms, which are also uncertain and recorded with error. These issues are exacerbated by asymptomatic infections, as seen often seen in infectious diseases, including recently for Covid-19. Therefore, statistical methods are combined with data to infer these missing variables alongside parameter values in the underlying transmission model. Especially in the case of emerging diseases, typically it is also unclear how to structure models e.g., in terms of disease progression, or what spatial and temporal heterogeneities should be accounted for (Marion et al. (2021)). Therefore, and regardless of whether a model is deterministic or stochastic, statistical inference is used to quantify uncertainty in model structures, assess and select models, and handle multi-model ensembles. If these models are used to support decisions, then these challenges also need to be addressed in real-time.

Deterministic, state-and-transition transmission models can be fitted relatively efficiently to data, by assuming transitions between states are a continuous process, ignoring intrinsic uncertainty associated with the underlying epidemiological history. Methods such as least-squares fitting are often used to find a set of input parameters that minimise the residual error between simulated event curves and observed data. More sophisticated methods, such as using explicit stochastic observation processes that account for discrepancies between the simulated event curves and the observed data points can also be used to construct likelihood functions that (depending on how they are implemented) can produce exact inference for a given transmission/observation model (McKinley et al. (2018)). However, deterministic models are at best an approximation to the average behaviour of an underlying stochastic system, and as such are applicable only in certain scenarios, for example, with high infection levels in large, well-mixed populations. In highly heterogeneous populations, such as

those with spatial or network structures (Eames et al. (2015)), these models are less appropriate, or indeed when the numbers of infections are low, then predictions from these models can deviate dramatically from their stochastic counterparts.

Stochastic transmission models offer more realism at the cost of significant increases in computational complexity. Here events are modelled probabilistically. For example, models of infections such as foot-and-mouth disease or E coli might choose to model transmission between herds, or alternatively at the individual animal-level, with coupled processes modelling within- and between-herd spread (Touloupou et al. (2020)). Similar considerations apply to human diseases. Some frameworks model individual-level interactions, while others model transmission among and between groups, such as meta-population models. Since transmissions are rarely observed, the amount of missing information that needs to be imputed in the inference process is linked to the model, so that an individual-based model for every individual in the UK would correspond to many millions of unobserved stochastic events, making inference and predictions highly computationally intensive.

It is clear that there are multiple challenges to developing timely epidemiological models. One challenge that seems common to all approaches is the need to develop infrastructure to conduct more comprehensive uncertainty analyses in real-time, whether through availability of more efficient algorithms, general software, computational power or knowledge and expertise. This in turn will facilitate urgent decision-making, so simple and fast estimation procedures will remain desirable. In all circumstances, decisions must be made in the face of considerable uncertainty and often at speed, and this uncertainty needs to be communicated effectively to enhance decision making by those (typically non-quantitative experts) responsible. Thus, uncertainty quantification also presents challenges for expert elicitation, and communication (including visualisation).

In this paper, to prepare for future pandemics, we highlight a series of key challenges pertaining to estimation, uncertainty quantification and expert elicitation that are relevant to pandemic modelling. In section 2, we outline challenges in the Uncertainty Quantification paradigm for estimation of uncertainties and sensitivities coupled with model calibration for large-scale pandemic models. Section 3 identifies challenges of using real-world data in estimation procedures in real-time. Section 4 suggests challenges for parameter estimation and model selection in pandemic modelling, and finally, Section 5 discusses the challenges of using expert judgement in pandemics when evidence and data are less readily available than is required by the models.

2. UQ for estimation

As mentioned above, one of the principal aims of estimation is to measure and account for the various aspects of potential bias and uncertainty inherent in the mathematical and statistical modelling of real-world systems. We begin by discussing the UQ framework, which in its fullest interpretation is a formal set of statistical methodologies accounting for the discrepancies present in the use of computer models to represent the real world, and their associated calibration to data and forecasting for future outcomes. Aspects of UQ are applicable at all stages of the modelling process, specifically pre-, during- and post-pandemic, and can therefore underpin or inform the sections that follow.

Simulators and emulation

The mathematical and statistical analysis of complex numerical models or simulators and their connection to the real world, is often referred to as Uncertainty Quantification (UQ). Although the modelling of pandemics faces clear challenges that could be addressed by using these methods, with a few exceptions (Andrianakis et al. (2015); McCreesh et al. (2017); McKinley et al. (2018); Gugole et al (2021)), there has been little application of UQ methodology to epidemic models. A major such challenge is that of estimation, in UQ often referred to as model calibration (sometimes model tuning). Nonetheless, the problem being solved is the same: can real-world data corresponding to model outputs (say hospital admissions) be used to tell us something about the model inputs

(transmission rates, say), and how can this be achieved efficiently within a coherent framework that incorporates all appreciable sources of uncertainty?

One of the main tools employed in complex UQ tasks is an emulator, often a Gaussian (or second order) process. A Gaussian Process (GP) is a stochastic process that gives smooth continuous functions that can be fitted to model runs as a surrogate for the true (unknown) analytical solution to the model. The key here is speed: such GP emulators are typically several orders of magnitude faster to evaluate than the epidemiological model they are mimicking, and hence they facilitate otherwise infeasible UQ calculations, including a comprehensive exploration of the model's parameter (input) space and behaviour. A second substantial advantage of GPs over other possible surrogate models (such as polynomials) is that the GP includes an estimate of its own uncertainty. This can be formally included in any subsequent calculation, inflating any uncertainty calculations to account for the fact that a surrogate model has been used rather than the true model. The fit of the GP and the validity of its estimated uncertainty can be tested using additional model runs (Bastos and O'Hagan, 2009). The GP emulator has many applications in the analysis of computer models, for instance predicting a new value (with uncertainty) and performing sensitivity and uncertainty analyses efficiently.

Sensitivity analysis

An additional stage of the UQ framework is sensitivity analysis, in which the impact of changes to inputs or parameters of the model on the outputs of that model are studied. This can be done as part of the model construction process (Marion et al. (2021)) but can also be useful in estimation. In particular, it can be useful in reducing the dimension of the estimation problem, by avoiding focus on parameters that have little importance for the model; in determining important parameters to focus estimation and calibration procedures on; or highlighting areas where data may be particularly useful in obtaining inference or uncertainty reduction. Frequently this sensitivity is not done as a routine part of the estimation procedure, meaning that time can subsequently be wasted on non-identifiable or nuisance parameters that are of little statistical interest. Sensitivity analyses of stochastic models also cause computational and algebraic challenges that can be prohibitive for their general uptake.

Calibration and history matching

One substantial difference between the UQ and more conventional estimation approaches is explicit acknowledgement that the model will never be a perfect representation of the real world, no matter what model parameters are used. This has profound implications. For example, simply using a method such as least squares with no discrepancy term will 'overfit' the model and have poor predictive performance. Including a structural model discrepancy term, in both the past and in the future, can result in vastly improved predictions. This solves two problems: overfitting in the past and being overly confident in the future. The inclusion of model discrepancy elevates the analysis from that of the model to the analysis of the real world itself and provides a (partial) defence against the question "Why should we use these models to make decisions?".

There are two current methods for calibrating models. The first builds an emulator for the model and an emulator for the discrepancy simultaneously (Kennedy and O'Hagan (2001)). If the interest is only in prediction, then the Kennedy and O'Hagan method works well, but there is an identifiability problem between the two emulators. Their sum can be estimated but the two components are difficult to separate (Brynjarsdottir and O'Hagan (2014)). Several solutions to this problem have been proposed but are subject to severe limitations.

An alternative is termed history matching (HM—Vernon et al. (2010)). HM aims to identify those inputs that give predicted model outputs so far removed from the data that they can be regarded as implausible. HM proceeds by producing and validating an emulator, that is trained on a carefully-designed set of model runs (using theory from optimal experimental design). Then the distance

between the data and emulated model output (called the implausibility) is calculated and scaled by three 'variance' terms: the emulator variance (which is known), the data variance (supplied by the data collector) and a model discrepancy term (elicited from the model developer, in combination with a series of carefully-designed experiments on the model, see section 5 below). If this implausibility is greater than a defined threshold the set of model inputs is ruled implausible. It is worth noting that the implausibility measure is a normalised unimodal variable, and as such these cut-off thresholds can be informed by theory, most notably Pukelsheim's three-sigma rule (Pukelsheim, (1994)). By adding extra model runs, as a new *wave*, inside the not ruled out yet (NROY) region, increasingly more accurate emulators can be produced, which reduce the NROY region further. Eventually, either the NROY space becomes so small that further reduction is unhelpful (adding extra waves makes no difference to the NROY space, and better data are needed to reduce it any further), or the NROY space vanishes as all sets of model inputs are implausible. The implication of the latter is that, regardless of the model inputs, the model cannot be made to agree with the data. Analysts then need either to find another model, or a higher tolerance value for the model discrepancy is required (Runge et al. (2016)). Common usage of conventional estimation methods can miss the fact that the model may not fit the data well. This is especially problematic because, as the number of model runs is increased, the estimated uncertainty on a bad fit can be reduced: in essence bad model fits can lead to misleadingly tight posterior distributions.

Such methods for model calibration from the UQ field offer many advantages over conventional estimation methods. The use of fast GP emulators allows the use of Monte Carlo or other sampling-based methods that would be impossible with a full model. The inclusion of model discrepancy in the calibration/estimation methodologies acknowledges that models are not perfect representations of the real world, in the same way that data are not—both contain biases and uncertainties.

Model discrepancy

Formal separation of model and reality within the UQ framework opens many further possibilities, including construction of an overarching framework that incorporates multiple epidemiological models in a coherent fashion (Goldstein and Rougier (2009)). This is virtually impossible without such structural model discrepancy terms. This framework allows the predictive power of multiple models to be combined coherently, while acknowledging their various strengths, weaknesses and differences. Similarly, fast, simple models (for which many runs can be evaluated to train the emulator with high accuracy) can be combined with slower, more detailed models (for which far fewer runs are available). Furthermore, these methods allow separation of the inference and simulation frameworks, so that the same techniques can be used to fit a wide range of different models, without having to make fundamental changes to the nature of the inference algorithm. Hence such a separation could represent a step change in epidemiological analysis.

The major challenges for using these approaches for real-time pandemic modelling are:

1. **Efficient Model Calibration support**: the provision of efficient and robust UQ methods and code to aid the epidemiologists' model calibration efforts. The efficiency is achieved via the use of emulation, allowing epidemiologists to calibrate current models, and to explore more complex/higher-dimensional models when needed.
2. **Acknowledging the difference between the model and reality**: calibration methods should be robust in the sense that they incorporate structural model discrepancy, and hence guard against the dangers of treating an imperfect model as perfect. They should also exhibit robustness to (mis-)specification of distributional forms in the likelihood and associated error structures.
3. **Scaling**: the current GP based emulators do not scale well to large numbers of parameters or outputs (unless treated independently). Appropriate methods exist when these parameters correspond to spatial fields or time series. Increasing the number of inputs via a hierarchy of

models, for example, by adding spatial effects to a non-spatial model as described in (5) below, is a possible simple solution.

4. **Uptake of these methods**: a substantial challenge is the paradigm shift required for the uptake of these methods. Going from traditional ideas of statistical model fitting to ideas such as using fast emulators or representing all major sources of uncertainty in and around the models, including the structural model discrepancy terms, is new and unsettling for the modelling community, so it is not surprising if take up is slow during a pandemic when time is short. It also requires that modellers become familiar with fitting and validating (GP) emulators, which are currently not widely taught. This is exacerbated by the lack of suitable easy-to-use software or a lack of familiarity with software that is available, an issue addressed by point 1). An expository paper is currently in preparation for publication in this series to assist with the adoption of these methods.

Other more sophisticated challenges, of no less importance are:

5. **Multilevel Model Emulation and Calibration**: the incorporation of multiple levels of fidelity of epidemiological model (e.g., using fast, medium and slow versions) within a UQ emulation and calibration framework. This, as described above, is the most efficient way to emulate and calibrate very detailed epidemiological models (Craig et al. (1997); Kennedy and O'Hagan (2000); Cumming and Goldstein (2010)).

6. **Coherent Overarching Structure for Combining Multiple Models**: the provision of techniques to aid the combination of models from multiple research groups into a coherent structure to give more powerful predictions and subsequent decision support, underpinned by more realistic uncertainty statements. While some progress has been made on this front during the SARS-CoV2 pandemic, far more must be done. Suitable UQ frameworks for this, are ready to be employed (Rougier et al. (2013); Goldstein and Rougier (2009)).

7. **Generalising UQ to Stochastic Models**: UQ methodology was traditionally designed with deterministic models in mind. While much of it has been generalised to stochastic models, a setting closer to traditional statistics where many more tools are available, key challenges remain, e.g., issues around bi-modality and quantile emulation in complex stochastic models, motivating further research into the set of requisite statistical methods.

We have focused here on estimation/calibration, but the above challenges and UQ solutions also pertain to the critical issues of prediction and decision support (Marion et al. (2021); Hadley et al. (2021)).

3. Data challenges for estimation during a pandemic

Mathematical modelling works by simulating historic behaviour to understand better the current behaviour of the system, which can be used to make estimates and future predictions. The level of uncertainty in estimates and model outputs depends on several aspects, often closely related to the data. In this section, we describe some key estimation challenges that arise from use of data available during a pandemic. This discussion is general but draws on experience of the SARS-CoV2 pandemic.

Data availability and indirectness

During a pandemic, particularly in the early stages, scarcity of data can make it challenging to fit models and estimate parameters. However, during these early stages, policy decisions must be made despite scarce data, requiring models and estimation to use the data available efficiently, typically entailing a compromise between model complexity and parsimony, to make best use of available data

whilst not running into issues of non-identifiability. As more data are collected, across multiple layers, models can be refined and complexity can be increased, if required. If models are non-identifiable in the early stages, further attention needs to be given to exploring the parameter space. This can be computationally intensive but is vital to ensure correct communication of limitations and uncertainty in estimation.

Typically, even when scarce, epidemiological data can inform indirectly on the transmission process; however, complex data imputation techniques are needed even in the presence of abundant data. A major challenge is computational complexity and time. Care is needed to assess how much information the data contain about the parameters of interest, to ensure that the data are driving estimates (Section 4).

Inferences of the transmission process may be biased by missing data. During the early stages of an epidemic, when outbreaks are spatially distinct, estimation of epidemiological parameters can be biased by factors such as travel out of outbreak areas (Overton et al. (2020)), which may result in cases being missed, or inconsistent reporting rates across spatial regions, leading to different estimates of relationships between observed data and the underlying epidemic.

Using multiple layers of data can help to reduce uncertainty, such as combining sequencing data with surveillance data to obtain more direct estimates of a chain of transmission events. In the SARS-CoV-2 pandemic, appearance of different strains brought the possibility of higher relative transmissibility. This is hard to measure without detecting cases among contacts of an infected person, which relies on contact tracing or sequencing data. Challenges here relate to both the availability of data and accounting for biases in these. For example, there may be no systematic testing, producing challenges in what data to calibrate to or test model predictions against. If it is not possible to collect these data within the necessary timeframe, the challenge arises of how to deal with biases in predictions that may depend on these missing data. Although data collection from contact tracing and contact patterns are continuously improving, challenges remain in how to estimate the level of risk associated with different types of contact (Kretzschmar et al. (2021).

The pandemic has given rise to may new sources of data, each bringing their own challenges in estimation. One example is the use of phone apps that allow users to submit symptoms or movement activities on a daily basis. These data provide resolution that would not be possible through more direct experimental designs, but such 'community/citizen science' data is known to come with many issues in potential biases (Dickinson et al. (2010)). The use of waste water to sample for genetic viral material has also come to light, having previously been used to detect presence or absence of polio (O'Reilly et al. (2010)). Individual host variation in shedding is a specific challenge in developing more accurate prevalence of infection in populations, as is the tracking of the original source of the genetic material.

Even when the right type of data is available in sufficient quantities, it might not be at the correct resolution. For example, most mathematical epidemiology is based on continuous-time models, but data are always discrete, so a choice of whether to use a discrete-time model or how to discretise a continuous-time model is important. Continuous-time models may help with issues of censored data (see below). Similarly, time series data could be weekly rather than daily or fluctuate based on weekly reporting patterns, so the choice of how to aggregate or smooth data will affect estimation, requiring models that are robust to these systematic data issues. Resolution can also affect definitions at data used, such as whether to count all deaths where the patient tested positive for a pathogen, or only those where it was the primary cause of death. Discrepancies across regions can make it hard to estimate consistent fatality rates. Similarly, hospital occupancy data may count occupancy from time of admission or from time of returning a positive test, which can lead to challenges in estimating length of stay. To address these issues, better meta-data is needed to provide clarity into the definitions used. Data missingness can substantially affect the benefits of high-resolution data. For example, during the COVID-19 pandemic, high volumes of testing data have

been collected. However, high levels of missingness in the numbers of negative cases make the data challenging to use, due to changes in testing rates over time (Shadbolt et al. (2021)).

Noisy data, truncation and aggregation

Noisy signals arise from imperfections in observations and fluctuations in natural and human-mediated processes, requiring models to separate trends from residual effects. Aggregating over short time scales is prone to significant noise or delays, but if a signal is strong enough, the increased resolution may increase the usefulness of estimates. Aggregating over longer time scales can provide more stable estimates and less uncertainty, but estimates are affected by older data points so signal can be more "delayed", and rapid changes in signal can be missed. It is important to determine a suitable balance between flexibility and timeliness of estimates, and robustness and reliability of such estimates.

Lack of information due to gaps in data in space and/or time creates uncertainty in data streams. In these cases, imputation or smoothing between points relies on good understanding of biological processes to avoid introducing bias resulting from poor mechanistic representation and model discrepancy. Attention should be given to ensure that information is not being lost in the interpolation – for example on behaviour from mobility data if smoothing the relevant curve or from aggregating time series data (all cases vs age or risk-group stratified data). When an outbreak is unmitigated, such aggregation may be reasonable since the relative contribution across different units may be constant. However, for example, interventions may affect spatial or demographic groups differently.

The choice of aggregation level reflects which sources of heterogeneity are considered (Marion et al. (2021)). Many parameters, such as symptom duration and outcome probabilities, vary substantially with factors such as age, sex, socio-economic context or ethnicity. Aggregation across multiple covariates provides bigger sample cohorts, so estimates can be generated with seemingly lower uncertainty. If important covariates are not accounted for, estimated trends may be misleading. For example, data might suggest temporal changes in some parameter estimates that are driven by demographic changes over time. Data may be aggregated at a regional or national level, but this may fail to capture local heterogeneity, and local outbreaks might be very severe even if other areas are still apparently unaffected. However, disaggregating with multiple covariates may result in small sample sizes, inflating uncertainty, which could cause identifiability issues if estimates are used as model inputs.

During a pandemic, reporting events such as the transition from infection to hospital admission (Pellis et al. (2020)) or from hospital admission to death is often subject to significant delays. This leads to many observations being incomplete, lacking information regarding the duration of the delay and which outcome is observed. Such bias needs to be carefully adjusted for when estimating key epidemiological parameters (Commenges (1999)). It is possible only to consider cases where all events of interest have been observed. However, this introduces a truncation bias, whereby observed distributions are shortened as they approach the most recent time points (Kalbfleisch and Lawless (1991); Sun (1995)). The effect of delayed information on measures of uncertainty often is overlooked. Estimation will produce larger uncertainty intervals for recent events and even larger intervals when forecasting, which can make decision making more complex and subjective. To account for this, one can use data based on date of report rather than date of occurrence. However, this can lead to further complications in estimation. For example, hospital admission time-series may not be recorded by date of admission but by date of returning a positive test (https://coronavirus.data.gov.uk/), whereas length of stay estimates may be generated from the time of admission (Vekaria et al. (2020)). Fitting a relationship between time-series for admissions and bed occupancy will be inconsistent with hospital length of stay estimates.

Multiple data streams

Data collected during an outbreak may be generated as part of the emergency response, rather than a regular data collection process, which can lead to inconsistencies. This is particularly pertinent when data are requested from multiple sources. For example, during the Covid-19 pandemic, each NHS trust in England returned daily data on hospital admission and occupancy. However, being a novel request, it took a few months to ensure consistent data streams across the country. Such labour-intensive data are unlikely to be retrospectively corrected. Statistical models account for such issues, but more robustly when sources of errors are known. For example, if a model fitted to multiple data streams, a known bias in a data stream can be built into the model uncertainty. Many countries have different definitions of what measurements relate to, such as different measures of mortality or different numbers in the tested population (Shadbolt et al. (2021)). Random effects or latent variables can be used to account for individual variations in the data sources and there is increasing literature on integrated models combining data streams. One of the major estimation challenges here is developing methods that are sufficiently general to be of use to a wide range of scenarios.

When using multiple data streams, which are inevitably interrelated, a relationship between the streams (both observed and unobserved) can be estimated (De Angelis et al. (2015)). However, as an epidemic progresses, interventions and policy changes can alter this relationship. Interventions such as vaccination may alter the age distribution of cases, thereby changing hospitalisation/mortality risk. Similarly, treatment could reduce mortality in infected individuals. Dimension reduction techniques can be used to address this, however the interpretation of these procedures is often challenging. A further challenge might arise when attempting to provide a country-wide reproduction number, as one could aggregate potentially de-synchronised data streams from different regions or combine regional reproduction numbers. If these variations are not properly accounted for, inference about infections/prevalence may be biased. If a model does not accurately capture the impact of an intervention, inference regarding the transmission process may be inaccurate (Kretzschmar et al. (2021)). However, there may be insufficient data to quantify vaccine impacts on transmission/disease prevention accurately, creating a substantial modelling challenge.

Challenges:

In preparing for future pandemics, methods for dealing with the following estimation challenges should be considered:

1. Due to the **indirectness of data streams**, a challenge lies in assessing how much information the available data contains about the parameters of interest.
2. **Discrepancies in data collection procedures between spatial regions** leads to different relationships between observed data and the underlying epidemic. If this is not correctly accounted for, estimates can be severely biased.
3. **Data may not be at the desired resolution**, so a challenge lies in aligning model complexity to the available data or making the model robust in accounting for aggregated data.
4. **Temporal aggregation** creates a challenge in how to determine the right balance between flexibility and timeliness of current estimates, and the robustness and reliability of such estimates. This is important when investigating whether an apparent deviation from the previous trend should be considered trend or noise.
5. **Aggregating across demographic/regional groups** may obscure important trends in the data. For example, the effectiveness of a stay-at-home order may correlate with sociodemographic deprivation and therefore failing to account for deprivation may bias estimation of the impact of such orders.
6. Models and statistical methods need to account for **incompleteness in recent data, due to censoring and reporting delays.**
7. When using multiple levels of data, challenges remain in **connecting the various levels of data and accounting for potential biases**.

8. A challenge for future pandemics is **accounting for inconsistencies between different data streams in estimation procedures** to provide more accurate and robust quantification.
9. **Interventions and policy changes during a pandemic** can alter relationships between data streams. This needs to be understood and appropriately accounted for when developing estimation models and quantifying uncertainty.

4. Model-based inference and prediction/forecasting

At different stages in a pandemic, some types of estimation are more feasible than others. In data sparse periods at the start of the pandemic, reliance may be on formal model analysis or expert elicitation. Reliance on data can be more robust as the pandemic evolves and data sources grow and extend. The choice of how to account for uncertainty is made more complex by the fact that there is a general lack of understanding of different types of uncertainty, as discussed in detail above. These discussions notwithstanding, the principal estimation challenge is how to deal with large amounts of missing data and hidden states (e.g. pre-symptomatic infections) that are inherent in the modelling of epidemics.

Explicit likelihoods and data augmentation

Latent variable approaches (e.g., data-augmented MCMC: Gibson and Renshaw (1998); O'Neill and Roberts (1999)) represent unobserved epidemiological events in the statistical model, and these are estimated as part of the inference routine. These often Bayesian approaches can, in theory, use standard methods such as Markov chain Monte Carlo to explore the joint (high-dimensional) parameter space of hidden variables and parameters. Extensions, such as reversible-jump methodologies (Green (1995)) can be employed to allow for unknown numbers of hidden variables. When applicable, these approaches can yield a huge amount of information, e.g. by robustly integrating multiple sources of data including epidemiological observations and genetics (Lau et al. (2015)).

Implementing these techniques requires a close synergy between the underlying model and the inference algorithm to avoid complexities in updating the parameter values conditional on the data at each iteration. Standard random walk updating schemes do not work well with estimating hidden states due to inherent correlation between and within model components. Often, generic algorithms and poorly implemented code are extremely slow to explore parameter space. The development and optimisation of these approaches is thus very challenging and time-consuming, and for large systems with many hidden states, they can become computationally infeasible. However, some generic updating schemes have improved performance including non-centred parameterisations (Papaspiliopoulos et al. (2003)), tempered algorithms (Sacchi and Swallow (2021)) and model-based proposals (Pooley et al. (2015)). Sometimes, approximate models such as discrete-time models help reduce computational complexity, and recent research has exploited sophisticated computer hardware, such as Graphics Processing Units, to help alleviate some of the computational burden. Despite this, for high-dimensional models, computational efficiency, and the challenges in implementation and coding, remain a bottleneck that limits practical application. Some open software implementations of these methods have been developed, e.g., GEM, (https://gem.readthedocs.io/en/latest/), however much more is required before these can be widely used by domain experts.

Likelihood-free simulation-based approaches

An alternative to using latent variables to capture hidden states, is to simulate them directly from the underlying model of interest. Approaches such as maximum likelihood via iterated filtering (Ionides et al. (2006)), Approximate Bayesian Computation (Minter and Retkute (2019)), synthetic likelihoods

(Wood (2010)) and particle MCMC (Andrieu, Doucet and Holenstein (2010)) aim to approximate likelihood functions via simulations. In some cases, these methods can provide exact inference, conditional on the choice of transmission and observation models, but in practice the latter must often be replaced by a measure that penalises large deviations from the observed data in a somewhat arbitrary fashion. The interpretation of these approximations is discussed in more detail in Wilkinson (2013). Despite these issues, these approaches are attractive because they are much more straightforward to implement than latent variable methods, since coding simulation models is in general much easier than using data-augmentation approaches, and general-purpose software exists to implement these. Simulation-based approaches are thus often touted as "plug-and-play", but, in practice there are key challenges in scaling up these methods to large-scale systems.

The main challenge is that these approaches can require hundreds-of-thousands, if not millions of simulation runs to explore the parameter space adequately. If the simulation algorithms are highly stochastic, then this induces large variability in, for example, estimated likelihoods. Particle filter-based likelihood estimation typically scales poorly with data complexity. Thus, relative ease-of-implementation in practice often is superseded by extreme computational loads. Often the only computationally viable approach is to match to summary measures of the data, especially if the data are highly complex. This relies on the generation of informative summary measures, since in many cases it is not possible to identify and generate sufficient statistics (i.e., those that preserve the information in the likelihood). This introduces a loss-of-information, which introduces more uncertainty into parameter inference and prediction.

As discussed in Section 2, a statistical emulator may alleviate some of this computational burden by searching the parameter space exhaustively for areas of the space where good fits to the data are likely to be found, using techniques such as history matching. Alternatively, they can be used to emulate the likelihood directly. Since emulators are typically trained on individual outputs, it is necessary to reduce complex data sets to a lower dimensional set of informative summary statistics. Furthermore, expertise in fitting and validating emulators is required, and to date there is no general-purpose software for implementing these approaches. Moreover, some behaviours seen in stochastic infectious disease models, such as multimodal outcomes, are hard to emulate using standard approaches, and remain an area of ongoing research.

Model structure and inference

At different stages of a pandemic, the decision on which model structure to use may be forced by time constraints that govern when estimates need to be provided or by data availability/quality (Section 3), constrained by the familiarity of those responsible for model development with alternative approaches. However, even when sufficient data are available, the choice of which model to use and the potential implications of that decision on estimates of both parameters and uncertainty bounds are rarely apparent or considered. Stochastic and individual-based models are more realistic, and more widely applicable than deterministic models, particularly as more complex structures are introduced, such as meta-populations, spatial structures or network dynamics (Eames et al. (2015)). These structures may be critical to answering policy questions such as concerning contact tracing or vaccination strategies (Marion et al. (2021)). However, these models are inherently more difficult to fit to data than simpler deterministic models and are also more data- and computation-hungry. When there is need to quantify properties of an outbreak, to inform public health policy, it is important that relevant processes are included in the fitted model and that due consideration is given to the impact of model structure and its potential biases on estimation.

Model assessment and comparison

As discussed in Section 2, parameter and output uncertainties are conditional on the specific choice of model, and thus do not account for the discrepancies between the model and the reality it aims to represent. Incorporating terms into the model that can account for this discrepancy when conducting inference or predictions is an ongoing area of research, and although techniques exist for doing this

for certain approaches (including history matching using emulators), these ideas have not been readily implemented into standard statistical approaches, such as data-augmented or particle MCMC.

Associated with the idea of model discrepancy is the idea of model specification. Multiple model structures can be fitted to data, but new tools are needed to assess model fits, and either select between models or combine them in meaningful ways. Model assessment is frequently difficult with complex models, particularly with spatial and/or temporal components and especially for stochastic models. Latent residuals for spatio-temporal models of disease spread (Lau et al. (2014)) are an interesting move in this direction but much work is needed to develop tools that can be routinely applied across a range of models (Gibson et al. (2018)). Improved tools could provide significant advantages in tackling pandemics by identifying key characteristics of novel pathogens, although this will likely require better quality data than are currently routinely available in outbreak settings (Shadbolt et al. (2021)). Current methodologies, such as information criteria or calculation of marginal likelihood, are not well suited to disease transmission models or are computationally challenging (Pooley and Marion (2018)). For example, standard cross-validation (CV) methods may smooth over deficiencies in model structure if not conducted with care, and are difficult to employ in data sparse scenarios, or across highly structured data such as time-series, or spatially explicit or regional models. These approaches are also computationally demanding, since models need to be refitted multiple times for CV.

There can be a significant difference between models used for explanation and description (Shmueli (2010), Hanna (1969)) and those used for prediction or forecasting, both structurally and from a philosophical perspective. The treatment of uncertainty in each case is potentially different and active consideration needs to be given to what unknowns are being integrated over and/or which quantities could change beyond the data used to estimate the parameters. The reality is that in prediction, model structure and estimated parameter values are often considered to be constant, which will not be realistic in many settings. It seems that this distinction is often not made explicit or considered when developing statistical paradigms for estimation. Consistency across model types might not be feasible but little attempt appears to have been made to consolidate this.

Model ensembles

With a plethora of model types and structures, and many ways of estimating parameters within those models, differences in estimates are almost inevitable. Understanding why these differences occur and how and whether it is sensible to combine inferences is complex and an ongoing area of research (Berger et al. (2021)). Bayesian model averaging enables model aggregation in a statistically principled way, although it requires a close synergy between the specific aspects of the inference algorithm and the model. Expert elicitation may be required in this instance (Section 5). Interpretations of parameters might vary between models, meaning they are not directly comparable and cannot be averaged across models. Forecasts are often more straightforward to average, although outputs from different models may have different spatial and/or temporal granularity, precluding sensible averaging. Borrowing work from other application areas such as local-scale weather and population dynamic models may provide some ideas of how to advance, but models of pandemics are likely to be much more variable and stochastic than those, for example, that have been used to model long-term climate trends. Furthermore, the computational and time-constrained burdens of developing and fitting multiple models often means that individual research groups work with a single, or small sets of models.

Limits to formal estimation

Early in pandemics of novel viruses, knowledge about key parameters may be unreliable or non-existent. Data may be sparse or particularly noisy, making estimation of parameters especially challenging (Section 3). Bayesian inference enables models and data to be combined with prior distributions representing available information. Nonetheless, problems remain. Reliance on assumed knowledge from other viruses or pathogens may introduce biases. This may, however, be the only

option, and putting a distribution on the range of parameter values is preferable to fixing the unknowns to take specific values. As such, the use of systematic prior elicitation techniques (Section 5) to establish plausible prior distributions will help to inform model simulations in the early stages of an outbreak. Systematic sensitivity analyses can help to identify which outputs from the model are sensitive to which parameters and thus offer a means of targeting data collection and study design to identify key parameters better, where possible (Shadbolt et al. (2021)). Emulation and other techniques can also be employed to help perform systematic sensitivity analyses in high-dimensional systems.

Challenges

1. The principal challenge across this section is **the development of efficient and more generally applicable approaches to updating latent states within MCMC frameworks** for high-dimensional models and development of general-purpose software to implement latent variable approaches.
2. Methodological challenges remain in the **development of likelihood free methods based on informative summary statistics** to conduct inference in high-dimensional and stochastic systems.
3. **Implementation of High Performance Computing (HPC) and cloud-based procedures for running large numbers of simulations** from stochastic models. A key challenge is putting the infrastructure in place for groups to be able to respond quickly in the face of a future pandemic, as those often made available at institutional level cannot be made sufficiently flexible to be beneficial for all computational needs.
4. Challenges remain in **methodological approaches to model structure and inference**, as well as facilitating the uptake of these methods by modellers conducting suitable investigations as part of the estimation process. One important challenge is generating observation processes that consider causes for systematic biases in observed data.
5. **Model discrepancy/structural bias**: there are remaining challenges in ensuring model-reality/structural discrepancies are routinely accounted for within estimation processes.
6. **Separation of predictive and descriptive approaches for conducting inference** and estimation. Further challenges arise on separation of validation and assessment approaches for these different philosophical approaches.
7. An important challenge is to **develop more approaches that indicate poor fit** and point towards aspects of the model that are most deficient. Further work is also needed to enable routine application of model comparison methods including marginal likelihood, suitable information criteria and cross-validation.
8. Models are developed and fitted by different research groups; hence **generic ways of comparing and averaging models** that have been fitted in different ways to different data are an important challenge. Difficulties remain in estimating weights or relative beliefs for each of the competing models.

5. Challenges for Expert Judgement

It will often be the case that decisions or forecasts need to be made when the evidence base is limited, particularly in the early stages of a pandemic, or when assessing whether a novel outbreak might lead to a new epidemic. However, even after many months of experience with Covid-19 and

related data collection, analysis, modelling and scientific advances, there remain many important questions and data gaps. Therefore, at various stages during an epidemic, expert judgements may be required to fill gaps and support decisions; indeed, in some contexts this may be the only source of relevant information.

Roles for expert judgement

Expert knowledge has important roles to play in addressing many of the challenges of understanding and responding to a pandemic:

1. **Early warning to decision-makers:** experts often alert decision-makers to emerging pathogen outbreaks, providing models to explain the perceived cause-and-effect relationships, and helping to characterise the potential or relative risks in relation to other infectious diseases and current policies.
2. **Formulating useful and relevant questions**. During an outbreak it is important that policy measures are timely, and that subsequent research and assessments are focused on providing information while there is still time to act. However, as was evidenced by the recent COVID-19 pandemic, recognizing all the relevant factors can be challenging; those involved may not agree on the formulation or prioritisation of key research questions. This can lead to unfocused research and conflicting recommendations. If decisions are to be made on how to act and where to invest in further research, then decision-makers need to decide on the problems to be addressed and the objectives of importance. These decisions often require the balancing of multiple values and objectives, and are best addressed within a decision-analytic framework (Shea et al. (2020); Gregory et al. (2012)). Here, experts are most often required to help frame possible actions to meet objectives, identify information sources to evaluate the consequences of actions, and estimate parameters and model structures. Importantly, objectives in decision-making and policy often extend beyond scientific concerns, to include social, economic and cultural values. This requires an appropriate pool of experts and stakeholders (Hadley et al. (2021)).
3. **Developing models and identifying important parameters**. During an initial outbreak of a new infectious disease, expert opinion will be crucial to inform both model structure (e.g., transmission routes and the stages of the natural history of disease that should be considered) and parameter quantification (e.g., the distribution of latency times). Here, the combination of research question, expert knowledge, and available data will inform the required level of detail (e.g., explicit transmission networks vs. homogeneously mixing populations). In addition, expert knowledge may help disentangle unidentifiable sets of parameters (e.g., contact rates and transmission probabilities), by informing model parameters with prior distributions.

   Uncertainty about the appropriate model should be taken fully into account, and a range of models considered. There is increasing use of multiple models in disease forecasting and scenario projection to aid decision-making (e.g., Li et al. (2017); Viboud et al. (2018); Ray et al. (2020); Borchering et al. (2021)). Recently there have been moves to leverage Structured Expert Judgement approaches within multi-model analyses, to ensure full expression of scientific uncertainty (i.e., uncertainty about biological processes or parameters, or about interventions) while reducing linguistic misunderstandings and minimising cognitive biases in expert elicitation (Shea et al (2020)). This can be done by a curated discussion between modelling rounds, during which linguistic uncertainty about data streams, interventions and objectives can be discussed and clarified. Embedding these in structured decision-making approaches (Runge et al. (2020)) may also enhance and streamline the integration of modelling and policy efforts (Shea et al. (2020)).

4. **Predicting the expected impact of interventions.** This requires assumptions about the effects of postulated interventions, either in terms of model mechanics (e.g. a reduction in duration of infectivity due to treatment) or in terms of expected outcomes (e.g. a decrease in hospital admissions due to quarantining). Assumptions should be made explicit and informed by data, where available, and, where necessary, by expert judgment.

   A major challenge is that the outcomes of interventions will depend on the extent to which individuals and demographic groups participate in, or adhere to, required actions. Predictions about human behaviour are particularly challenging, especially in new and undocumented circumstances such as a pandemic. Timeliness is particularly important as participation and adherence patterns are likely to drift due to changes in risk perception and "policy fatigue" in the population.

5. **Communicating model assumptions and outputs.** Model predictions best represent what is currently known when they are based on a foundation of validated knowledge, and properly incorporate uncertainty. Involving expertise from diverse relevant disciplines will make model predictions more realistic and credible. Also, by involving experts from different disciplines, elements of a common taxonomy and technical language can be developed with which to discuss research questions across disciplines; this is particularly important when addressing emergent pathogen outbreaks and pandemics, which are high-dimensional, multi-disciplinary problems. Such an approach in turn can help to communicate underlying assumptions, results, and their associated uncertainties to policy makers and the public at large. Ideally, experts should be drawn from a range of stakeholder communities, to engender transparency and understanding, leading to increased support for and trust in models to inform policy. ~~, however,~~ Those with expertise in deliberative judgment and stakeholder engagement may help to engage different groups within society to increase awareness, trust and commitment to action.

How to capitalise on expertise?

While expert judgment is often required, there can be unease in using experts to inform decisions of importance, even when the data required are absent, contradictory or uninformative and even though decision-makers are quick to draw on trusted sources (e.g., informal discussions with those they perceive to be reliable experts). To some extent this unease is justified. Experience has repeatedly demonstrated that, under such circumstances, people are prone to make poor judgments, to be affected by contextual biases and other cognitive limitations (O'Hagan et al. (2006); Shanteau et al. (2003)). Even those with substantial knowledge and expertise in a domain typically have difficulty in formulating their judgments in precise, unbiased and meaningful ways (Burgman et al. (2011a); Hemming et al. (2018)). Real care is needed to minimise biases, inaccurate judgments and poor decisions.

Even when experts are asked to provide judgments that are limited to the estimation of facts or outcomes (i.e., not value judgements), they may reasonably disagree (e.g., because of different background and expertise) and may offer different estimates. For those relying on experts, this can be disconcerting.

However, insights from studies of expert judgment have identified ways to capitalise on expert judgements to generate reliable judgments. Many contextual biases and psychological frailties can be mitigated by offering suitable facilitation, training and assistance to experts, as is done in Structured Expert Judgment (SEJ) protocols (see below). It is entirely natural that -- in the face of real scientific uncertainty -- experts will provide apparently divergent judgments; these alternative views are the essence of scientific endeavour and progress, and should be viewed as an advantage for informed decision support, especially if they bring different information and understandings to the table (Cawson et al. (2020); Moon et al. (2019)). Eliciting expert knowledge from a diverse panel makes it more likely that the basic elements required to align research efforts and inform policy are

considered. A variety of methods, including SEJ, for synthesising the range of opinions from an expert panel have been developed, and it has been shown that such syntheses generally provide more accurate judgments than less formal approaches (Hemming et al. (2018, 2020a), Colson & Cook (2017)).

What quantities to elicit?

Expert elicitation is most easily focused on meaningful and in principle measurable outcomes or quantities, such as whether an emergent novel virus will escape its local area, or how many deaths there will be in a certain population and time-interval. If the quantity that has been forecast is later observed, predictive success can be formally evaluated and used to calibrate future forecasts and rank different forecasters.

In applications it is often desirable to express uncertainty about theoretical quantities, such as the basic reproduction number R_0, and other parameters of a model. A challenge is to find good ways to assess such parameters in terms of meaningful quantities. For instance, in a simple SIR model, it may be desirable to assess uncertainty about R_0, based on expert opinion on the duration of infectiousness, combined with data on disease incidence over time during an outbreak, using modelling and expert judgment about the relationships between these quantities.

How to express judgments?

It is important for judgements to be expressed probabilistically. It will seem natural to many practitioners and modellers to give only single point estimates of unknown quantities, but these can be very misleading: it is instead vital for experts to be open about the associated uncertainties and their judgements of these. For example, when faced with a new infectious disease, information gained from a previous disease may be all that is available, but its relevance will be questionable, and this should be represented explicitly. Any projections from such past experience must be carefully considered, taking into account similarities and differences between the past and the future, and, importantly, hedged with appropriate, typically high, uncertainty. Such uncertainty is most usefully expressed as probabilities (O'Hagan and Oakley (2004)).

It can be useful to conduct expert knowledge elicitation in an iterative fashion, asking experts first to make a private individual estimate, giving the experts feedback on how their estimates, knowledge and assumptions, and those of others, translate into expected outcomes, and allowing them to address any apparent discrepancies and misunderstandings. Feedback and iteration can reveal information or assumptions not considered by others and allow experts to see that their views of the problem, and even their interpretation of terms, may differ from their colleagues, so helping them to understand better the range of uncertainty. This is particularly important when experts are unpractised at expressing their knowledge in probabilistic terms.

Structured Expert Judgment

"Structured Expert Judgment" (SEJ, also known as "Expert Knowledge Elicitation") is a broad label for a set of systematic decision support tools for model development and parameter quantification, for use when data are absent or incomplete and critical decisions need to be made. SEJ supplies structured and repeatable methods for the selection and training of experts, and elicitation and aggregation of their uncertain opinions about parameters and the outcomes of events. It delivers probabilistic assessments that are realistic, credible, defensible and, importantly, imparts transparency to the process so that it is possible to critique and review how outcomes based on judgments were derived (O'Hagan (2019)). There are several well-established implementations of SEJ, the most widely used being the Cooke (Cooke's Classical Method) (Cooke, 1991), SHELF (Sheffield Elicitation Framework) (Oakley and O'Hagan (2019); Gosling (2018)) and IDEA (Investigate, Discuss, Estimate and Aggregate) (Hemming et al. (2018)) protocols.

SEJ protocols share a number of features. They emphasise the need to elicit the judgment of more than one expert, encourage diversity in the group of experts convened, ask questions about meaningful events and quantities, request experts to quantify their uncertainty when expressing their judgments, and encourage open expression of judgements by anonymising the contributions of individual experts. While aggregation is not required (Morgan (2015)) many protocols provide processes to derive an aggregate estimate from expert judgments. Validation studies have shown these aggregated estimates are typically more accurate and better calibrated than those of a single, well-credentialled expert (Burgman et al. (2011b); Colson and Cooke (2017); Hemming et al. (2018, 2020)). While there are many subtle differences in how the protocols guide experts through an elicitation, the primary differences relate to the level of interaction between experts, and the approach for aggregation (Hanea et al. (2021); O'Hagan (2019)). We briefly elaborate on these differences among the three protocols listed above.

All three protocols begin by asking the experts to make judgements individually and privately. Cooke then aggregates the individual judgements by forming a weighted average. In order to derive weights, the experts are also asked for judgements about some additional quantities called seed variables, whose true values are known to the investigator but not to the experts. Weights are computed based on how well each expert's judgements accord with the known true values. The Cooke protocol does not include discussion between experts, except possibly to confirm the aggregated distribution. In contrast, group discussion is a feature of both SHELF and IDEA, with the objective of exploring differences in the initial judgements by sharing opinions and interpretations of the evidence. IDEA then asks the experts to revise their initial judgements, privately, after which they are aggregated, usually by an equally-weighted average. SHELF, however, asks the experts themselves to agree on judgements that will represent what a rational, impartial observer would believe after hearing their opinions and their reasoning.

An excellent, if slightly dated, overview of SEJ, with detailed practical guidance, may be found in the 2014 report of the European Food Safety Authority on Expert Knowledge Elicitation (EFSA (2014)). While it is targeted to a different field of application, it is relevant to infectious diseases modelling. For more recent overviews see Dias et al. (2018), O'Hagan (2019), Hanea et al. (2021), Williams et al. (2021)**.**

SEJ has been applied successfully in a wide range of contexts, including for interventions to control spread of wildlife diseases (Szymanski et al. (2009)) and human infectious disease applications (McAndrew et al. (2021); McAndrew and Reich (2020)).

Some of the epidemiological and infection models developed in the UK, in response to the Covid-19 pandemic, have – inevitably – had to make use of expert judgment, in one form or another. But, this said, most, if not all, of these judgments were elicited informally and were untested, their sources undocumented, and associated uncertainties and assumptions not made explicit nor adequately reported. This state of affairs unquestionably increased the risk of introducing serious biases and the lack of openness, transparency and scientific validation undoubtedly helped undermine public and political trust in expert judgment in an evolving crisis.

Adopting SEJ in epidemiology would help create more reliable model assumptions and parameter estimates, would support the advancement and credibility of the science and provision of scientific advice and, ultimately, lay foundations for better decisions and public health outcomes.

<u>Challenges</u>

1. Building awareness of, expertise in, and familiarity with, structured expert elicitation in the epidemiological and modelling communities.

2. Encouraging experts to become engaged in SEJ in a way that they feel they are contributing for the greater good. During the Covid-19 pandemic, many people were willing to give of their time and expertise, but this cannot be taken for granted, especially if they perceive a risk of being identified personally and abused on social media
3. Training suitable experts and facilitators so that they are ready to go when required. This includes having one or more expert panels, with administrative support available, especially at the start of an epidemic when a fast response is needed. Standing panels of facilitators and administrators could also be used for other kinds of emergencies, though expert panels would need to be relevant to each specific task.
4. Developing guidelines regarding which elicitation procedures can best serve different types of questions and uncertainties.
5. Building and regularly updating an expert elicitation manual and toolbox for emergent zoonose and viral pandemic preparedness and rapid response, and ensuring its relevance, quality and readiness.
6. Developing methods for efficient, appropriate and timely integration of expert judgments and accruing empirical data, and – perhaps most critical - continual revision and updating of estimates as conditions and circumstances vary when policy changes are implemented, or infection resurgences occur.
7. Identifying formats for the clear presentation of the probabilistic expressions of knowledge that are the outcomes of SEJ exercises, and training modellers and decision makers to understand, utilise and communicate these effectively.
8. Extending and consolidating advice for structured expert judgment beyond parameter estimation to guidance for full probabilistic methods, as well as guidance for the elicitation of multiple or consolidated models from experts.
9. Developing principled methods for quantitative expert judgment of structural model discrepancy, whether inherent in the internal configuration of the model itself or reflecting its limitations in representing the real pandemic.

6. Conclusion

There is a large amount of research on modelling and estimation for epidemics and pandemics, as well as the development of the appropriate estimation and uncertainty quantification paradigms to conduct that research. However, the current Covid-19 pandemic has highlighted many remaining challenges in method development, application and uptake within the wider epidemiological community that should be treated as priorities in preparing for future pandemics.

Collating themes across the dimensions of this paper, major difficulties often revolve around the building of infrastructures necessary for conducting necessary analyses or communicating results on a large, rapidly changing and noisy system that rarely follows the format that ideal simulations prepare researchers for. These infrastructures cover data accessibility and computational resource availability and software development that is flexible enough to be useful for the wider community. Infrastructure issues also incorporate difficulties of open communication and knowledge exchange between differing groups, where there is frequently a conflict between open science and rapid response and demands of academic careers.

The current pandemic has highlighted the necessity of open communication routes between researchers, data providers and practitioners in each of these areas and priorities going forward should be in facilitating those open pathways, consolidating research engineers and other subject

matter experts within the estimation pipeline, as well as making open software available such that uptake of robust uncertainty quantification and parameter and model estimation can be conducted by a wider community of epidemiological modellers. Often useful methods exist either within the wider field of epidemiology or in related application areas, but the potential has not come to the attention of those on the front line. Synthetic reviews, such as the ones in this special feature that draw on the varied expertise of many scientists, provide a critical repository of wide-ranging knowledge for novices and experts alike, and save researchers from having to reinvent the wheel in times of crisis.

It is impossible to discuss challenges in estimation without also making references to challenges in the components that estimation depends on, namely the mechanistic models and data that feed into estimation approaches. Challenges within these areas inadvertently have knock-on effects on the ability of statisticians and modellers to conduct robust estimation, and hence challenges in all these areas should not be considered in isolation. Estimation also feeds into many other dimensions of pandemic preparedness and response, such as modelling interventions, informing policy and politics and determining emergence of new pathogens and/or virus strains. Without combining these different domains, estimation remains a purely academic affair and fails to reach its full potential in directly or indirectly informing public health responses.

Acknowledgements

The co-authors acknowledge helpful comments from Valerie Isham, Denis Mollison, Tony O'Hagan, Jim Smith, Willy Apsinall, Stephen Sparks and Marissa McBride.

The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support during the Infectious Dynamics of Pandemics programme where work on this paper was undertaken. This work was supported by EPSRC grant no. EP/R014604/1.

LEC is a member of the Dutch Covid-19 Monitoring Consortium, which is funded by ZonMw (https://www.zonmw.nl, Grant 10430022010001). Daniel Villela is a fellow from National Council for Scientific and Technological Development (Ref. 309569/2019-2, 441057/2020-9). Katriona Shea acknowledges NSF COVID-19 RAPID award 2028301. TJM is supported by an "Expanding Excellence in England" award from Research England and UKRI grants: EP/V051555/1 and MR/V038613/ (JUNIPER Consortium). LP and CO are funded by the Wellcome Trust and the Royal Society (grant 202562/Z/16/Z). LP is also supported by the UKRI through the JUNIPER modelling consortium (grant number MR/V038613/1) and by The Alan Turing Institute for Data Science and Artificial Intelligence.

G.M. is supported by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS).

Authors' contributions

All authors took part in discussions and wrote sections of the manuscript. B.S. coordinated discussions throughout and compiled the final version of the manuscript. All authors edited the manuscript and approved the final version for publication.

References

I. Andrianakis, I. R. Vernon, N. McCreesh, T. J. McKinley, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White, "Bayesian history matching of complex infectious disease models using emulation: A tutorial and a case study on HIV in Uganda," PLOS Computational Biology, vol. 11, e1003968, 2015.

C. Andrieu, A. Doucet, and R. Holenstein, “Particle Markov chain Monte Carlo methods,” Journal of the Royal Statistical Society. Series B: Statistical Methodology, vol. 72, 2010.

L. S. Bastos and A. O’Hagan, “Diagnostics for Gaussian Process Emulators”, Technometrics, 51:4, 425-438, 2009.

L. Berger, N. Berger, V. Bosetti, I. Gilboa, L. P. Hansen, C. Jarvis, M. Marinacci, and R. D. Smith, “Rational policymaking during a pandemic,” PNAS, vol. 118, no. 4. 2021.

R. K. Borchering et al. “Modeling of Future COVID-19 Cases, Hospitalizations, and Deaths, by Vaccination Rates and Nonpharmaceutical Intervention Scenarios — United States, April–September 2021,” Morbidity and Mortality Weekly Report (MMWR) 70(19):719–724, 2021.

J. Brynjarsdottir and A. O’Hagan, “Learning about physical parameters: The importance of model discrepancy,” Inverse Problems, vol. 30, 2014.

M. Burgman, A. Carr, L. Godden, R. Gregory, M. McBride, L. Flan- der, and L. Maguire, “Redefining expertise and improving ecological judgment,” Conservation Letters, vol. 4, pp. 81–87, 2011a.

M. A. Burgman, M. McBride, R. Ashton, A. Speirs-Bridge, L. Flander, B. Wintle, F. Fidler, L. Rumpff, and C. Twardy, “Expert status and performance,” PLoS ONE, vol. 6, 2011b.

J. G. Cawson, V. Hemming, A. Ackland, W. Anderson, D. Bowman, R. Bradstock, T. P. Brown, J. Burton, G. J. Cary, T. J. Duff, A. Filkov, J. M. Furlaud, T. Gazzard, M. Kilinc, P. Nyman, R. Peacock, M. Ryan, J. Sharples, G. Sheridan, K. Tolhurst, T. Wells, P. Zylstra, and T. D. Penman, “Exploring the key drivers of forest flammability in wet eucaypt forests using expert-derived conceptual models,” Landscape Ecology, vol. 35, 2020.

A. R. Colson and R. M. Cooke, “Cross validation for the classical model of structured expert judgment,” Reliability Engineering & System Safety, vol. 163, pp. 109–120, 2017.

D. Commenges, “Multi-state models in epidemiology,” Lifetime Data Analysis, vol. 5, 1999.

R. M. Cooke, Experts in Uncertainty: Opinion and Subjective Probability in Science (Environmental Ethics and Science Policy), vol. 44. 1991.

P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith, “Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes linear strategies for large computer experiments,” 1997.

J. Cumming and M. Goldstein, “Bayes Linear Uncertainty Analysis for Oil Reservoirs Based on Multiscale Computer Experiments,” in A. O’Hagan and M. West (eds.), The Oxford Handbook of Applied Bayesian Analysis, pp. 241–270. Oxford University Pres, 2010.

D. De Angelis, A. M. Presanis, P. J. Birrell, G. S. Tomba, and T. House, “Four key challenges in infectious disease modelling using data from multiple sources,” Epidemics, vol. 10, pp. 83–87, 2015.

L. C. Dias, A. Morton, and J. Quigley, “Elicitation - the science and art of structuring judgement,” 2018.

J. L. Dickinson, B. Zuckerberg, and D. N. Bonter, “Citizen science as an ecological research tool: Challenges and benefits,” Annual Review of Ecology, Evolution, and Systematics, vol. 41, pp. 149–172, 2010.

K. Eames, S. Bansal, S. Frost, and S. Riley, “Six challenges in measuring contact networks for use in modelling,” Epidemics, vol. 10, pp. 72–77, 2015.

EFSA, “Guidance on expert knowledge elicitation in food and feed safety risk assessment,” doi:10.2903/j.efsa.2014.3734, 2014

G. J. Gibson and E. Renshaw, “Estimating parameters in stochastic compartmental models using Markov chain methods,” IMA Journal of Mathemathics Applied in Medicine and Biology, vol. 15, 1998.

G. J. Gibson, G. Streftaris, and D. Thong, “Comparison and assessment of epidemic models,” Statistical Science, vol. 33, no. 1, pp. 19 – 33, 2018.

M. Goldstein and J. Rougier, “Reified Bayesian modelling and inference for physical systems,” Journal of Statistical Planning and Inference, vol. 139, no. 3, pp. 1221–1239, 2009.

J. P. Gosling, SHELF: The Sheffield Elicitation Framework, pp. 61–93. Cham: Springer International Publishing, 2018

P. J. Green, “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination,” Biometrika, vol. 82, 1995.

R. Gregory, L. Failing, M. Harstone, G. Long, T. Mcdaniels, and D. Ohlson, Structured Decision Making: A Practical Guide to Environmental Management Choices. John Wiley Sons, 2012.

F. Gugole, L. E. Coffeng, W. Edeling, B. Sanderse, S. J. d. Vlas, and D. Crommelin, “Uncertainty quantification and sensitivity analysis of covid-19 exit strategies in an individual-based transmission model,” medRxiv, 2021.03.24.21254218, 2021.

L. Hadley, P. Challenor, C. Dent, V. Isham, D Mollison, D. A. Robertson, B Swallow, and C Webb, “Challenges on the interaction of models and policy for pandemic control”, under submission to Epidemics special issue, 2021.

A. M. Hanea, V. Hemming, and G. F. Nane, “Uncertainty quantification with experts: Present status and research needs,” 2021.

J. F. Hanna, “Explanation, prediction, description, and information theory,” Synthese, vol. 20, 1969.

V. Hemming, N. Armstrong, M. A. Burgman, and A. M. Hanea, “Improving expert forecasts in reliability: Application and evidence for structured elicitation protocols,” Quality and Reliability Engineering International, vol. 36, 2020.

V. Hemming, T. V. Walshe, A. M. Hanea, F. Fidler, and M. A. Burgman, “Eliciting improved quantitative judgements using the idea protocol: A case study in natural resource management,” PLoS ONE, vol. 13, 2018.

E. L. Ionides, C. Breto, and A. A. King, “Inference for nonlinear dynamical systems,” Proceedings of the National Academy of Sciences of the United States of America, vol. 103, 2006.

J. D. Kalbfleisch and J. F. Lawless, “Regression models for right truncated data with applications to aids incubation times and reporting lags,” Statistica Sinica, vol. 1, 1991.

M. C. Kennedy and A. O’Hagan, “Bayesian calibration of computer models,” Journal of the Royal Statistical Society. Series B: Statistical Methodology, vol. 63, 2001.

M. C. Kennedy and A. O’Hagan, “Predicting the output from a complex computer code when fast approximations are available,” Biometrika, vol. 87, no. 1, pp. 1–13, 2000.

M. Kretzschmar et al. “Challenges for modelling interventions for future pandemics,” in preparation for Epidemics special issue, 2021.

M. S. Y. Lau, G. Marion, G. Streftaris, and G. J. Gibson, “New model diagnostics for spatio-temporal systems in epidemiology and ecology,” Journal of The Royal Society Interface, vol. 11, no. 93, p. 20131093, 2014.

M. S. Y. Lau, G. Marion, G. Streftaris, and G. J. Gibson, “A systematic Bayesian integration of epidemiological and genetic data,” PLOS Computational Biology, vol. 11, pp. e1004633, 2015.

S. L. Li, O. N. Bjørnstad, M. J. Ferrari, R. Mummah, M. C. Runge, C. J. Fonnesbeck, M. J. Tildesley, W. J. Probert, and K. Shea, “Essential information: Uncertainty and optimal control of Ebola outbreaks,” Proceedings of the National Academy of Sciences of the United States of America, vol. 114, 2017.

J. O. Lloyd-Smith, D. Mollison, C. J. E. Metcalf, P. Klepac, and J. A. P. Heesterbeek, “Challenges in modelling infectious disease dy- namics: Preface,” Epidemics, vol. 10, pp. iii–iv, 2015.

T. C. McAndrew and N. G. Reich, “An expert judgment model to predict early stages of the Covid-19 outbreak in the united states,” 2020.

T. McAndrew, N. Wattanachit, G. C. Gibson, and N. G. Reich, “Aggregating predictions from experts: A review of statistical methods, experiments, and applications,” 2021.

N. McCreesh, I. Andrianakis, R. N. Nsubuga, M. Strong, I. Vernon, T. J. McKinley, J. E. Oakley, M. Goldstein, R. Hayes, and R. G. White, “Universal test, treat, and keep: improving art retention is key in cost- effective hiv control in uganda,” BMC Infectious Diseases, vol. 17, no. 1, 2017.

T. J. McKinley, I. Vernon, I. Andrianakis, N. McCreesh, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White, “Approximate Bayesian computation and simulation-based inference for complex stochastic epi- demic models,” Statistical Science, vol. 33, 2018.

G. Marion et al., “Modelling: understanding pandemics and how to control them,” in preparation for Epidemics special issue, 2021.

A. Minter and R. Retkute, “Approximate Bayesian computation for infectious disease modelling,” Epidemics, vol. 29, 2019.

K. Moon, A. M. Guerrero, V. M. Adams, D. Biggs, D. A. Blackman, L. Craven, H. Dickinson, and H. Ross, “Mental models for conservation research and practice,” 2019.

M. G. Morgan, “Our knowledge of the world is often not simple: Policymakers should not duck that fact, but should deal with it,” Risk Analysis, vol. 35, pp. 19–20, 2015.

A. O’Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow, Uncertain Judgements: Eliciting Experts’ Probabilities. John Wiley and Sons, Chichester, 2006.

A. O’Hagan, “Expert knowledge elicitation: Subjective but scientific,” The American Statistician, vol. 73, pp. 69–81, 03 2019.

A. O’Hagan and J. E. Oakley, “Probability is perfect, but we can’t elicit it perfectly,” Reliability Engineering & System Safety, vol. 85, no. 1, pp. 239–248, 2004.

P. D. O’Neill and G. O. Roberts, “Bayesian inference for partially observed stochastic epidemics,” Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 162, pp. 121–129, 1999.

K. M. O’Reilly, D. J. Allen, P. Fine, and H. Asghar, “2sampling for sars-cov-2 must be met: lessons from polio eradication,” The Lancet Microbe, vol. 1, pp. e189–e190, 2020.

J. E. Oakley and A. O’Hagan, SHELF: the Sheffield Elicitation Frame- work (version 4). School of Mathematics and Statistics, University of Sheffield, http://tonyohagan.co.uk/shelf, 2019.

C. E. Overton, H. B. Stage, S. Ahmad, J. Curran-Sebastian, P. Dark, R. Das, E. Fearon, T. Felton, M. Fyles, N. Gent, I. Hall, T. House, H. Lewkowicz, X. Pang, L. Pellis, R. Sawko, A. Ustianowski, B. Vekaria, and L. Webb, "Using statistics and mathematical modelling to understand infectious disease outbreaks: Covid-19 as an example," Infectious Disease Modelling, vol. 5, 2020.

O. Papaspiliopoulos, G. O. Roberts, and M. Sk¨old, "Non-centered parameterisations for hierarchical models and data augmentation," Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting, 2003.

L. Pellis, F. Scarabel, H. B. Stage, C. E. Overton, L. H. Chappell, K. A. Lythgoe, E. Fearon, E. Bennett, J. Curran-Sebastian, R. Das, M. Fyles, H. Lewkowicz, X. Pang, B. Vekaria, L. Webb, T. A. House, and I. Hall, "Challenges in control of Covid-19: Short doubling times and long delay to effect of interventions," 2020.

C. M. Pooley, S. C. Bishop, and G. Marion, "Using model-based proposals for fast parameter inference on discrete state space, continuous-time Markov processes," Journal of the Royal Society Interface, vol. 12, 2015.

C. M. Pooley and G. Marion, "Bayesian model evidence as a practical alternative to deviance information criterion," Royal Society Open Science, vol. 5, 2018.

F. Pukelsheim, "The three sigma rule," The American Statistician, vol. 48, pp. 88–91, 05, 1994.

E. L. Ray, N. Wattanachit, J. Niemi, A. H. Kanji, K. House, et al. "Ensemble forecasts of coronavirus disease 2019 (Covid- 19) in the u.s," 2020.

J. Rougier, M. Goldstein, and L. House, "Second-order exchangeability analysis for multimodel ensembles," Journal of the American Statistical Association, vol. 108, no. 503, pp. 852–863, 2013.

M. C. Runge, S. J. Converse, J. E. Lyons, and D. R. Smith (eds), Structured Decision Making: Case Studies in Natural Resource Management. Johns Hopkins University Press, 2020.

G. Sacchi and B. Swallow, "Towards efficient Bayesian approaches to inference in hierarchical hidden Markov models for inferring animal behaviour," Frontiers in Ecology and Evolution, 2021.

N. Shadbolt et al., "Data challenges for pandemic modelling", in preparation for Epidemics special issue, 2021.

J. Shanteau, D. J. Weiss, R. P. Thomas, and J. Pounds, "How can you tell if someone is an expert? Performance-based assessment of expertise," 2012.

K. Shea, R. K. Borchering, W. J. Probert, E. Howerton, et al., "Covid-19 reopening strategies at the county level in the face of uncertainty: Multiple models for outbreak decision support," 2020.

G. Shmueli, "To explain or to predict?" Statistical Science, vol. 25, 2010.

J. Sun, "Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to aids studies," Biometrics, vol. 51, 1995.

J. A. Szymanski, M. C. Runge, M. J. Parkin, and M. Armstrong, "White-nose syndrome management: report on structured decision making initiative.," 2009.

P. Touloupou, B. Finkenstädt, and S. E. F. Spencer, "Scalable Bayesian inference for coupled hidden Markov and semi-Markov models," Journal of Computational and Graphical Statistics, vol. 29, 2020.

B. Vekaria, C. Overton, A. Wi, S. Ahmad, A. Aparicio-Castro, J. Curran-Sebastian, J. Eddleston, N. A. Hanley, T. House, J. Kim, W. Olsen, M. Pampaka, L. Pellis, D. P. Ruiz, J. Schofield, N. Shryane, and M. J. Elliot, "Hospital length of stay for Covid-19 patients: Data-driven methods for forward planning," Research Square, 2020.

I. Vernon, M. Goldstein, and R. G. Bower, “Galaxy formation: a bayesian uncertainty analysis,” Bayesian Analysis, vol. 5, pp. 619–669, 12, 2010.

C. Viboud, K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, and A. Vespignani, “The rapid Ebola forecasting challenge: Synthesis and lessons learnt,” Epidemics, vol. 22, 2018.

R. D. Wilkinson, “Approximate Bayesian computation (ABC) gives exact results under the assumption of model error,” vol. 12, no. 2, pp. 129– 141, 2013.

C. J. Williams, K. J. Wilson, and N. Wilson, “A comparison of prior elicitation aggregation using the classical method and shelf,” Journal of the Royal Statistical Society: Series A (Statistics in Society), 2021.

S. N. Wood, “Statistical inference for noisy nonlinear ecological dynamic systems,” Nature, vol. 466, 2010.

Conflict of Interest

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: