Statistical Science
2022, Vol. 37, No. 2, 183–206
https://doi.org/10.1214/22-STS854

Interoperability of Statistical Models in Pandemic Preparedness: Principles and Reality

George Nicholson[+], Marta Blangiardo[+], Mark Briers[+], Peter J. Diggle[+], Tor Erlend Fjelde[+], Hong Ge[+], Robert J. B. Goudie, Radka Jersakova[+], Ruairidh E. King[+], Brieuc C. L. Lehmann[+], Ann-Marie Mallon[+], Tullia Padellini[+], Yee Whye Teh[+], Chris Holmes[+] and Sylvia Richardson[+]

Abstract. We present *interoperability* as a guiding framework for statistical modelling to assist policy makers asking multiple questions using diverse datasets in the face of an evolving pandemic response. Interoperability provides an important set of principles for future pandemic preparedness, through the joint design and deployment of adaptable systems of statistical models for disease surveillance using probabilistic reasoning. We illustrate this through case studies for inferring and characterising spatial-temporal prevalence and reproduction numbers of SARS-CoV-2 infections in England.

Key words and phrases: Bayesian graphical models, Bayesian melding, COVID-19, evidence synthesis, interoperability, modularization, multisource inference.

George Nicholson is Senior Researcher, University of Oxford, Oxford, UK (e-mail: george.nicholson@stats.ox.ac.uk). Marta Blangiardo is Professor, MRC Centre for Environment and Health, Dept. of Epidemiology and Biostatistics, Imperial College London, London, UK. Mark Briers is Honorary Turing Fellow, The Alan Turing Institute, London, UK. Peter J. Diggle is Distinguished Professor, CHICAS, Lancaster Medical School, Lancaster University, Lancaster, UK. Tor Erlend Fjelde is PhD student, Department of Engineering, University of Cambridge, Cambridge, UK. Hong Ge is Senior Research Fellow, Department of Engineering, University of Cambridge, Cambridge, UK. Robert J. B. Goudie is Senior Investigator Statistician, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. Radka Jersakova is Senior Research Data Scientist, The Alan Turing Institute, London, UK. Ruairidh E. King is Data Wrangler, MRC Harwell Institute, Harwell, UK. Brieuc C. L. Lehmann is Assistant Professor, UCL, London, UK. Ann-Marie Mallon is Head of Bioinformatics, MRC Harwell Institute, Harwell, UK. Tullia Padellini is Research Associate, MRC Centre for Environment and Health, Dept. of Epidemiology and Biostatistics, Imperial College London, London, UK. Yee Whye Teh is Professor, University of Oxford, Oxford, UK. Chris Holmes is Professor, University of Oxford, Oxford, UK, Programme Director for Health and Medical Sciences, The Alan Turing Institute, London, UK, and Programme Leader, MRC Harwell Institute, Harwell, UK. Sylvia Richardson is Professor of Biostatistics and Emeritus Director, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK (e-mail: sylvia.richardson@mrc-bsu.cam.ac.uk).

1. BACKGROUND AND KEY PRINCIPLES OF INTEROPERABILITY

Faced with the coronavirus disease 2019 (COVID-19) pandemic that posed an urgent and overwhelming threat to global population health, policy makers worldwide sought to muster reactive and proactive analytic capabilities in order to track the evolution of the pandemic in real time, and to investigate potential control strategies. In the United Kingdom, governmental health data analytic capabilities were strengthened in the midst of the pandemic with the creation in May 2020 of the Joint Biosecurity Centre (JBC), whose mission was to provide evidence-based, objective analysis, assessment and advice so as to inform the response of local and national decision-making bodies to current and future epidemics. Early on, the JBC established links with external academic or institutional groups, and in particular the health programme (lead Chris Holmes) and the digital technology project within the defence and security programme (lead Mark Briers) of The Alan Turing Institute (herein Turing).

[+]Members of The Alan Turing Institute and Royal Statistical Society's "Turing-RSS Health Data Lab", in partnership with the UK Health Security Agency, part of NHS Test and Trace within the Department of Health and Social Care, https://www.turing.ac.uk/research/research-projects/turing-rss-health-data-lab, formerly known as the "Turing-RSS Statistical Modelling and Machine Learning Laboratory".

In parallel, institutions such as the Royal Society and the Royal Statistical Society (RSS) established task force groups to contribute their collective expertise to the UK government and public bodies. The RSS Covid-19 Task Force [11] was created in April 2020. Besides intervening on statistical issues, co-chairs and members of its steering group were mindful of ensuring coordination and avoiding duplication with other initiatives. Under the leadership of Chris Holmes (Turing and RSS Covid-19 Task Force) and with the support of Sylvia Richardson (RSS president-elect, co-chair of the RSS Covid-19 Task Force) and Peter Diggle (steering group of RSS Covid-19 Task Force), a partnership to provide additional capacity to the JBC was established between Turing and the RSS. This resulted in the creation in October 2020 of the "Statistical Modelling and Machine Learning Laboratory" within the JBC, now known as the "Turing-RSS Health Data Lab" (the Lab). The Turing-RSS Lab's aims are to work within the JBC to provide additional capacity through independent, open-science research based on rigorous statistical modelling and inference directed at JBC priority areas. In October 2021, the United Kingdom Health Security Agency (UKHSA) was created, which incorporated the JBC, and the Lab's partnership then became a partnership with UKHSA.

Established against the background of a fast moving pandemic, this partnership brought into focus a number of interesting challenges to conventional statistical practice arising, in particular, from the need to model real-time, messy data from diverse sources, in order to efficiently address rapidly evolving public health demands. The dynamic nature of the pandemic and the resulting public health priorities led to frequent changes in the specific questions being asked of the data, with focus often shifting unpredictably and suddenly. This challenged conventional statistical analysis protocols that target specific research questions, as these would take too long to deliver, and carry an associated risk of redundancy. Instead, it was necessary to develop robust, easy-to-update and reuseable modules which could be integrated into arbitrarily complex models to provide analyses useful for decision-making.

We will often use the terminology "model" alongside the distinct but related terminology "module/modular/modularity", so it is useful to compare and contrast these concepts at the outset:

- *Model.* "A statistical model is a probability distribution constructed to enable inferences to be drawn or decisions made from data" [13]. A model can be comprised of one or more modules.
- *Module.* A module is a model component or, more precisely, a joint probability distribution linking some or all of: observed data, latent (unobserved) parameters or random effects, and user-specified hyperparameters.

Of course the distinction between model and module is not strict, as a simple model may also be a module of a more complex model. Rather than trying to be pedantic with the above definitions, we are simply aiming to convey the spirit and sense of the vocabulary we use throughout the paper.

Our practice and strategic thinking led us to develop a set of inter-connected health protection models, and to articulate the principle of "*interoperability*" as an important statistical concept and goal for future disease surveillance systems.[1] In this article, we discuss the emerging principles of interoperability of statistical models that we have operationalised since 2021, and we illustrate these on case studies carried out in the Lab. Interoperability can be usefully, and loosely, characterised as "an operational statistical and computational data-driven framework based on modularized inference designed to provide timely analyses and future preparedness for decision making on related questions relevant to a common process". At its core, interoperability is driven by the need to optimally fulfil the following complementary principles:

P.1 Shared latent quantities. When building models to answer different questions, harmonize the models' essential elements by incorporating common key latent quantities, for example, disease prevalence.

P.2 Modularity. Use modular statistical approaches, such as Bayesian graphical models, as a foundation for inference and computations to ensure both modelling agility and principled propagation of uncertainty.

P.3 Structural robustness. When one or more modules or data sources are considered potentially unsound, consider specifying robust model structures, in which the failure or misspecification of one module has a limited impact on other modules; aim for transparent and appropriate sharing of information across diverse data modules, paying particular attention to possibility of conflict between data sources.[2]

P.4 Dynamic model validation. Identify opportunities for ongoing dynamic assessment of model performance, ideally that can be applied to multiple models sharing latent quantities and/or addressing similar questions.

P.5 Composability.[3] Strive for efficient transfer of results of analyses between different computational tools, with the freedom to implement each module of analysis primarily in its most convenient software package.

[1] We stress at the outset that when we refer simply to "interoperability", we always mean *statistical interoperability*. The concept of interoperability referred to in engineering or for computer hardware is not the subject of this paper.

[2] We distinguish structural robustness from the more general statistical notion of robustness, namely good performance across of wide class of possible true data generating mechanisms.

[3] We use the term composability by analogy with its definition in system design as a principle that deals with the inter-relationships of components.

P.6 Probabilistic programming. Fit models using modular, efficient, high-level probabilistic programming languages, allowing flexible and fast implementation of complex methods and nimble repurposing of scripts as new data and questions arise.

P.7 Data pipelines. Maintain version-controlled data streams synchronised across a system of models and modules, accompanied by vigilant quality control, for example through global visualization of data inputs and results.

P.8 Reproducibility. Ensure the generation of stable reproducible results via an open-source code base, coupled with tight control on code and data versioning.

There are obvious connections between these objectives. The key desirability of modularity informed our choice of Bayesian inference methods and Bayesian graphical models as the core statistical framework. We have heavily relied on the known flexibility of Bayesian hierarchical models (BHMs) with latent (Gaussian) process components to deliver predictive and/or explanatory inferences while accommodating complex data structures and measurement processes and enabling principled data synthesis (for examples, see [1] and the introductory chapter of [25]). Recent work on Markov melding [24] and statistical learning, and on cutting or restricting information flow between modules [8, 29, 35, 39, 54] is particularly relevant to anchor and operationalize our goals.

2. INTRODUCTION TO INTEROPERABILITY AS A STRATEGIC GOAL, INFORMED BY CURRENT APPROACHES TO DISEASE SURVEILLANCE

Integrated infectious disease surveillance has unusual characteristics from a statistical analysis perspective. Multiple research questions on disparate outcomes are targeted towards better understanding *of a common underlying process*, for example, of the disease spread and its evolution in time and space. Asking multiple questions of a single process opens up a spectrum of modelling approaches. At one extreme, different models using potentially partially overlapping data could be built independently to estimate common latent characteristics of the disease under consideration. At the other extreme, we can look to build a single universal joint model covering every facet of the disease process that is theoretically able to answer any question. Our conjecture is that, in the face of the constraints and operational challenges outlined above, there is an optimal middle ground in which we develop models and analysis plans and couple them according to the guiding principles of interoperability.

2.1 Building Separate Models for Common Key Latent Quantities

Many flavours of epidemic models have been developed for tracking COVID-19 disease transmission ranging from agent-based simulation models to age-structured compartmental models or discretized semi-mechanistic models using renewal equations [2, 20]. It is not our purpose to give a comprehensive review of these but simply to take stock of the rich diversity of modelling approaches and data sources chosen by different teams to inform the calibration or estimation of epidemic parameters. In the UK, a number of academic modelling groups have been actively participating in the expert advisory panel Scientific Pandemic Influenza Group on Modelling (SPI-M), a subgroup of the Scientific Advisory Group for Emergencies (SAGE), which has advised the UK government from the start of the pandemic. The adopted collegiate mode of working of SPI-M has fostered the development of a set of distinct models for estimating key epidemic quantities and producing short-term forecasts. Besides the differences between the modelling approaches, the choice of primary data sources and the ways to embed this information into the modelling framework may also differ, creating an ensemble of models with complex connections. This parallel model development step is then followed by a meta-analysis of the estimates of the key epidemic parameters at regular intervals, and the synthesized results are communicated to the public, as well as used by policy makers [18].

Such a strategy has the benefit of structural robustness for inference on latent quantities, such as the much quoted effective reproduction number, here denoted $\mathcal{R}_t^{\text{eff}}$, aimed at protecting against misspecification of one or more of the models in the ensemble, as well as countering undue influence of artefacts connected to particular data sources. But it also raises unresolved statistical issues on how best to formulate criteria for including suitable models in the ensemble, and how to weight a set of inter-connected models in the final estimate. The UK government website lists ten academic groups producing models which contribute to the ensemble [18], and pooled estimates are computed via a random effects meta-analysis in which all models are given equal weight [40]. From a statistical perspective, using weights based on some measure of short-term predictive performance for each model would be a natural alternative, but could be potentially challenging to operationalize, as it requires all models to produce comparable predictive outputs; see [5] for a discussion of forecast evaluation metrics. Hence, meta-analysing results from an ensemble of models, while attractively operationally simple, results in overall estimates with unclear statistical properties, due to models using the same or overlapping inputs, and uncertainty not being fully propagated.

2.2 Building a Full Joint Model

At the other end of the spectrum, one could strive to develop a single "uber-model" which contains all latent quantities of interest and a comprehensive set of data

sources. The corresponding full joint posterior distribution of all the parameters of interest could then be derived simultaneously, using the paradigm of Bayesian graphical models. While theoretically optimal (provided the model is well specified) a full joint model is particularly challenging in a fast moving epidemic situation because it requires expansion, adaptation and revision each time a new piece of information needs to be integrated, whether it is a new data source (e.g., presence of SARS-CoV-2 in geolocalised wastewater data), or a new policy or intervention influencing the structure of the model (e.g., vaccination), or the behaviour of people (e.g., mobility).

Such increasing model complexity is accompanied by a heightened risk of misspecification and conflict between sources of information. This makes it hard to understand how different data sources are balanced in their contribution to the overall results, and difficult to track the influence of the assumptions made on how information is shared [58], for example, across space and time. Moreover, inevitable errors and quirks that occur in real time epidemiological data can result in contamination of inference whereby misspecification in one part of the model adversely affects analysis in another part. This may be hard to diagnose and correct for. Computationally, fitting a full joint model typically requires intricate and dedicated programming, though this can be mitigated if full modularity is embedded in the programming language; in Section 3.6 we will discuss such computational strategies with particular reference to the `Turing.jl` language [21], one of the probabilistic programming languages we are using in the Lab. Building a full joint model is also likely to be computationally demanding as data accrues and the associated number of model parameters increases, requiring frequent fine tuning of inference algorithms.

In spite of these implementational challenges, full joint modelling has been applied successfully in a number of inferential contexts during this pandemic. For example, the University of Cambridge Medical Research Council Biostatistics Unit (BSU) and Public Health England (PHE) model uses a deterministic age-structured compartmental model [4], data on daily COVID-19 confirmed deaths, and published information on the risk of dying and the time from infection to death, as primary sources from which they estimate the number of new severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infections over time. Starting in the early months of 2020 from a pre-existing flu transmission model, the BSU-PHE model has continually been adapted and complexified. In its December 2021 release, the model accounts for the ongoing immunisation programme and latest estimates of vaccine efficacy, differential susceptibility to infection in each adult age group, and incorporates estimates of community prevalence from the Office of National Statistics COVID-19 Infection Survey [10, 57].

2.3 Interoperability of Models—the Middle Ground

Between building separate models and building a single full joint model, the Lab experience of the COVID-19 pandemic has motivated us to adopt a strategy of interoperability of models for disease surveillance. Our overarching goal is to ensure adaptability of models and modules so they can be repurposed as needed, while maintaining a consistent treatment of uncertainty, and ensuring our approach is as robust to misspecification as possible. We see interoperability as a journey, and the case studies presented in Section 4 are there to illustrate the principles and to show the direction of travel, not the fully equipped arrival lounge.

A software engineering analogy is pertinent here: we believe that there is benefit from moving from a "parallel" approach, in which several separate models are analysed simultaneously and then the results integrated by model averaging, towards a soft "serial" approach, where input and output components and loose chains of models are considered. In a straightforward serial process, one could use posterior output as input into the next model. However, doing this in a fully Bayesian manner can be as computationally demanding as a full uber-model. Furthermore, as we will illustrate in our first case study (Section 4.2), there are instances where it is beneficial to cut feedback [29]. In other cases, the serial process will be akin to approximate Bayesian melding where posterior outputs are approximated by a suitable parametric distribution [24].

3. THE MANY INGREDIENTS OF INTEROPERABILITY

Interoperability is driven by a desire to deliver timely and robust statistical inference to answer several related research questions on a common process. As such, interoperability intersects and affects many aspects of the statistical workflow, from model specification and inference to computations and data deployment.

3.1 Shared Latent Quantities

We consider there to be an important distinction between an underlying model for the scientific process of interest, S say, with parameters θ whose interpretation does not depend on what data are available, and an observation model for data D given S, with parameters ϕ. Dawid [14] calls θ and ϕ the *extrinsic* and *intrinsic* parameters, respectively. Making this distinction clarifies how a new data-source can be added to an existing model without the need to re-build the model from scratch. In the current context, our inferential focus is on extrinsic latent quantities such as incidence, prevalence or the growth rate. We note, however, that care is needed in defining precisely the inferential target to answer any particular question; for example, an unqualified reference to "prevalence" as a latent

quantity is open to multiple interpretations. In the case studies in Section 4, we use point prevalence defined as the number of individuals in the population who would be found to be PCR-positive if tested, averaged over a specified time interval (e.g., over a week for the debiasing model described in Section 4.2.1).

3.2 Modularity

Once latent quantities and scenarios of interest are specified, a common modelling framework is desirable for building each model, to facilitate principled propagation of uncertainty. We have chosen to formulate our models within the flexible framework of Bayesian hierarchical models (BHM) as it brings to the foreground conditional independence assumptions between the quantities of interest (whether observables or not) and encodes the probabilistic relationships between them. BHMs clarify information flows through the use of Directed Acyclic Graphs (DAGs), key assumptions made on exchangeability and ways of borrowing information. They also enable inclusion of new data sources in the DAGs in a coherent and computationally efficient manner.

It is often helpful to decompose a complicated model into smaller modules. For example, modules could represent different parts of the prior, the evolving dynamics of the latent epidemic, and the likelihood model for different data sources. Modularity creates a spectrum of choices between the full joint model approach and the interoperability approach.

Working within a Bayesian inference framework enables the possibility of coherent propagation of uncertainty without resorting to the direct specification of a full joint model. We wish to be able to freely specify each module separately, and then subsequently join them into a single full joint model via their common parameters. This can be accomplished via *Markov melding* [24, 41], which builds upon the ideas of Markov combination [15, 42] and Bayesian melding [56]. Markov melding can be used to join several Bayesian models that involve a common parameter, with the prior combined using a "pooling function". The joint Markov melded model is the product of the conditional distributions of each module (given the common parameter) and the pooled prior for the common parameter [24].

We apply Markov melding at various points in our case studies in Section 4. In Section 4.2, our core model combines two modules and two sources of data to infer the posterior distribution of debiased prevalence; there we use a form of Markov melding for the model's bias parameter, creating its prior via *product-of-experts* pooling, that is, the combined prior is proportional to the product of module-specific priors.

3.3 Structural Robustness

When models are constructed from multiple modules, a desirable property is that misspecification of one module does not adversely affect others. Cut models [29, 35, 54] and semi-modular inference [8] can be deployed to block, or respectively regulate, the information flow from misspecified modules to more trustworthy modules. Cut models have been applied in a broad range of areas, such as pharmacokinetic-pharmacodynamic data modeling [39, 54, 71], complex computer models [35] and epidemiology [44]. Cut and semi-modular models can be computationally challenging to fit, requiring nested MCMC by default [29], though there is currently much active work on developing more effective computationally strategies [30, 37, 54, 55, 70] as well as clarifying such models' technical properties [36, 49, 55].

We provide details of our use of cut posterior distributions in Section 4.2.2, as an example of structural robustness. In this case we are giving more weight to an unbiased-by-design source of evidence within our interoperability framework, relative to a data module that may suffer from misspecification.

As a further point of note on the use of cut posterior distributions, many epidemiological models make use of knowledge of generation intervals, serial intervals, and incubation periods. These are typically estimated from small-scale but direct studies of transmissions, producing broad confidence/credible intervals reflecting the inherent uncertainties (e.g., [3]). If these are used as priors for generation/serial intervals or incubation periods for a large scale epidemiological model such as Epimap ([66]; see Section 4.5), the lower quality yet larger amounts of information in the large scale data (e.g., test counts in case of Epimap, but can also include hospitalisation and death counts) can easily overwhelm the priors. Instead a common approach (besides Epimap, see also [7]) is to draw multiple samples from the priors of these quantities, compute the posterior predictive distribution for each sample, then aggregate across prior samples to capture the uncertainties over the generation/serial interval and incubation period. This can be equivalently viewed as nested Monte Carlo estimation for a cut posterior distribution.

3.4 Dynamic Model Validation

Adopting a common inference framework for all models (here we choose to use Bayesian inference) allows a unified interpretation of their results. Additionally, it facilitates the adoption of common ways to validate outputs, including the type of metrics to use, for example, the use of marginal predictive distributions on observables. As an example, in Section 4.2.4 and Figure 2, we are able to identify which model is performing best in terms of prediction by validating the resulting posterior estimates

against held-out gold-standard data. More generally, using a modular approach has the benefit to make each modelling component easier to validate using similar criteria on shared parameters.

3.5 Composability

Here we are concerned with computational strategy. We use the term "composability" for an approach which is sometimes called "recursive" or "two stage" in a Bayesian framework, [28, 38], and has been advocated as a way to enable computationally-efficient model exploration and cross-validation [23] and integrate different statistical software packages [31]. As examples, in our last two case studies (Sections 4.4 and 4.5), we transfer results between analyses via a Normal approximation, which can also be viewed as an approximation to Markov melding, as discussed in [24]; we discuss this approach in more detail in Section 4.2.5.

From an operational point of view, it is also easier to perform unit testing, debugging and identification of computational bottlenecks within small, self-contained modules. For example, one can perform the prior, posterior or conditional predictive checks for these components independently. Then, using the modules, one can freely choose between building a joint model by assembling these components or treating them as separate models, perform inference, and connect them using a cut or melding mechanism as introduced in Sections 3.2 and 3.3.

Considerations of interoperable modularity suggest there are benefits in targeting the marginal distributions of core quantities that feature within multiple models. This is because inference from these marginal models can then be fed into multiple downstream analyses, providing a common and coherent representation of key parameters.

3.6 Probabilistic Programming

One powerful way of performing statistical inference in an automated and timely manner is probabilistic programming, which allows one to write models in a concise, modular, intuitive syntax and automate Bayesian inference by using generic inference strategies (e.g., Gibbs sampling, Hamiltonian Monte Carlo). This significantly speeds up iterating models[4] during a preliminary data analysis phase. However, most probabilistic programming languages lack native support for interoperability. We thus based our implementation on the `Julia` programming language [32], a very fast language, specially designed for numerical computations, together with `Turing.jl` [21] an independently developed software package[5] implemented in `Julia`.

`Julia` has the advantage that it contains highly specialized implementations for specific computations, for example, convolutions, that are essential for our epidemic models, drastically improving computational performance. `Turing.jl` provides a convenient syntax for defining, for example, a standard `Julia` function that computes the log-probability of any desired generative model and samples from that model. Moreover, to fully embrace the modularisation principle, we added a new `module` feature in the `Turing.jl` language. Each specified `module` behaves like an independent model and allows us to perform all kinds of operations and diagnostics available to a full model. Moreover, the programming language allows these `modules` to be combined into more complex models, similarly to assembling elemental probability distributions. Thus, the `module` feature allows us to break a highly complex model into many composable, reusable modules. Note that the `Julia` community has extensively tested many of the dependencies of the internal packages, and that we have ourselves tested "under the hood" implementations of numerical computations such as "convolution" to ensure reproducibility.

In principle, any general-purpose programming language can implement the optimisations available in `Turing.jl`. However, Julia makes such optimisations accessible by providing high-level dynamic language features similar to Python without sacrificing performance. For example, we can implement convolutions in `Stan`, but the code is not very concise and much less reusable. These issues make `Turing.jl` more suitable for complex and computationally demanding models. Modularity via high-level language abstraction in these models is required to keep the implementation composable, interoperable, and reusable.

We experimented with the `module` feature on the Epimap model of the local reproduction number [66]. The module mechanism enables us to implement interoperability between Epimap and the Debiasing model with a minimal amount of extra work (i.e., one author implemented the code for interoperability of Epimap and Debiasing models within a day). It also allowed us to spot and fix a computational bottleneck arising from distributional changes. We discuss more details of this example of interoperability in Section 4.5.

3.7 Data Pipelines

The data deployment and its synchronisation is of paramount importance to ensure interoperable modelling, requiring meticulous data synthesis pipelines and high quality data curation and tracking. To ensure that interoperable models provide a coherent and robust set of outputs, consistent, high quality data feeds must be used. This is challenging when the generation of datasets is evolving rapidly, the data are frequently changing, and the

[4]The process of specifying and estimating models, making it practicable to explore a range of models.

[5]The name Turing given to this software package has nothing to do with The Alan Turing Institute.

downstream synthesis is continually being updated and modified. The complex datasets generated are typically being made available to a diverse user group resulting in data proliferation and redundancy. Parallel, superficially redundant data feeds may be maintained to increase resilience, analogous to the Redundant Arrays of Inexpensive Disks (RAID) storage framework [9]. Note that differing levels of information governance, dependent upon the proposed data use, dictate the granularity of data available to specific users, requiring strict data management to ensure data synchronisation.

Data curation processes may be performed at the source of the data and centrally for both basic data cleaning and sophisticated curation [69]. The tracking of metadata describing the data granularity and data curation are important to assess outputs that are comparable. All of these factors mean that the same core data may be required to be provided to users in several different forms, with varied provision and annotation of metadata. The data selection requirements for our interoperable models may also vary (e.g., the level of granularity of age or geography may be different) and so transparency of the data transformation is key to informing interoperability. To reduce data redundancy, "Single-source-of-truth" or "Master Data Management principles" could be applied to the provision of datasets. These principles suggest that only a single, master copy of each dataset is maintained, and that datasets are combined into, for example, a data warehouse by linking rather than duplication. The deployment of resilient ETL (Extract-Transform-Load) processes that implement the differing business logic (e.g., data transformations) to capture and integrate data from multiple feeds into a single data repository facilitates downstream consistent data feeds, data synchronisation and reduces redundancy. In this manner, the ETL facilitates the definition of the data transformation and processes required to manage rapidly evolving data and its underlying structure. The data warehouse is then a stable research-ready dataset that facilitates data releases and snapshots of the primary data, for example, COVID-19 test results, to specified users.

3.8 Reproducibility

Reproducibility is the principle that all policy decisions or publication should be based on analysis outputs that can be generated exactly by other researchers in their own computing environments, given the same data inputs and hyperparameter settings. Reproducible, reliable, and transparent results comprise one of the key ingredients of the data life science framework for veridical data science, that is, "principled inquiry to extract reliable and reproducible information from data, with an enriched technical language to communicate and evaluate empirical evidence in the context of human decisions and domain knowledge", as proposed in [69]. To achieve reproducibility, both input data and software code must be versioned, and each version must be retrievable at a later date. Code versioning is easily achieved using a source code management tool such as Git [22]. Data versioning can be achieved through a variety of methods, depending on the stability of the data structures. For static data structures, temporal tables (or manually timestamped records) provide the ability to query data as of a specific time. The normalised data structures described above promote static data structures; when new fields are required, a new table is used instead of adding fields to existing tables. When changes in schemas are unavoidable, regular data dumps with named versions may be more appropriate.

4. CASE STUDIES

Our case studies are chosen to illustrate some of the benefits and issues relating to interoperability that the Lab has encountered in its work. The concept of interoperability first crystallised through our work on debiased prevalence [51]. In Section 4.2, we describe this work and demonstrate how careful information synthesis (in this case cutting feedback) between data modules improved prevalence estimation. In Section 4.3, we demonstrate how we can coherently integrate and propagate the uncertainty of resulting estimates of debiased prevalence into a compartmental epidemic model.

In the final two case studies, in Sections 4.4 and 4.5, we provide additional examples of model interoperability, showing different methods of inputting debiased prevalence into other models to answer new questions. In each case study, we outline some additional advantages (e.g., computational ease) but also new questions that arise in the process (e.g., appropriate handling of uncertainty in the prevalence estimates or how to link models built for different time scales).

Overall, the case studies demonstrate how analyses in a demanding, fast-moving, public health context can be operationalised within a framework that flexibly combines a set of independent analysis modules.

4.1 Data

We first describe the data used across the case studies. All scripts and data to reproduce the results of our case studies are available online.[6]

4.1.1 *Randomized surveillance data*. These record u positive tests out of U total subjects tested. The REal-time Assessment of Community Transmission (REACT) study is a nationally representative prevalence survey of SARS-CoV-2 based on repeated polymerase chain reaction (PCR) tests of cross-sectional samples from a representative subpopulation defined through stratified random sampling from England's National Health Service patient register [60].

[6] https://github.com/alan-turing-institute/ukhsa-turing-rss-interoperability

4.1.2 *Targeted surveillance data.* These record n positive tests of N total subjects tested. Pillar 1 tests comprise *"all swab tests performed in Public Health England (*PHE*) labs and National Health Service (*NHS*) hospitals for those with a clinical need, and health and care workers"*, and Pillar 2 is defined as *"swab testing for the wider population"* [17]. Pillar $1+2$ testing has more capacity than REACT, but the protocol incurs ascertainment bias as those at higher risk of being infected are more likely to be tested, such as front-line workers, contacts traced to a COVID-19 case, or the subpopulation presenting with COVID-19 symptoms, such as loss of taste and smell [17]. The ascertainment bias potentially varies over the course of the pandemic as the testing strategy and capacity changes. We exclude lateral flow tests and use only test data from Pillar $1+2$ PCR tests.

4.1.3 *Population metadata.* We enrich the testing data by population characteristics related to the following measures of ethnic diversity and socio-economic deprivation in each local area:

- *Ethnic diversity.* The proportion of BAME (Black, Asian and Minority Ethnic) population is retrieved from the 2011 Census.
- *Socio-economic deprivation.* The 2019 Index of Multiple Deprivation (IMD) score is retrieved from the Department of Communities and Local Governments [46]. IMD is a composite index calculated at Lower Super Output Area level (LSOA) and based on several domains representing deprivation in income, employment, education, crime, housing, health and environment. For other geographies, for example, Lower Tier Local Authority (LTLA) which we use in the case studies, the IMD is obtained as the population weighted average of the corresponding LSOAs, using the 2019 mid-year population counts.

4.1.4 *Commuter flow data.* As one of its data inputs, the Epimap model [66] uses commuting flow data from the 2011 Census [67]. After preprocessing, these data are used to create a flux matrix F that determines how transmission events occur within and between lower tier local authoritys (LTLAs); see [66] for full details. In Section 4.5.4, we will discuss the relationship between flux matrix F and estimates of $\mathcal{R}^{\text{eff}}_{i,w}$.

4.2 Estimating and Adjusting for Ascertainment Bias in Pillar $1+2$ Data

We now summarise our debiasing model, which combines targeted surveillance data with randomized surveillance data to obtain local estimates of prevalence. A useful way of motivating the debiasing model is to compare the properties of the randomised and targeted surveillance datasets. REACT provides unbiased information on prevalence, though with limited sample size and publicly accessible only at a coarse spatiotemporal scale. In contrast, Pillar $1+2$ has a large sample size providing strong and biased information about fine-scale spatiotemporal variation in prevalence, but which can be useful for prevalence estimation when the bias is smooth and estimable. In other words, our model is aimed at using Pillar $1+2$ to extend REACT to a finer spatiotemporal scale. Full details can be found in [51] along with accompanying R code.[7]

4.2.1 *Debiasing model.* The REACT data provide accurate but relatively imprecise estimates of prevalence at the PHE region level (i.e., coarse scale). Note that REACT total test counts U tend to be much smaller than Pillar $1+2$ test counts N, with U/N of order 10^{-2}. The REACT data likelihood for the PCR-positive prevalence proportion π is

$$\begin{aligned}&\mathbb{P}(u \text{ of } U \mid \pi)\\ &\quad = \text{HyperGeometric}(u \mid M, \pi M, U),\\ &\qquad \text{for } \pi M \in \mathbb{Z},\end{aligned} \tag{1}$$

based on observing u positive tests out of a total of U randomly allocated in a population of known size M; our inference under (1) and (2) is based on a latent integer number of infected πM (see [51] for details). Note that, for simplicity, we are ignoring the stratified sampling design in likelihood (1). We also assume here for simplicity that the PCR-tests have perfect specificity and sensitivity; it is straightforward to incorporate known type I and II error rates within the test debiasing framework (see [51] for further details).

In contrast, test positivity rates in Pillar $1+2$ data are strongly biased upwards relative to the population prevalence proportion, as the testing is directed at the higher risk population (e.g., symptomatics, frontline workers). We show how careful modelling of the ascertainment process allows us to estimate prevalence accurately, and with good precision, even at a fine-scale level such as LTLA.

We introduce the following causal model for the observation of n of N positive targeted (e.g., Pillar $1+2$) tests:

$$\begin{aligned}&\mathbb{P}(n \text{ of } N \mid \pi, \delta, \nu)\\ &\quad = \text{Binomial}\big(n \mid \pi M, \mathbb{P}(\text{Tested} \mid \text{Infected})\big)\\ &\qquad \times \text{Binomial}\big(N-n \mid (1-\pi)M,\\ &\qquad\qquad \mathbb{P}(\text{Tested} \mid \text{NotInfected})\big)\,,\end{aligned} \tag{2}$$

where δ and ν parameterise (on log odds scale) the binomial success probabilities $\mathbb{P}(\text{Tested} \mid \text{Infected})$ and $\mathbb{P}(\text{Tested} \mid \text{NotInfected})$:

$$\delta := \log\left(\frac{\text{Odds}(\text{Tested} \mid \text{Infected})}{\text{Odds}(\text{Tested} \mid \text{NotInfected})}\right), \tag{3}$$

$$\nu := \log \text{Odds}(\text{Tested} \mid \text{NotInfected}). \tag{4}$$

[7] https://github.com/alan-turing-institute/jbc-turing-rss-testdebiasing

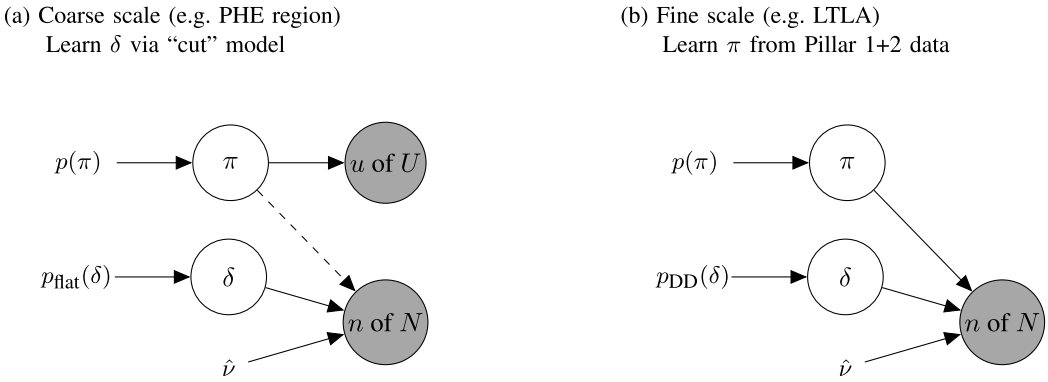

FIG. 1. *Models for debiasing.* (a) *Cut model where the dashed line represents cutting feedback from Pillar* $1+2$ *to* π. (b) *Data-dependent prior* $p_{\text{DD}}(\delta)$ *has been created from cut posterior for* δ *from stage* (*a*), *and now only Pillar* $1+2$ *data are used to infer* π *at the local* (LTLA) *level.*

By causal, we mean that we explicitly take into account the way the Pillar $1+2$ data was generated by inferring, and conditioning in (2), on the probability of individuals in the population being tested. We provide a detailed description of the model in [51]. The parameter requiring most careful treatment is δ, that is, the log odds ratio of being tested in the infected versus the noninfected subpopulations.

Our default approach for the other parameter, ν, is to use the plug-in estimator $\hat{\nu} := \text{logit}[(N-n)/M]$ in likelihood (2), because it allows fast and tractable computation, and is precise with little bias when prevalence is low. While this allows for efficient computation, heuristically using the sub-optimal plug-in estimator can lead to model misspecification because of its bias when prevalence is high, a point we return to in Section 4.2.3.

4.2.2 *Structurally robust estimation of ascertainment bias* δ. In Figure 1(a), the joint posterior distribution (without cutting any information flow) is

$$\begin{aligned} &p(\pi, \delta \mid u \text{ of } U, n \text{ of } N, \nu) \\ &\quad = p(\pi \mid u \text{ of } U, n \text{ of } N, \nu) \times p(\delta \mid \pi, n \text{ of } N, \nu), \end{aligned} \tag{5}$$

and we use the plug-in estimator $\nu = \hat{\nu} \equiv \text{logit}[(N-n)/M]$ for computational convenience, as introduced in Section 4.2.1. This provides effective inference when the model space contains the true data generating mechanism and $\hat{\nu}$ is not too biased. However, if the Pillar $1+2$ likelihood (2) is misspecified for any reason, then inference on π, and hence on δ, can be adversely affected. In the current context, the consequences of misspecification are particularly severe because, conditional on δ, the relative sample sizes lead to the Pillar $1+2$ likelihood (2) typically containing far more information on π than the REACT likelihood (1).

With this in mind, we use a cut posterior distribution, as described in [29]:

$$\begin{aligned} &p^{\text{cut}}(\pi, \delta \mid u \text{ of } U, n \text{ of } N, \nu) \\ &\quad := p(\pi \mid u \text{ of } U) \times p(\delta \mid \pi, n \text{ of } N, \nu) \end{aligned} \tag{6}$$

again with plug-in estimator $\nu = \hat{\nu}$. where the first distribution on the RHS of (6) is no longer conditioning on the n of N from Pillar $1+2$. Switching from model (5) to (6) "cuts feedback" from Pillar $1+2$ to π in inference on (π, δ). Figure 2(a)–(b) compares the joint full posterior (5) with the joint cut posterior (6) for Pillar $1+2$ and REACT data gathered in London during the week commencing (W/C) 14th Jan 2021. In this week, the Pillar $1+2$ data were n of $N = 60{,}749$ of 326,986 and the REACT data were u of $U = 101$ of 3,778, that is, the Pillar $1+2$ data have 87 times as many tests as REACT that week.

The joint posteriors in Figure 2(a)–(b) have clearly quite different support. The mean (95% CI) for δ under the full posterior is 3.5 (3.2 −3.9) whilst under the cut posterior it is 2.4 (2.2 −2.7). Which model, full or cut, is estimating δ more accurately? Note that in the cut posterior (6) the marginal distribution of the prevalence proportion π depends only on the REACT data,

$$p^{\text{cut}}(\pi \mid u \text{ of } U, n \text{ of } N) = p(\pi \mid u \text{ of } U), \tag{7}$$

and that the maximum likelihood estimator for π based on model $\mathbb{P}(u \text{ of } U \mid \pi)$ at (1) is approximately unbiased, since REACT is a designed, randomised study. We say that the estimator is approximately unbiased as, for simplicity, we do not account for the stratified sampling design nor for the nonresponse ($> 75\%$), which we assume is noninformative. Thus, under a weakly informative prior $p_{\text{flat}}(\pi)$, the cut-posterior π-marginal mean (95% CI) of 2.7 (2.2 −3.2) estimates π reasonably accurately. However, the full-posterior π-marginal mean of 1.4 (1.1 −1.6) is quite different, suggesting that the Pillar $1+2$ causal testing model at (2) is misspecified and, having a relatively large amount of data, swamps the accurate information in the REACT data.

4.2.3 *Exploring model misspecification.* We initially posit two explanations for the model misspecification identified in Figure 2(a)–(b):

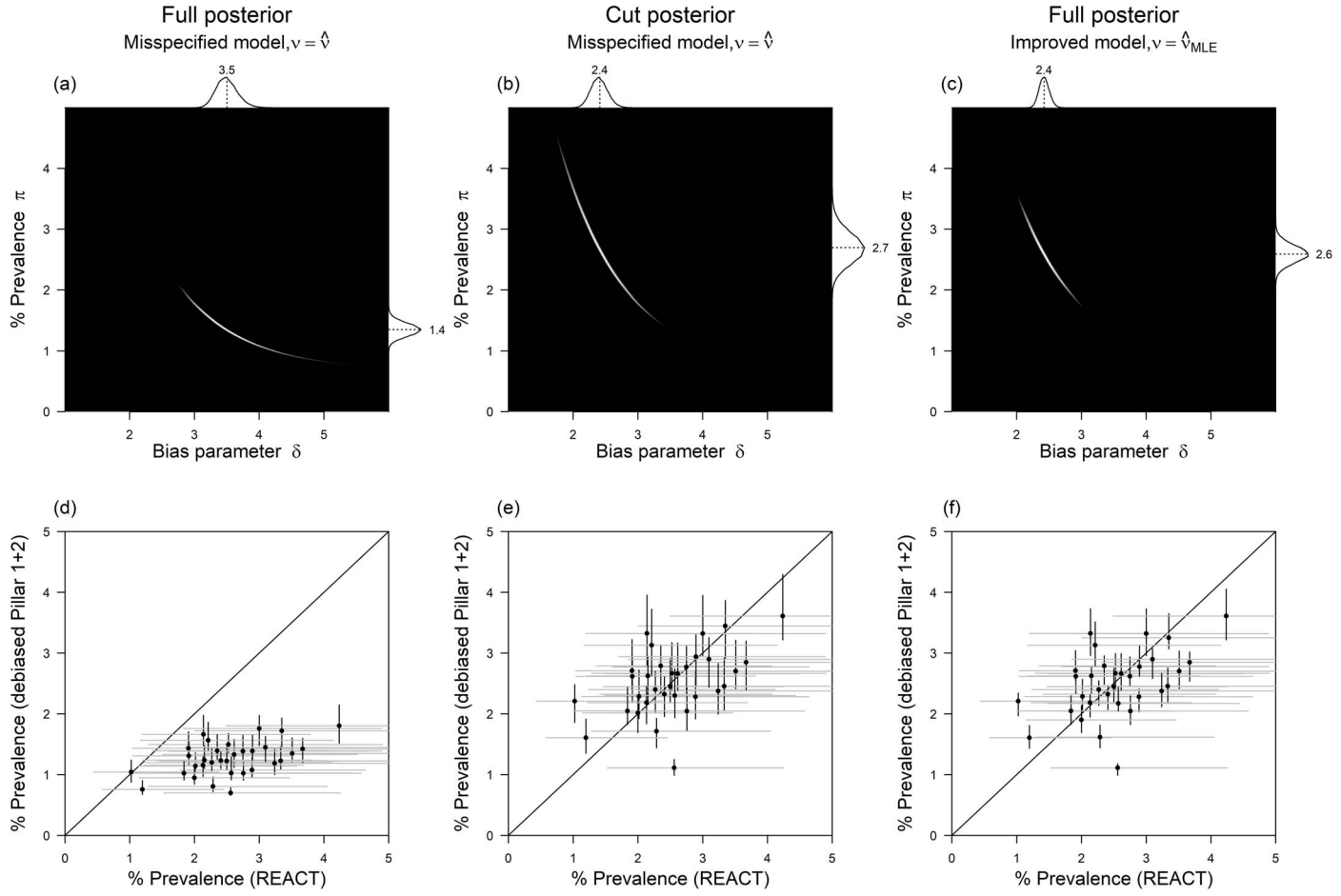

FIG. 2. *Comparison of full vs cut posterior inference for bias parameter* δ *under different inference strategies for* ν. (*a*) Full, misspecified *joint log posterior—logarithm of equation* (5) *with* $\nu = \hat{\nu}$. (*b*) Cut, misspecified *joint log posterior—logarithm of equation* (6) *with* $\nu = \hat{\nu}$. (*c*) Full, improved *joint log posterior—logarithm of equation* (5) *with* $\nu = \hat{\nu}_{\text{MLE}}$. (*d*) LTLA-*level debiased % prevalence based on* full, misspecified *posterior* (*vertical*) *vs gold-standard* REACT *% prevalence* (*horizontal*), *one point per* LTLA. (*e*) LTLA-*level debiased % prevalence based on* cut, misspecified *posterior* (*vertical*) *vs gold-standard* REACT (*horizontal*). (*f*) LTLA-*level debiased % prevalence based on* full, improved *posterior* (*vertical*) *vs gold-standard* REACT (*horizontal*). *In panels* (*a-c*) *the marginal posterior distributions are plotted at the panel top and right edges, with marginal means labelled. Inference is based on* REACT *and Pillar* $1+2$ *data for London and its constituent* LTLAs *for* W/C 14*th Jan* 2021. *The gold-standard* LTLA-*level estimates used for validation on the horizontal axes in* (*d-f*) *are based on* REACT *round* 8 *aggregated data* (6*th*-22*nd Jan* 2021).

(E1) Violation of the Binomial distributional assumption in (2) through over-dispersion caused by systematic variation of δ, ν and π across subpopulations within an LTLA having different COVID-19 beliefs and behaviours;

(E2) Bias in the plug-in estimator $\hat{\nu} = \text{logit}[(N - n)/M]$ introduced in Section 4.2.1.

Upon further investigation, we conjecture that explanation (E2) is the main source of misspecification because, while it allows efficient computation, using the sub-optimal plug-in estimator $\nu = \hat{\nu}$ can lead to model misspecification because of its bias when prevalence is high. To illustrate this, we note that a better plug-in approach would instead use the unbiased MLE estimator $\hat{\nu}_{\text{MLE}} := \text{logit}[(N - n)/\{M(1 - \pi)\}]$. However, estimator $\hat{\nu}_{\text{MLE}}$ is more computationally expensive as it requires the unknown prevalence π as an input (and this is why we do not apply $\hat{\nu}_{\text{MLE}}$ as our default).[8] To target $\hat{\nu}_{\text{MLE}}$ we implement Algorithm 1, which estimates ν according to its π-conditioned maximum likelihood estimator $\hat{\nu}_{\text{MLE}} := \text{logit}[(N - n)/\{M(1 - \pi)\}]$ alternately with estimating π. For the current example, this leads to the estimate $\hat{\nu}_{\text{MLE}} = -3.460$ which is to be contrasted with the biased, computationally convenient estimate $\hat{\nu} = -3.486$.

The joint (π, δ) full posterior downstream of this improved estimator $\hat{\nu}_{\text{MLE}}$ is shown in Figure 2(c), and has π-marginal mean (95% CI) of 2.6 (2.3 −2.9), which is now compatible with the cut-posterior's estimate of 2.7 (2.2 −3.2) shown in in Figure 2(b). It is interesting to compare the width of these intervals, noting that the pos-

[8]The most principled approach would be jointly to infer (π, δ, ν), though again at a greater computational cost (we do not fit the joint model here).

Algorithm 1 Fixed point algorithm targeting $\hat{\nu}_{\text{MLE}}$

Input: REACT u of U, Pillar $1+2$ n of N, and population size M

$\hat{\pi} \leftarrow 0$

repeat

$\hat{\nu} \leftarrow \text{logit}\, \frac{N-n}{(1-\hat{\pi})M}$

$\hat{\pi} \leftarrow \underset{\pi}{\text{argmax}}\, p(\pi \mid u \text{ of } U, n \text{ of } N, \hat{\nu})$

where $p(\pi \mid \cdot)$ here is (5) marginalized wrt δ

until convergence

$\hat{\nu}_{\text{MLE}} \leftarrow \hat{\nu}$

Output: $\hat{\nu}_{\text{MLE}}$

terior distribution in Figure 2(c) is relatively concentrated compared to the one in Figure 2(b). This reflects the extra information on π provided by the Pillar $1+2$ data in the full model, relative to the cut model in which information flow from Pillar $1+2$ to π is blocked, as illustrated in Figure 1(a).

It is a matter of taste and expediency whether one chooses to fit the full joint model, or to use the unbiased plug-in $\hat{\nu}_{\text{MLE}}$, or to use the biased plug-in $\hat{\nu}$ repaired by cut modelling. In our case, the heuristic estimator $\hat{\nu}$ applied with a cut model is far more computationally efficient, and we are able to validate the overall method to be unbiased. It is an interesting choice philosophically to incorporate a deliberately misspecified module, which is then repaired using cut modelling, into the the overall model. Our choice reflected our priority on computationally tractability.

Note that the bias of $\hat{\nu}$ relative to $\hat{\nu}_{\text{MLE}}$ tends to be larger for high prevalence and will have a greater impact when the Pillar $1+2$ sample size is large (both of which apply to the illustrative example of London w/c 14th Jan 2021, where prevalence is estimated at 2.7% and $N = 326{,}986$). We further compare and contrast inference under the full and cut posteriors after having applied them to estimate local prevalence in Section 4.2.4.

4.2.4 *Inferring cross-sectional local prevalence*. We specify a data-dependent prior $p_{\text{data-dependent(DD)}}(\delta_{J,w})$ for each week w in each region J, based on the posterior distribution at (6). We begin by approximating the marginal posterior distribution for $\delta_{J,w}$ with a moment-matched Gaussian based on the cut posterior:

$$p_{\text{DD}}(\delta_{J,w}) := \text{Normal}(\delta_{J,w} \mid \hat{\mu}_{J,w}, \hat{\tau}^2_{J,w}) \tag{8}$$

$$\approx p^{\text{cut}}(\delta_{J,w} \mid u_{J,w} \text{ of } U_{J,w}, n_{J,w} \text{ of } N_{J,w}). \tag{9}$$

For example, in the case of the marginal density at the top of Figure 2(b), we specify $p_{\text{DD}}(\delta_{J,w})$ by setting $\hat{\mu}_{J,w} = 2.4$ and $\hat{\tau}_{J,w} = 0.1$ for London W/C 14th Jan 2021. Each of the three inferential approaches shown columnwise in Figure 2 will yield its own data-dependent prior in the form of (8) which approximates the δ-marginal posteriors at the top of panels Figure 2(a)–(c), and which then feeds forward into the cross-sectional debiased % prevalence estimates on the vertical axes of Figure 2(d)–(f) respectively.

While this approach provides a prior for weeks at which both Pillar $1+2$ and REACT data are available (since we are able to estimate δ), we also wish to interpolate and/or extrapolate information on δ to weeks at which REACT data are unavailable (e.g., between sampling rounds). We achieve this by introducing a smoothing component into a product-of-experts prior ([27]; details in Appendix A), thereby allowing us to specify independent priors $p_{\text{DD}}(\delta_{J,w})$ of the form (8) for all weeks, including those without REACT data.

Having specified a prior on δ at the coarse-scale regions J, we proceed to perform full Bayesian inference at a fine-scale LTLA i using the prior from its corresponding region, $p_{\text{DD}}(\delta_{J[i],w})$. We plot cross-sectional % prevalence posterior medians (with 95% posterior CIs) on the vertical axes of Figure 2(d)–(f), with each point corresponding to the estimated % prevalence for one LTLA in London for W/C 14th Jan 2021. Observe in Figure 2(d)–(f) that the horizontal REACT CIs are relatively broad, compared to the vertical debiased Pillar $1+2$ CIs, exemplifying the relatively large amount of information in the Pillar $1+2$ data if δ can be inferred.

The results in Figure 2(d)–(f) are also informative for our discussion of inference under the various models, because we are able to validate against "gold-standard" LTLA-level randomised surveillance data aggregated across round 8 of the REACT study (6th–22nd Jan 2021). On the horizontal axis, we plot the REACT unbiased prevalence estimates (with 95% exact binomial CIs). Compared to the accurate REACT estimates, the debiased prevalence estimates based upon the full posterior in Figure 2(d) are biased downwards—the estimated bias across the LTLAs plotted is -1.28% (SE $= 0.11\%$). In contrast, the debiased prevalence estimates based on the cut posterior in Figure 2(e) appear to be accurate, having estimated bias of -0.01% (SE $= 0.11\%$). The full posterior estimates based on improved plug-in estimator $\hat{\nu}_{\text{MLE}}$ in Figure 2(f) also appear accurate, having estimated bias of -0.04% (SE $= 0.11\%$).

4.2.5 *General interoperability between models*. The graph in Figure 3(a) relates the debiasing model (in the i, w plate) to another arbitrary model with parameters (π, θ) and data/covariates Y (with Y not containing Pillar $1+2$). One approach to fitting Figure 3(a)'s model would be to perform full Bayesian inference directly, that is, sampling from (θ, π, δ). However, as will be illustrated with the SIR model in Section 4.3, we can reduce the computational complexity by first marginalising with respect

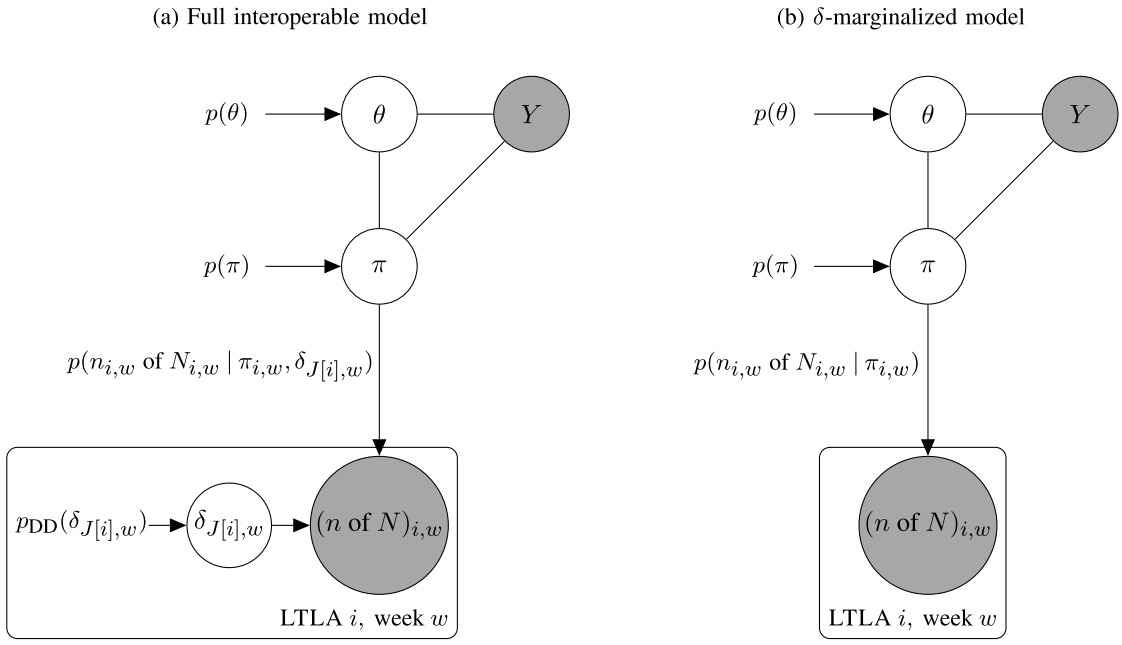

FIG. 3. *Interoperable interface of debiasing output with a model parameterised by θ and with other data Y. (a) Full interoperable model. (b) Collapsed representation, in which the full model has been marginalized wrt δ. We use π to denote the entire spatiotemporal collection of prevalence proportions $\pi_{1:I,1:W}$.*

to $\delta_{J[i],w}$ yielding the marginal likelihood

$$
\begin{aligned}
&p(n_{i,w} \text{ of } N_{i,w} \mid \pi_{i,w}) \\
&\quad = \int p(n_{i,w} \text{ of } N_{i,w} \mid \pi_{i,w}, \delta_{J[i],w}) \\
&\qquad \times p_{\mathrm{DD}}(\delta_{J[i],w})\, d\delta_{J[i],w},
\end{aligned} \tag{10}
$$

which is an unnormalized function of $\pi_{J,w}$ encapsulating all information on $\pi_{J,w}$ that results from observing $n_{i,w}$ of $N_{i,w}$ positive Pillar $1+2$ tests. As illustrated in Figure 3(b), we then need sample only (θ, π) in the δ-marginalised interoperable model.

While the marginal likelihood $p(n_{i,w}$ of $N_{i,w} \mid \pi_{i,w})$ at (10) can easily be evaluated pointwise, it does not have a closed parametric form. We can further simplify inference at the interface between models by approximating (up to a multiplicative constant) the marginal likelihood with a parametric distribution. A Gaussian moment-matched approximation on log odds scale (Figure 3(b)),

$$
\begin{aligned}
&\hat{p}(n_{i,w} \text{ of } N_{i,w} \mid \pi_{i,w}) \\
&\quad \overset{\pi_{i,w}}{\propto} \mathrm{Normal}\big(\operatorname{logit} \hat{\pi}_{i,w} \mid \operatorname{logit} \pi_{i,w}, \hat{\sigma}^2_{i,w}\big),
\end{aligned} \tag{11}
$$

is a natural choice here because it provides an empirically good fit and also integrates conveniently with methods and software for hierarchical generalised linear models with logit link function, as we shall see in Section 4.4. This approach of summarising a module with a moment-matched Gaussian distribution for use as an approximate likelihood term in subsequent modules has been previously widely used in simpler contexts, including evidence syntheses [68] and more general hierarchical models [12, 38].

Depending on the context, either one of (10) or (11) may be preferred. Using the exact marginal likelihood at (10) avoids making a Gaussian approximation, but (10) can be computationally unwieldy as it is a mass function on integer prevalence πM. The Gaussian approximated marginal likelihood at (11) is often more computationally convenient to integrate with other models. We use (10) in Section 4.3 and we use (11) in Sections 4.4 and 4.5.

4.3 Interoperability with an SIR Model

The marginal likelihood $p(n_{i,w}$ of $N_{i,w} \mid \pi_{i,w})$ from (10) can be used to link the Pillar $1+2$ data directly to latent prevalence nodes in a graphical model. As a concrete example, we implemented a full Bayesian version of the standard stochastic SIR model [6, 51, 63]. Figure 4 illustrates the SIR model DAG, relating prevalence proportion $\pi_{i,w}$, effective reproduction number $\mathcal{R}^{\mathrm{eff}}_{i,w}$, and test data $(n$ of $N)_{i,w} \equiv n_{i,w}$ of $N_{i,w}$. Note that Figure 4 is a special case of Figure 3(b) with data node Y empty and $\theta = \mathcal{R}^{\mathrm{eff}}_{i,1:W}$. In Figure 4, each $\pi_{i,w}$ is linked to its corresponding test data $(n$ of $N)_{i,w}$ via the marginal likelihood $p(n_{i,w}$ of $N_{i,w} \mid \pi_{i,w})$ of (10). Consecutive $\pi_{i,w}$ and $\mathcal{R}^{\mathrm{eff}}_{i,w}$ nodes are related by a discrete time Markov chain, in which the stochastic change in the number infected at week w, relative to week $w-1$, is modelled as the difference between a Poisson-distributed number of new infections and a Binomial-distributed number of new

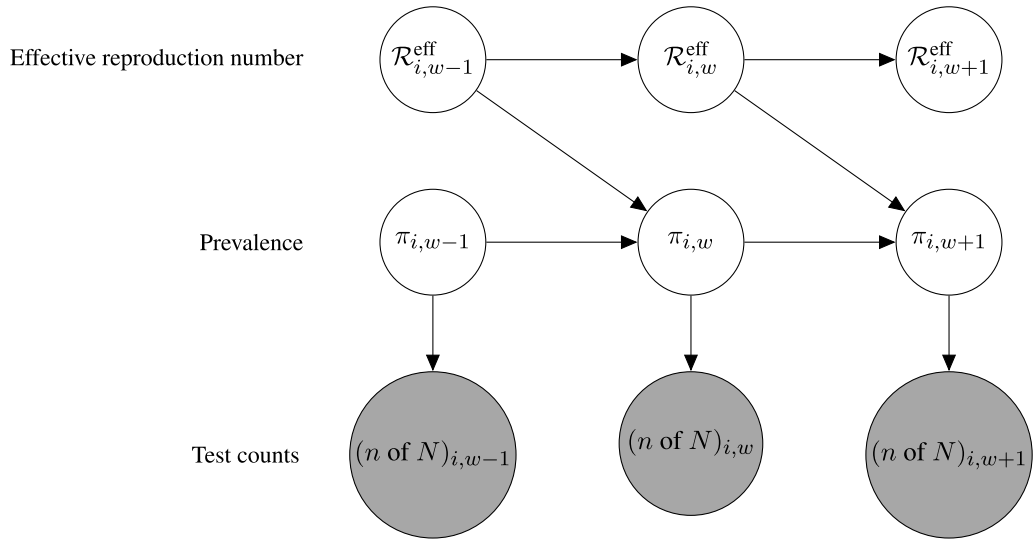

FIG. 4. *Directed acyclic graph showing interoperability of the debiasing model output with a stochastic SIR epidemic model. Latent nodes for* LTLA *i's prevalence and effective reproduction number in week* w, $\pi_{i,w}$ *and* $\mathcal{R}^{\text{eff}}_{i,w}$ *respectively, are shown, along with weekly Pillar* $1+2$ *test counts, integrated via the debiasing model-outputted marginal likelihood at* (10). *Details of the discrete time Markov chains on the latent nodes are given in equations* (12)-(14).

recoveries:

$$\begin{aligned}(12)\qquad &\pi_{i,w}M_i - \pi_{i,w-1}M_i \mid \pi_{i,w-1}, \mathcal{R}^{\text{eff}}_{i,w-1}\\ &\quad = \#\text{ new infections} - \#\text{ recoveries},\end{aligned}$$

$$\begin{aligned}(13)\qquad &\#\text{ new infections} \mid \pi_{i,w-1}, \mathcal{R}^{\text{eff}}_{i,w-1}\\ &\quad \sim \text{Poisson}\big(\gamma \mathcal{R}^{\text{eff}}_{i,w-1}\pi_{i,w-1}M_i\big),\end{aligned}$$

$$\begin{aligned}(14)\qquad &\#\text{ new recoveries} \mid \pi_{i,w-1}\\ &\quad \sim \text{Binomial}(\pi_{i,w-1}M_i,\ \gamma);\end{aligned}$$

γ is the (pre-specified) probability of recovery from one week to the next; and the $\mathcal{R}^{\text{eff}}_{i,w}$ are modelled sequentially by an AR1 process. We sample from the full Bayesian posterior for this SIR model using MCMC methods (see [51] for details).

Figure 5 compares, for three example LTLAs, the cross-sectional posteriors for π with the SIR-model MCMC-sampled longitudinal posteriors for π and $\mathcal{R}^{\text{eff}}$. The width of the SIR posterior 95% credible intervals are often much narrower than the cross-sectional posterior CI width, illustrating the benefit of sharing prevalence information across time points within the framework of an epidemiological model. Fitting the full Bayesian SIR model provides posterior credible intervals on the effective reproduction number $\mathcal{R}^{\text{eff}}$ (Figure 5 bottom panels), which is an important measure of spatiotemporally local rates of transmission. In Section 4.5, we compare these estimates of local $\mathcal{R}^{\text{eff}}$ (based on Pillar $1+2$ from a single LTLA only) with spatially smoothed estimates of local $\mathcal{R}^{\text{eff}}$ from Epimap [66].

4.4 Interoperability Between Debiasing and Space–Time Equality Analysis

There is extensive evidence to suggest that ethnically diverse and deprived communities have been differentially affected by the COVID-19 pandemic in the UK [43, 47, 61]. It is thus very important to be able to relate unbiased prevalence estimates such as those introduced in Section 4.2 to area level covariates, in order to assess associations between the spread of the virus and particular population characteristics. Given the infectious nature of the disease, residual heterogeneity is likely to have a spatio-temporal structure, which needs to be accounted for in the model. Failing to account for autocorrelation may result in narrower credible intervals for parameters of interest and may inflate covariates effects [33]. One option to incorporate sources of spatial autocorrelation in estimates of disease surveillance metrics such as prevalence, would be to run the debiasing model presented in Section 4.2.4 augmented by a spatio-temporal prior structure both on the $\pi_{i,w}$ and on the ascertainment bias δ_i, while simultaneously adding the population characteristics as covariates to assess their impact on the spread of the disease. However, such a strategy would entail a prohibitive computational cost, not only due to the large number of latent variables typically involved in the specification of even a relatively simple spatio-temporal model, but also because of implementing a cut model in such a high-dimensional setting. As an alternative, in this section we illustrate two interoperable models which differ in how they treat the uncertainty of prevalence estimates.

We focus on ethnic composition and socio-economic deprivation as covariates of interest, to assess the impact

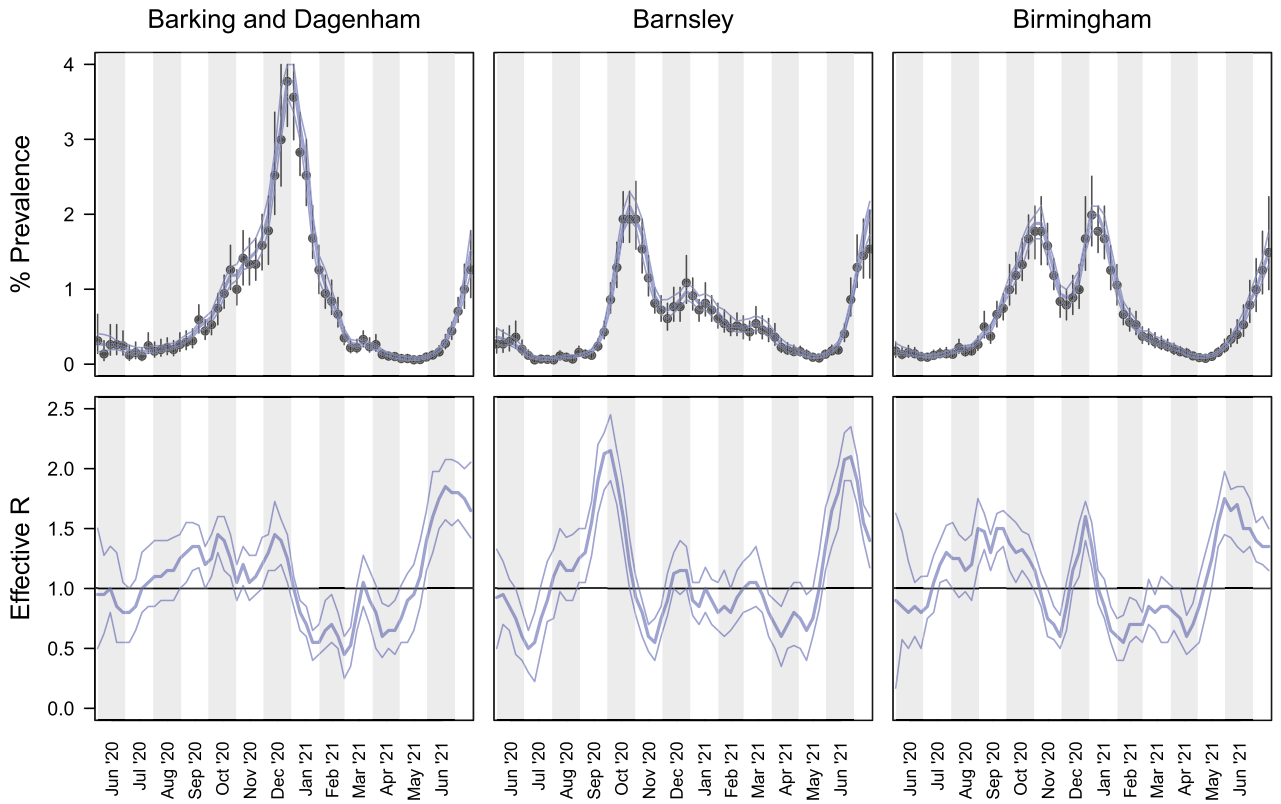

FIG. 5. *SIR longitudinal posterior (% prevalence and effective reproduction number) compared with cross-sectional posterior for % prevalence. Top panels: cross-sectional % prevalence posterior median and 95% CIs (black points and whiskers), and SIR modelled % prevalence posterior median and 95% CIs (blue curves). Bottom panels: SIR modelled longitudinal $\mathcal{R}_{i,w}^{\text{eff}}$ posterior median and 95% CIs.*

of socio-economic factors on the evolution of the pandemic. As our aim is to assess a link between prevalence proportion and these two factors, first we specify a probability distribution for the outcome, $\pi_{i,w}$ as

$$\text{logit}(\pi_{i,w})|\eta_{i,w},\sigma^2 \sim \text{Normal}(\eta_{i,w},\sigma^2), \tag{15}$$

where $\eta_{i,w}$ is a linear predictor containing the area level variables of interest as well as space and time structured random effects defined at (18) below, while σ^2 is the variance of the error term.

Since $\pi_{i,w}$ is unknown, it is not possible to fit this model directly. A first interoperable approach, which we call the *naive model*, simply consists of plugging in the debiased prevalence proportion point estimates (e.g., median) $\widehat{\pi}_{i,w}$, outputted by the debiasing model of Section 4.2:

$$\text{logit}(\widehat{\pi}_{i,w})|\eta_{i,w},\sigma^2 \sim \text{Normal}(\eta_{i,w},\sigma^2). \tag{16}$$

This model considers the prevalence estimates as fixed quantities, neglecting their uncertainty.

A second interoperable approach accounts for the posterior uncertainty from the debiased prevalence model via (11), similarly to [19, 53]. In practice, for the ith LTLA and wth week this *heteroscedastic model* reformulates (16) to include the variance component $\hat{\sigma}_{i,w}^2$:

$$\text{logit}(\widehat{\pi}_{i,w})|\eta_{i,w},\sigma^2 \sim \text{Normal}(\eta_{i,w},\hat{\sigma}_{i,w}^2+\sigma^2). \tag{17}$$

In both models, we use the following specification for the linear predictor $\eta_{i,w}$, in order to assess the effect of the area level variables of interest (proportion of BAME and IMD score) on the prevalence estimates:

$$\eta_{i,w} = \beta_0 + \beta_1 \texttt{BAME}_i + \beta_2 \texttt{IMD}_i + \lambda_i + \epsilon_w, \tag{18}$$

where β_0 is the global intercept, $\{\beta_1,\beta_2\}$ quantify the effects of the covariates of interest, λ_i denotes the area specific random effect accounting for spatial autocorrelation, and ϵ_w represents the temporal random effect. Details on the model specification can be found in Appendix B.

Output from the two interoperable modelling approaches presented above is used to examine the effects of socio-economic factors (i.e., ethnic diversity and socio-economic deprivation) on COVID-19 unbiased prevalence.[9] Additionally, the models allow us to characterise the baseline spatial distribution of prevalence across LTLAs in England, *after accounting for their socio-economic and ethnic profiles*. This enables us, for example, to identify LTLAs that have been particularly badly affected by COVID-19 relative to their level of deprivation and size of BAME population, and as a result may require further consideration by policy-makers.

[9]For a detailed analysis of the effect of time varying association between deprivation, ethnicity and SARS-CoV-2 infections in England, please see [52].

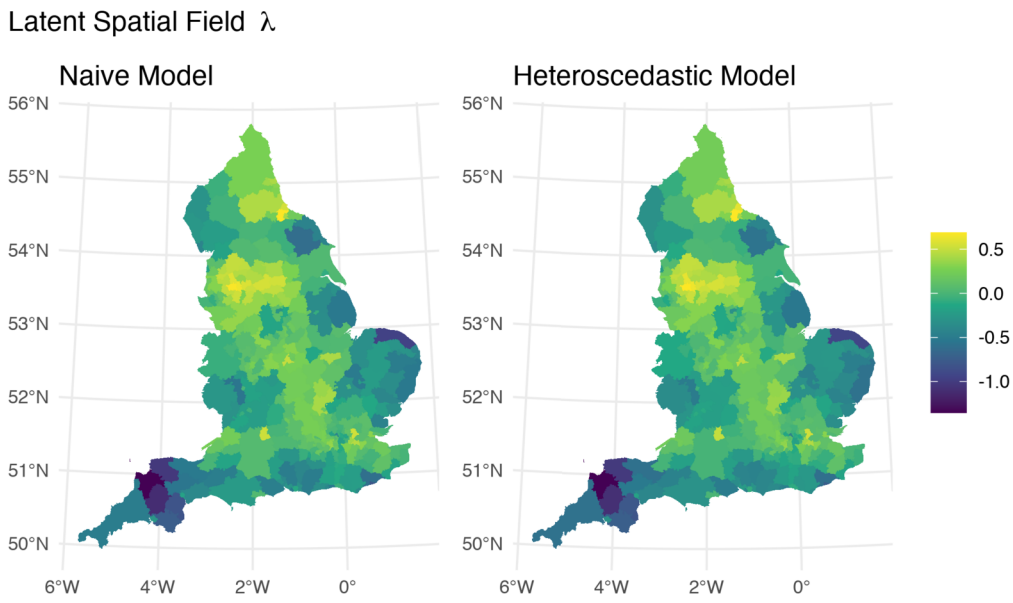

FIG. 6. *Posterior median of the spatial random effect* $\boldsymbol{\lambda}$.

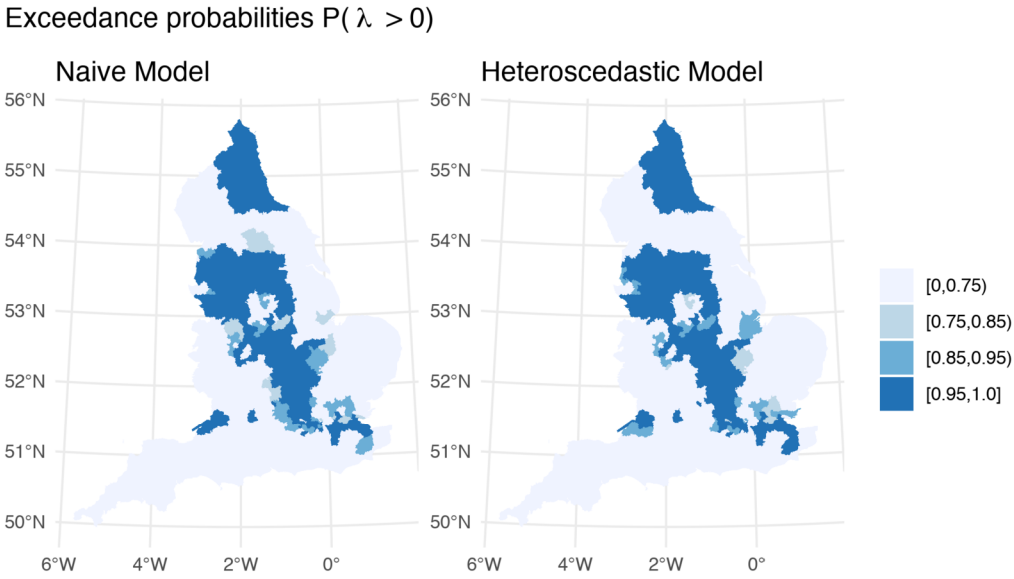

FIG. 7. *Posterior probability of having a positive spatial residual.*

TABLE 1

Posterior median and corresponding 95% *Credible Interval for the parameter estimates of the Naive (left) and Heteroscedastic (right) space–time equality analyses. Estimates for the fixed effect coefficients are on the Odds Ratio scale. The precision of the time and spatial random effects,* $1/\sigma_\epsilon^2$ *and* τ*, are defined in Appendix B*

	Naive		Heteroscedastic	
	Median	95% CI	Median	95% CI
BAME effect (β_1)	1.22	(1.17, 1.28)	1.20	(1.15, 1.25)
IMD effect (β_2)	1.11	(1.07, 1.16)	1.11	(1.07, 1.15)
Precision of the Gaussian residuals ($1/\sigma^2$)	2.67	(2.59, 2.73)	4.05	(3.95, 4.15)
Precision of time random effect ($1/\sigma_\epsilon^2$)	30.41	(13.72, 48.51)	31.37	(19.38, 52.29)
Precision of spatial random effect (τ)	10.89	(9.03, 13.28)	12.26	(10.00, 14.56)

While both models are examples of interoperability, they differ in how they handle uncertainty in the debiased prevalence proportion outcome measure; the naive model treats the prevalence estimates as fixed quantities, while the heteroscedastic model propagates their uncertainty through an additional variance term. This difference is predominantly reflected in the uncertainty of parameter estimates. Although the posterior median point estimates for the effects of IMD and BAME (Table 1) and the underlying spatial fields (Figure 6) are largely indistinguishable, the precision for the global error term (i.e., $1/\sigma^2$) is almost twice as large in the heteroscedastic model. Precision for the spatial component is also larger in the heteroscedastic model compared to the naive model, suggesting that the global error and the spatial random effect in the naive model were capturing part of the spatial variability present in the prevalence proportion estimates. It is also interesting to note that the exceedance probability map corresponding to the heteroscedastic model is sharper, showing that the ranking of areas deviating from the national average is influenced by the inclusion of uncertainty on the debiased prevalence estimates (Figure 7).

4.5 Interoperability Between the Debiasing and Epimap Models

Epimap is a hierarchical Bayesian method for estimating the local instantaneous reproduction number $\mathcal{R}_{i,t}^{\text{eff}}$ that models both temporal (day t) and spatial (LTLA) dependence in transmission rates [66]. Epimap incorporates information on population flows to model transmission between local regions, and performs spatiotemporal smoothing on the $\mathcal{R}_{i,t}^{\text{eff}}$. The data inputted to Epimap are daily positive Pillar $1+2$ test counts at LTLA level, the $n_{i,t}$ in our notation.

4.5.1 *Epimap model overview.* The observation model for $n_{i,t}$ is an overdispersed negative binomial model with mean $E_{i,t}$ given by a convolution of a testing delay distribution $D_s^{\text{Test+}}$ with the past incidences $X_{i,1:t}$ [20], that is,

$$n_{i,t} \mid X_{i,1:t} \sim \text{NegBin}(E_{i,t}, \phi_i), \tag{19}$$

$$E_{i,t} \equiv \sum_{s=0}^{t} X_{i,t-s} D_s^{\text{Test+}}, \tag{20}$$

where $\phi_i \sim \mathcal{N}_+(0,5)$.

LTLA i's incidence $X_{i,t}$ is probabilistically modelled conditional on past incidences $X_{1:n,1:t-1}$ via local and cross-coupled infection loads, denoted $Z_{i,t}$ and $\tilde{Z}_{i,t}$ respectively. Specifically, the local infection load $Z_{i,t}$ is given by a convolution of W_s with the past incidences $X_{i,1:t}$, where the generation distribution W_s is the probability that a given transmission event occurs s days after the primary infection. The local infection loads contribute transmission events not only locally but also to other regions, giving the cross-coupled infection load $\tilde{Z}_{i,t}$, with inter-regional transmission defined via a flux matrix F, built on the commuter flow data introduced in Section 4.1.4, in which F_{ji} denotes the probability that a primary case based in area j infects a secondary case based in area i. In summary, the local and cross-coupled infection loads are defined as

$$Z_{i,t} \equiv \sum_{s=1}^{t} X_{i,t-s} W_s, \qquad \tilde{Z}_{i,t} \equiv \sum_{j=1}^{n} F_{ji}^{(t)} Z_{j,t}. \tag{21}$$

The incidence $X_{i,t}$ follows an overdispersed negative binomial distribution with mean given by the product of the reproduction number $\mathcal{R}_{i,t}^{\text{eff}}$ with the cross-coupled infection load $\tilde{Z}_{i,t}$:

$$X_{i,t} \mid \mathcal{R}_{i,t}^{\text{eff}}, X_{1:n,1:t-1} \sim \text{NegBin}\big(\mathcal{R}_{i,t}^{\text{eff}} \tilde{Z}_{i,t}, \phi\big). \tag{22}$$

Note that the $\mathcal{R}_{i,t}^{\text{eff}}$ are our primary inferential target, and are estimated via the posterior distribution of the ratio $X_{i,t}/\tilde{Z}_{i,t}$. In Figures 8 and 9 below, we present the posterior distribution of the weekly averaged ratio, $\mathcal{R}_{i,w}^{\text{eff}} := \frac{1}{7}\sum_{t \in w} X_{i,t}/\tilde{Z}_{i,t}$. A final aspect of the Epimap model is the smoothing on $\mathcal{R}_{i,t}^{\text{eff}}$, whereby information is shared across time and space through specification of various Gaussian processes on $\log \mathcal{R}_{i,t}^{\text{eff}}$ (see [66] for details).

4.5.2 *Probabilistic interface between Epimap $\mathcal{M}_{\text{E}}$ and debiasing model $\mathcal{M}_{\text{D}}$.* There are three immediate and important differences between Epimap and the debiasing model that require attention at the model interface. First, Epimap's measure of infection burden is incidence (i.e., the number of new infections contracted in a time interval), whilst the debiasing model is based on point prevalence, as we defined in Section 3.1. Second, Epimap is at daily frequency, indexed t, while the debiasing model is at weekly frequency, indexed w. Third, some LTLAs were recently merged to create a more coarse-scale local geography, and the debiased model works with the newer coarser LTLA geography, while Epimap still works with the older finer-scale LTLA geography.

To map from incidence to prevalence, we draw from the existing COVID-19 literature on the probability of testing PCR positive when swabbed s days post infection [26]; we denote this as

$$D_s^{\text{PCR+}} := \mathbb{P}(\text{would test PCR+ on day } s \mid \text{contract virus day 0}). \tag{23}$$

To address the daily-to-weekly mapping, we average the daily prevalence proportion (mapped from daily incidence) across the days of each week. To address the models' differences in LTLA geography, we are able straightforwardly to preserve the geographies of both models by deterministically aggregating Epimap's latent incidences as part of the mapping between models described below

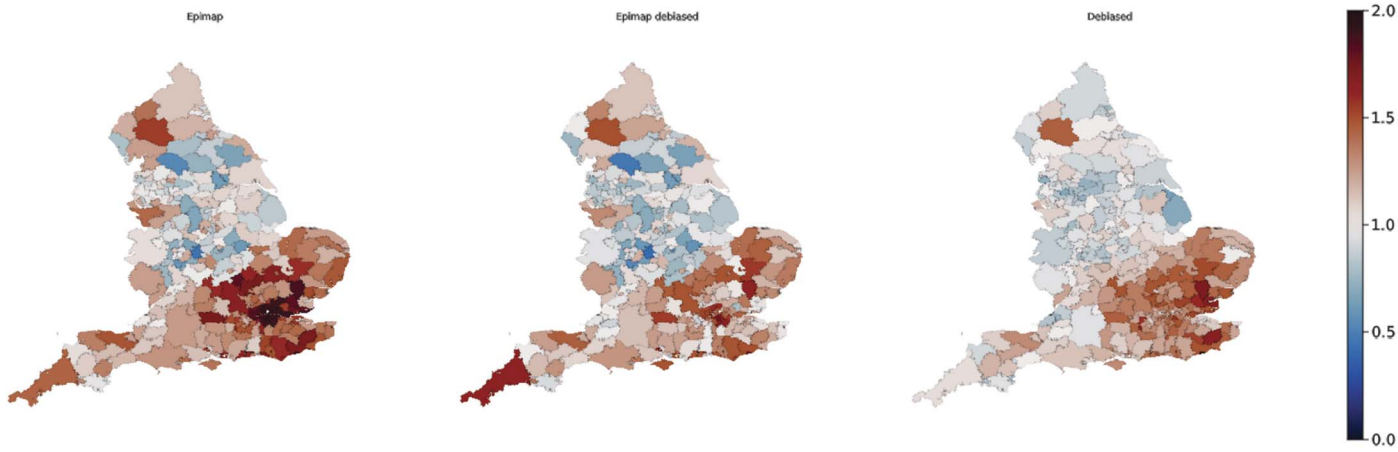

FIG. 8. *Maps of estimated* $\mathcal{R}^{\text{eff}}_{i,w}$ *for three different models across all* LTLAs *in England* W/C 4*th December* 2020. (*a*) *Epimap* ($\mathcal{M}_{\text{E}}$). (*b*) *Epimap debiased interoperable model* ($\mathcal{M}_{\text{ED}}$). (*c*) *Debiased SIR model* ($\mathcal{M}_{\text{D}}$).

in (25) (details omitted for simplicity), thereby creating a seamless interface between models. In summary, the mapping from daily incidence $X_{i,t}$ to weekly prevalence proportion $\pi_{i,w}$ is

$$\pi_{i,w} \equiv \frac{1}{7}\sum_{t\in w}\pi_{i,t}, \tag{24}$$

$$\pi_{i,t} \equiv \frac{1}{M_i}\sum_{s=0}^{t} X_{i,t-s}\, D_s^{\text{PCR+}}, \tag{25}$$

with M_i denoting local population size. Finally, the latent weekly prevalence proportions $\pi_{i,w}$ in the interoperable Epimap model are related to the debiasing model outputs $(\hat{\pi}_{i,w}, \hat{\sigma}^2_{i,w})$ via the approximate π-marginal likelihood of (11) (see also Figure 3(b)):

$$\begin{aligned}&\text{logit}(\hat{\pi}_{i,w}) \mid X_{i,1:t[w]}\\ &\quad\sim \text{Normal}\big(\text{logit}(\hat{\pi}_{i,w}) \mid \text{logit}(\pi_{i,w}), \hat{\sigma}^2_{i,w}\big),\end{aligned} \tag{26}$$

where $t[w]$ denotes the final day in week w.

In summary, the interface between Epimap and debiasing models is created by removing Epimap's observation model of (19) and (20), and replacing it with the debiasing-model outputted marginal likelihood defined at (24), (25) and (26).

4.5.3 *Comparing and contrasting estimated* $\mathcal{R}^{\text{eff}}_{i,w}$ *across models*. We estimated $\mathcal{R}^{\text{eff}}_{i,t}$ under each of the three models: Epimap ($\mathcal{M}_{\text{E}}$) described in Section 4.5.1; the debiased SIR model ($\mathcal{M}_{\text{D}}$) output as described in Section 4.3; and the interoperable combination of Epimap and debiased models ($\mathcal{M}_{\text{ED}}$) described in Section 4.5.2. Movie 1 in the Supplementary Material [50] provides a global perspective on the results, showing the longitudinal changes via maps evolving through time. Figure 8 shows one snapshot of this movie, for all LTLAs in England W/C 4th December 2020. The general theme is one of consistency in $\mathcal{R}^{\text{eff}}_{i,w}$ estimates across models, but there are points in space and time at which results differ. We will discuss these differences, demonstrating that they can help us to characterise and understand model performance.

4.5.4 *Interpreting model outputs via data synchronisation*. We turn first to the maps in Figure 8, visually comparing and contrasting the three models across all LTLAs in W/C 4th December 2020. One interesting feature here is the presence of a few LTLAs with low $\mathcal{R}^{\text{eff}}_{i,w}$ in both $\mathcal{M}_{\text{E}}$ and $\mathcal{M}_{\text{ED}}$, but with relatively high $\mathcal{R}^{\text{eff}}_{i,w}$ in $\mathcal{M}_{\text{D}}$. The two most prominent, coloured blue in Figure 8(a)–(b), are (North Warwickshire, Craven), which have $\mathcal{R}^{\text{eff}}_{i,w}$ estimated to be (0.37, 0.47) by $\mathcal{M}_{\text{E}}$ and (0.33, 0.42) by $\mathcal{M}_{\text{ED}}$, but estimated to be higher at (0.80, 0.90) by $\mathcal{M}_{\text{D}}$. When we compare two models that differ in both data inputs and in probabilistic structure (as do $\mathcal{M}_{\text{E}}$ and $\mathcal{M}_{\text{D}}$) any difference in results cannot immediately be attributed solely to either data or model structure. However, by constraining the two models to have the same data inputs—as we have here by using the prevalence outputs of $\mathcal{M}_{\text{D}}$ as inputs to $\mathcal{M}_{\text{ED}}$—we can potentially learn more. Observing that $\mathcal{M}_{\text{ED}}$ agrees with its "model twin" $\mathcal{M}_{\text{E}}$, but disagrees with its "data twin" $\mathcal{M}_{\text{D}}$, leads us to conclude that differing results in (North Warwickshire, Craven) between $\mathcal{M}_{\text{D}}$ and $\mathcal{M}_{\text{E}}$ arise because of differences in model structure rather than because of differences in data inputs.

We examine the hypothesis that the differences in (North Warwickshire, Craven) $\mathcal{R}^{\text{eff}}_{i,w}$ between $\mathcal{M}_{\text{E}}$ and $\mathcal{M}_{\text{D}}$ are attributable to Epimap's cross-coupled infection load, $\tilde{Z}_{i,t}$ in (21), which allows transmission across regional boundaries; in contrast, the debiased SIR model has only within-LTLA transmission. Note from (22) that Epimap's expected number of new infections is represented as the product $\mathbb{E}[X_{i,t} \mid \mathcal{R}^{\text{eff}}_{i,t}, \tilde{Z}_{i,t}] = \mathcal{R}^{\text{eff}}_{i,t}\tilde{Z}_{i,t}$, so that low estimates of $\mathcal{R}^{\text{eff}}_{i,t}$ will arise when the cross-coupled infection load $\tilde{Z}_{i,t}$ is large relative to the latent

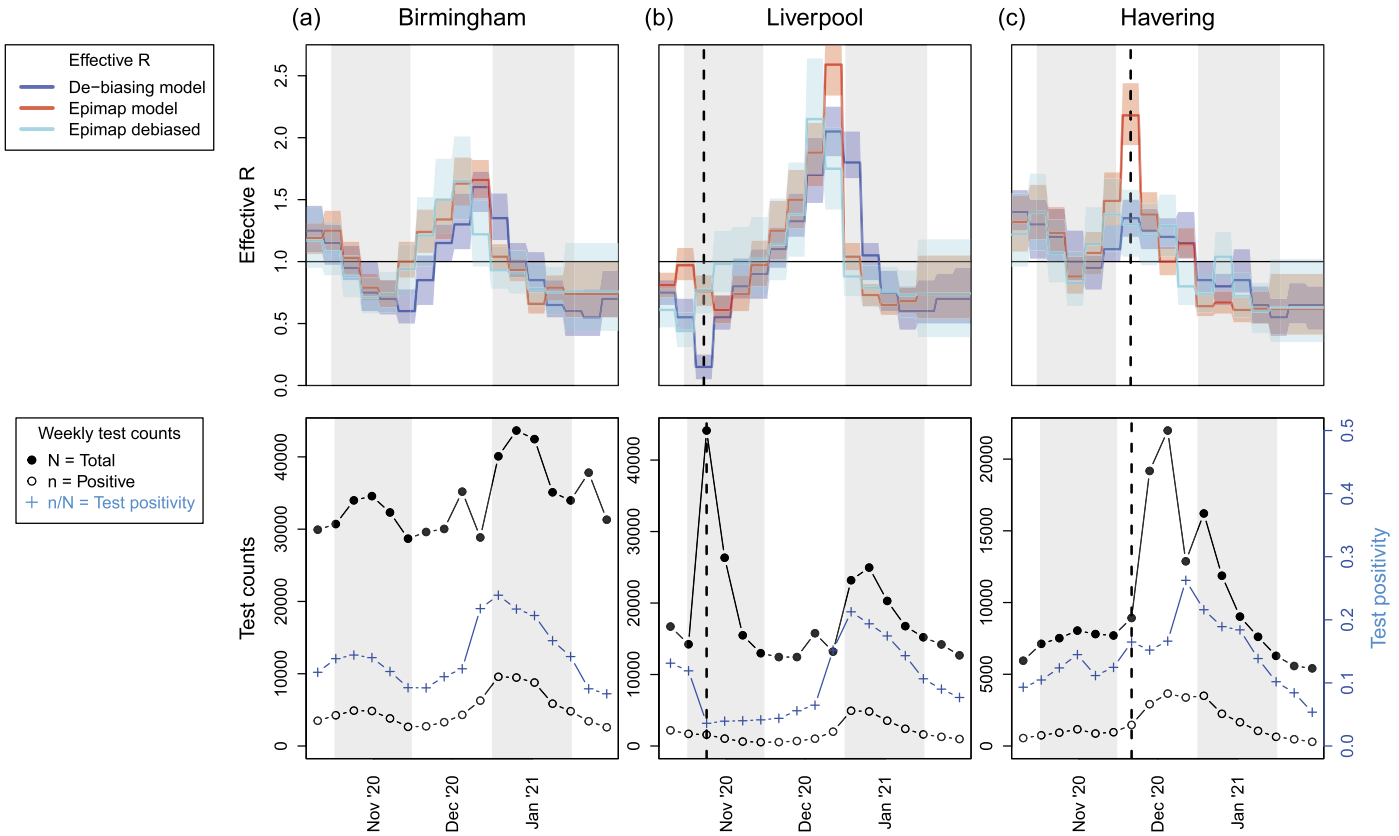

FIG. 9. *Longitudinal $\mathcal{R}^{\text{eff}}_{i,w}$ trajectories superimposed above corresponding Pillar $1+2$ test data. Top: posterior median and 95% CIs of $\mathcal{R}^{\text{eff}}_{i,w}$ for three models (see upper legend, where the "De-biasing", "Epimap" and "Epimap debiased" models are referred to in the text as $\mathcal{M}_{\text{D}}$, $\mathcal{M}_{\text{E}}$ and $\mathcal{M}_{\text{ED}}$ respectively). Bottom: weekly test counts, positive $n_{i,w}$ and total $N_{i,w}$, as well as test positivity $n_{i,w}/N_{i,w}$ (see lower legend). Vertical dashed lines in panels (b-c) indicate instances of model discordance in $\mathcal{R}^{\text{eff}}_{i,w}$ estimates, preceding or coinciding with local surges in testing capacity.*

incidence $X_{i,t}$. For 4th December 2020 in (North Warwickshire, Craven) $\mathcal{M}_{\text{ED}}$ outputs a posterior median for $X_{i,t}$ of (42.7, 24.8) and for $\tilde{Z}_{i,t}$ of (114.5, 81.1), consistent with the low $\mathcal{R}^{\text{eff}}_{i,t}$ estimates of (0.33, 0.42). We can further decompose the cross-coupled infection load $\tilde{Z}_{i,t}$ into infection load $Z_{i,t}$ originating from within LTLA i (27.5, 31.1), and $Z_{-i,t}$ originating from other LTLAs (86.7, 49.9). See Figure 10 in Appendix C for a map of the proportion of infection load arising external to each LTLA. It is clear that for (North Warwickshire, Craven) the majority of the infection load in the $\mathcal{M}_{\text{E}}$ (and $\mathcal{M}_{\text{ED}}$) models is *external*, and that these are among only a handful of LTLAs having external load at $> 50\%$ of the infection burden. Through data synchronisation, theorizing on salient differences between models, and examining confirmatory diagnostic plots, we have increased our understanding of the operational differences between the $\mathcal{M}_{\text{E}}$ and $\mathcal{M}_{\text{D}}$ models.

4.5.5 *Illustrating synergy between models.* The top panels in Figure 9 present longitudinal $\mathcal{R}^{\text{eff}}_{i,w}$ curves for three selected LTLAs for each of the three models; the bottom panels display the corresponding Pillar $1+2$ weekly test counts and positivity rate. First note that the $\mathcal{R}^{\text{eff}}_{i,w}$ plot for Birmingham in Figure 9(a) exhibits reassuring similarity between models; indeed this is what we observe for the majority of LTLAs that are not shown in Figure 9. But the LTLAs in Figure 9(b)–(c) display some interesting and contrasting behaviour between models.

The vertical dashed line in each panel 9(b)–(c) coincides with, or immediately precedes, a surge in community testing capacity—see the sharp increase in total test counts in the bottom panels at or after each dashed line. Figure 9(b) includes a mass testing pilot study in Liverpool beginning 6th November 2020 which led test positivity to drop sharply from 11.9% in W/C 29th October 2020 to 3.6% the following week. Such an abrupt and localized change in testing ascertainment is at odds with the spatiotemporally smooth parameterization of the debiasing model, leading to artefactual deflation of $\mathcal{R}^{\text{eff}}_{i,w}$ in $\mathcal{M}_{\text{D}}$ in W/C 5th November (marked with a dashed line). The $\mathcal{R}^{\text{eff}}_{i,w}$ deflation is not however evident in $\mathcal{M}_{\text{E}}$ or $\mathcal{M}_{\text{ED}}$.

Turning to Figure 9(c) we note that, in mid-December 2020, parts of Essex and London, including the illustrated example of Havering, were moved into Tier 3—the very high alert level—and earmarked for extra community testing. This led to a large spike in testing capacity and uptake, but test positivity rates remained relatively constant

in the two weeks following the dashed line (in contrast to the mass testing pilot in Liverpool a month earlier: see bottom panels of Figure 9(b)–(c)). This caused artefactual inflation of $\mathcal{R}_{i,t}^{\text{eff}}$ in $\mathcal{M}_{\text{E}}$, since positive cases surged in line with testing capacity, and $\mathcal{M}_{\text{E}}$ takes as input only the positive cases $n_{i,t}$ (but not the total tests $N_{i,t}$). The other two models, $\mathcal{M}_{\text{ED}}$ and $\mathcal{M}_{\text{D}}$, are not obviously affected by the December 2020 testing surge in Figure 9(c).

The steady performance of the interoperable model $\mathcal{M}_{\text{ED}}$ across Figure 9 points to a desirable synergistic form of structural robustness—we aspire to synthesise models in such a way that they support one another, with the strengths of one model stabilising inference if and when the other model shows any weakness with respect to the data generating mechanism. The suboptimal $\mathcal{R}_{i,w}^{\text{eff}}$ estimation observed in 9(b)–(c) occurred when there were sudden changes in the ascertainment mechanism; it is reassuring that both adversely affected models ($\mathcal{M}_{\text{D}}$ in Figure 9(b) and $\mathcal{M}_{\text{E}}$ in 9(c)) apparently return to agreement with the other models just one week after these extreme shifts in ascertainment bias.

5. DISCUSSION

Based on our experience, we believe that striving for interoperability across all facets of the delivery of statistical projects will provide:

- *Agility*: the ability to rapidly interlink and recycle statistical modelling outputs across analyses, with components transferable across health security problems;
- *Robustness*: the structural assembly of modules that can be tested independently and connected in such a way as to mitigate any widespread impact of model misspecification;
- *Sustainability*: a shareable, high-quality, reusable, open-source analytic code base of modules that grows over time;
- *Transferability*: a way to facilitate co-ownership of projects with public health and health policy teams, allowing rapid impact from academia and industry to be delivered against relevant, time-sensitive problems;
- *Preparedness*: solutions built for a specific health emergency, such as the COVID-19 pandemic, can be repurposed to meet future public health challenges. In particular, the necessary generic structural links between the data engineering architecture and the analytic and modelling side will have already been built.

Many challenges lie ahead on the path towards an effective, interoperable, and comprehensive disease surveillance system. From a statistical point of view, it is most relevant to focus our attention on issues that are generic and likely to recur when addressing a range of questions. For brevity we will only mention three particularly challenging ones.

A major hurdle that interoperability will face is the need to integrate evidence from data collected at different time steps and spatial scales. For example, the time granularity of the randomised survey data used in our debiased prevalence model is a week, yet most epidemic models have been built on the basis of daily case numbers, thus necessitating an additional time-alignment interface. Similarly, it will be common to have to integrate different geographies into a single model, constrained by the data sources. Misaligned geographies is a recurrent statistical issue that has been much discussed in environmental sciences [48], and for which pragmatic but robust solutions need to be investigated.

A second challenge is situations when moment-matched Gaussian distributions, as used here, do not provide adequate approximations; alternatives include particle-based approaches, in which posterior samples from a module are used as a proposal within an MCMC scheme [24, 38] or for importance sampling [45] or within a sequential Monte Carlo scheme [34].

A final challenging issue that we have already encountered, and dealt with in Section 4.2.2 by using a cut posterior, is how best to balance or weight different sources of evidence, to take into account prior knowledge; see [16] for a discussion of evidence weighting from an epidemiological perspective. Rather than completely preventing feedback, as per a cut posterior, it may be desirable to only partially down-weight, as proposed by [8]. For example, in the case of diagnostic tests, weights might take into consideration their *modus operandi* and context of use.

Although the analyses presented here were motivated by the specific example of the COVID-19 pandemic, the overarching principle of interoperability is pertinent in a variety of contexts with complex modelling requirements in a dynamic environment, such as climate change or natural disaster management.

APPENDIX A: LONGITUDINAL SMOOTHING PRIOR FOR BIAS PARAMETER δ

We evaluate the cross-sectional cut posterior for $\delta_{J,w}$, the bias at week w in PHE region J (Figure 1(a)), and use these to construct a prior to take forward to full Bayesian inference at each LTLA in region J. To induce smoothness we construct a "product-of-experts" prior [27]:

$$\begin{aligned} p(\boldsymbol{\delta}_{J,1:W}) \propto \text{Normal}(\boldsymbol{\delta}_{J,1:W} \mid \mathbf{0}, \boldsymbol{\Sigma}_{\delta}) \\ \times \prod_{w \in \mathcal{W}} \text{N}(\delta_{J,w} \mid \hat{m}_{J,w}, \hat{s}^2_{J,w}) \\ \times \prod_{w \notin \mathcal{W}} \text{N}(\delta_{J,w} \mid 0, v_{\text{flat}}). \end{aligned} \tag{27}$$

The first term on the right of (27) is a subjective prior on the longitudinal smoothness of $\delta_{1:T}$ encoding an AR1

process in $\boldsymbol{\Sigma}_\delta$, defined to be only very weakly informative with respect to average location of $\boldsymbol{\delta}_{J,1:W}$; the second term is the product of (approximations to) the cross-sectional cut-posterior marginals from (6) at weeks $\mathcal{W}$ for which REACT data are available, that is,

$$\begin{aligned}(28)\quad &\prod_{w\in\mathcal{W}} p^{\text{cut}}(\delta_{J,w} \mid u_{J,w} \text{ of } U_{J,w}, n_{J,w} \text{ of } N_{J,w})\\ &\approx \prod_{w\in\mathcal{W}} \mathrm{N}(\delta_{J,w} \mid \hat{m}_{J,w}, \hat{s}^2_{J,w});\end{aligned}$$

and the third term is a product of noninformative cross-sectional priors when REACT data are unavailable.

The normalised form of (27) is MV Gaussian:

$$\begin{aligned}(29)\quad p(\boldsymbol{\delta}_{J,1:W}) &= \text{Normal}(\boldsymbol{\delta}_{J,1:W} | \hat{\boldsymbol{\mu}}^{\text{out}}_\delta, \hat{\boldsymbol{\Sigma}}^{\text{out}}_\delta),\\ \hat{\boldsymbol{\mu}}^{\text{out}}_\delta &:= \left(\boldsymbol{\Sigma}_\delta^{-1} + \boldsymbol{D}^{-1}\right)^{-1} \boldsymbol{D}^{-1} \hat{\boldsymbol{\mu}},\\ \hat{\boldsymbol{\Sigma}}^{\text{out}}_\delta &:= \left(\boldsymbol{\Sigma}_\delta^{-1} + \boldsymbol{D}^{-1}\right)^{-1}\end{aligned}$$

with ($\hat{\boldsymbol{\mu}}$, diagonal matrix $\boldsymbol{D}_{W\times W}$) having elements ($\hat{m}_{J,w}$, $\hat{s}^2_{J,w}$) for $w \in \mathcal{W}$ and $(0, v_{\text{flat}})$ for $w \notin \mathcal{W}$.

We denote the marginal distribution of (29) for week w by

$$(30)\quad p_{\text{DD}}(\delta_{J,w}) := \text{Normal}(\delta_{J,w} \mid \hat{\mu}_{J,w}, \hat{\tau}^2_{J,w}),$$

deploying these $p_{\text{DD}}(\delta_{J[i],w})$ as data-dependent priors independently at each week w in LTLA i in region $J[i]$ as described in Section 4.2.4.

APPENDIX B: FULL MODEL SPECIFICATION FOR THE SPACE–TIME EQUALITY ANALYSIS

Following [59, 64], we model $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_I)$, the random effect accounting for the spatial autocorrelation across LTLAs, as

$$(31)\quad \boldsymbol{\lambda} = \frac{1}{\sqrt{\tau}}(\sqrt{1-\rho}\boldsymbol{v} + \sqrt{\rho}\boldsymbol{u}).$$

The vector $\boldsymbol{u} = (u_1, \ldots, u_I)$ is a spatially structured random effect with prior distribution

$$(32)\quad \boldsymbol{u}|\tau, \rho \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{Q}^-),$$

where $\boldsymbol{Q}^-$, is the inverse of the precision matrix of a ICAR model, scaled in the sense of [65]. The vector $\boldsymbol{v} = (v_1, \ldots, v_I)$ is an i.i.d. Gaussian random effect, that is

$$(33)\quad \boldsymbol{v}|\tau, \rho \sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}),$$

where $\boldsymbol{I}$ is the identity matrix. To account for the time dependence, ϵ_w is modelled as a random walk of order 2. Given $\Delta^2\epsilon_w = \epsilon_w - 2\epsilon_{w+1} + \epsilon_{w+2}$, this can be formalized as

$$(34)\quad \Delta^2\epsilon_w|\sigma^2_\epsilon \sim \text{Normal}(0, \sigma^2_\epsilon).$$

Both $\boldsymbol{\lambda}$ and $\boldsymbol{\epsilon}$ imply a degree of smoothing in space and in time and help highlight persistent patterns in the data. Finally we set a noninformative Gamma$(1, 5 \times 10^{-5})$ prior on the inverse of σ^2_ϵ and a noninformative Normal$(0, 1000)$ prior for the fixed effect coefficients β_1 and β_2. All models in the case study presented in Section 4.4 are fitted using the R package `INLA` [62].

APPENDIX C: ILLUSTRATION OF EPIMAP EXTERNAL INFECTION LOAD

Figure 10 presents the proportion of infection load originating external to each LTLA on 4th December 2020. In Section 4.5.4, we discuss the relatively low estimates of $\mathcal{R}^{\text{eff}}_{i,w}$ seen for (North Warwickshire, Craven), and we attribute this to them experiencing a relatively large external infection load, (North Warwickshire, Craven) are among the approximately 10 LTLAs in this map with external infection load proportionally $> 50\%$.

ACKNOWLEDGMENTS

Chris Holmes and Sylvia Richardson contributed equally to this research.

FUNDING

MB acknowledges partial support from the MRC Centre for Environment and Health, which is currently funded by the Medical Research Council (MR/S019669/1). RJBG was funded by the UKRI Medical Research Council (MRC) [programme code MC_UU_00002/2] and supported by the NIHR Cambridge Biomedical Research Centre [BRC-1215-20014]. BCLL was supported by the UK Engineering and Physical Sciences Research Council through the Bayes4Health programme [Grant number EP/R018561/1] and gratefully acknowledges funding from Jesus College, Oxford. GN and CH acknowledge support from the Medical Research Council Programme Leaders award MC_UP_A390_1107. CH acknowledges support from The Alan Turing Institute, Health Data Research, U.K., and the U.K. Engineering and Physical Sciences Research Council through the Bayes4Health programme grant. SR is supported by MRC programme grant MC_UU_00002/10; The Alan Turing Institute grant: TU/B/000092; EPSRC Bayes4Health programme grant: EP/R018561/1. HG and TF acknowledge partial support from Huawei Research UK. Infrastructure support for the Department of Epidemiology and Biostatistics is also provided by the NIHR Imperial BRC. Authors in The Alan Turing Institute and Royal Statistical Society Statistical Modelling and Machine Learning Laboratory gratefully acknowledge funding from Data, Analytics and Surveillance Group, a part of the UKHSA. This work was funded by The Department for Health and Social Care (Grant

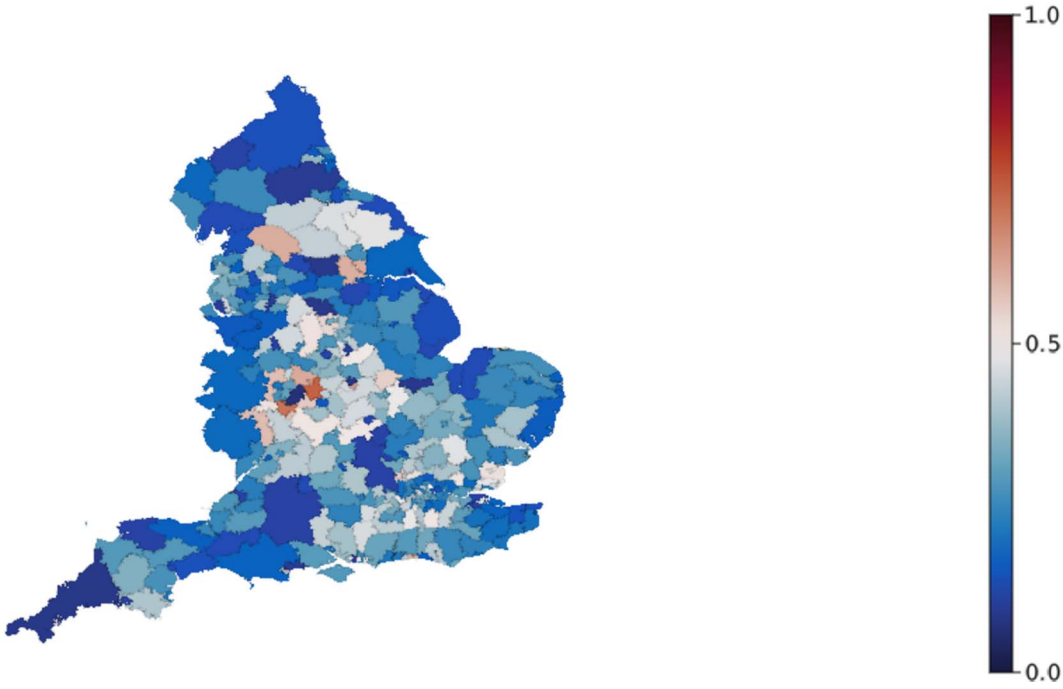

FIG. 10. *Map of the proportion of infection load originating external to each LTLA on* 4*th December* 2020. *In Section* 4.5.4 *we discuss the relatively low estimates of* $\mathcal{R}^{\text{eff}}_{i,w}$ *seen for* (*North Warwickshire*, *Craven*), *and we attribute this to them experiencing a relatively large external infection load*, (*North Warwickshire*, *Craven*) *are among the approximately* 10 LTLAs *in this map with external infection load proportionally* $> 50\%$.

ref: 2020/045) with support from The Alan Turing Institute (EP/W037211/1) and in-kind support from The Royal Statistical Society. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the National Health Service, NIHR, Department of Health, Joint Biosecurity Centre, or PHE.

SUPPLEMENTARY MATERIAL

Movie of $\mathcal{R}^{\text{eff}}$ for Three Different Models. (DOI: 10.1214/22-STS854SUPP; .zip). Movie showing maps of effective reproduction number $\mathcal{R}^{\text{eff}}_{i,w}$ for three different models across local authorities in England between W/C 31st October 2020 and W/C 2nd January 2021. *Left*: Epimap model described in Section 4.5.1. *Middle*: Epimap-Debiased interoperable model described in Section 4.5.2. *Right*: SIR model based on debiased prevalence outputs described in Section 4.3.

REFERENCES

[1] ADES, A. E. and SUTTON, A. J. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: Current approaches. *J. Roy. Statist. Soc. Ser. A* **169** 5–35. MR2222010 https://doi.org/10.1111/j.1467-985X.2005.00377.x

[2] ANDERSON, R., DONNELLY, C., HOLLINGSWORTH, D., KEELING, M., VEGVARI, C., BAGGALEY, R. and MADDREN, R. (2020). Reproduction number (R) and growth rate (r) of the COVID-19 epidemic in the UK: Methods of estimation, data sources, causes of heterogeneity, and use as a guide in policy formulation Technical Report London, UK: Royal Society.

[3] BI, Q. et al. (2020). Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: A retrospective cohort study. *Lancet Infect. Dis.* **20** 911–919.

[4] BIRRELL, P., BLAKE, J., VAN LEEUWEN, E., GENT, N. and DE ANGELIS, D. (2021). Real-time nowcasting and forecasting of COVID-19 dynamics in England: The first wave. *Philos. Trans. - R. Soc., Biol. Sci.* **376** 20200279. https://doi.org/10.1098/rstb.2020.0279

[5] BRACHER, J., RAY, E. L., GNEITING, T. and REICH, N. G. (2021). Evaluating epidemic forecasts in an interval format. *PLoS Comput. Biol.* **17** e1008618. https://doi.org/10.1371/journal.pcbi.1008618

[6] BRAUER, F., VAN DEN DRIESSCHE, P. and WU, J., eds. (2008) In *Mathematical Epidemiology*. *Lecture Notes in Math.* **1945**. Springer, Berlin. MR2452129 https://doi.org/10.1007/978-3-540-78911-6

[7] BRAUNER, J. M., MINDERMANN, S., SHARMA, M., JOHNSTON, D., SALVATIER, J., GAVENČIAK, T., STEPHENSON, A. B., LEECH, G., ALTMAN, G. et al. (2021). Inferring the effectiveness of government interventions against COVID-19. *Science* **371**. https://doi.org/10.1126/science.abd9338

[8] CARMONA, C. U. and NICHOLLS, G. K. (2020). Semi-modular inference: Enhanced learning in multi-modular models by tempering the influence of components. Available at arXiv:2003.06804.

[9] CHEN, P. M., LEE, E. K., GIBSON, G. A., KATZ, R. H. and PATTERSON, D. A. (1994). RAID: High-performance, reliable secondary storage. *ACM Comput. Surv.* **26** 145–185. https://doi.org/10.1145/176979.176981

[10] COVID-19 INFECTION SURVEY—OFFICE FOR NATIONAL STATISTICS. Available at https://www.ons.gov.uk/surveys/informationforhouseholdsandindividuals/householdandindividualsurveys/covid19infectionsurvey.

[11] COVID-19 TASK FORCE. Available at https://rss.org.uk/policy-campaigns/policy-groups/covid-19-task-force/.

[12] DANIELS, M. J. and KASS, R. E. (1998). A note on first-stage approximation in two-stage hierarchical models. *Sankhya, Ser. B* **60** 19–30. MR1717073

[13] DAVISON, A. C. (2003). *Statistical Models. Cambridge Series in Statistical and Probabilistic Mathematics* **11**. Cambridge Univ. Press, Cambridge. MR1998913 https://doi.org/10.1017/CBO9780511815850

[14] DAWID, A. P. (1985). Probability, symmetry and frequency. *British J. Philos. Sci.* **36** 107–128. MR0915922 https://doi.org/10.1093/bjps/36.2.107

[15] DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267 https://doi.org/10.1214/aos/1176349260

[16] DE ANGELIS, D., PRESANIS, A. M., BIRRELL, P. J., TOMBA, G. S. and HOUSE, T. (2015). Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics* **10** 83–87. https://doi.org/10.1016/j.epidem.2014.09.004

[17] DEPARTMENT OF HEALTH AND SOCIAL CARE (UK). COVID-19 testing data: Methodology note. Available at https://www.gov.uk/government/publications/coronavirus-covid-19-testing-data-methodology/covid-19-testing-data-methodology-note.

[18] DEPARTMENT OF HEALTH AND SOCIAL CARE GUIDANCE. REPRODUCTION NUMBER (R) AND GROWTH RATE: METHODOLOGY. Available at https://www.gov.uk/government/publications/reproduction-number-r-and-growth-rate-methodology.

[19] DOMINICI, F., SAMET, J. M. and ZEGER, S. L. (2000). Combining evidence on air pollution and daily mortality from the 20 largest US cities: A hierarchical modelling strategy. *J. Roy. Statist. Soc. Ser. A* **163** 263–302.

[20] FLAXMAN, S., MISHRA, S., GANDY, A., UNWIN, H. J. T., MELLAN, T. A., COUPLAND, H., WHITTAKER, C., ZHU, H., BERAH, T. et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584** 257–261.

[21] GE, H., XU, K. and GHAHRAMANI, Z. (2018). Turing: A language for flexible probabilistic inference. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (A. Storkey and F. Perez-Cruz, eds.). *Proceedings of Machine Learning Research* **84** 1682–1690. PMLR, Playa Blanca, Lanzarote, Canary Islands.

[22] GIT. Available at https://git-scm.com/.

[23] GOUDIE, R. J. B., HOVORKA, R., MURPHY, H. R. and LUNN, D. (2015). Rapid model exploration for complex hierarchical data: Application to pharmacokinetics of insulin aspart. *Stat. Med.* **34** 3144–3158. MR3402581 https://doi.org/10.1002/sim.6536

[24] GOUDIE, R. J. B., PRESANIS, A. M., LUNN, D., DE ANGELIS, D. and WERNISCH, L. (2019). Joining and splitting models with Markov melding. *Bayesian Anal.* **14** 81–109. MR3910039 https://doi.org/10.1214/18-BA1104

[25] GREEN, P. J., HJORT, N. L. and RICHARDSON, S. (2003). *Highly Structured Stochastic Systems. Oxford Statistical Science Series*. Oxford University Press, Oxford, New York.

[26] HELLEWELL, J., RUSSELL, T. W., THE SAFER INVESTIGATORS AND FIELD STUDY TEAM, THE CRICK COVID-19 CONSORTIUM, CMMID COVID-19 WORKING GROUP, BEALE, R., KELLY, G., HOULIHAN, C., NASTOULI, E. et al. (2020). Estimating the effectiveness of routine asymptomatic PCR testing at different frequencies for the detection of SARS-CoV-2 infections. *MedRxiv* 2020.11.24.20229948. https://doi.org/10.1101/2020.11.24.20229948

[27] HINTON, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14** 1771–1800. https://doi.org/10.1162/089976602760128018

[28] HOOTEN, M. B., JOHNSON, D. S. and BROST, B. M. (2021). Making recursive Bayesian inference accessible. *Amer. Statist.* **75** 185–194. MR4256123 https://doi.org/10.1080/00031305.2019.1665584

[29] JACOB, P. E., MURRAY, L. M., HOLMES, C. C. and ROBERT, C. P. (2017). Better together? Statistical learning in models made of modules. Available at arXiv:1708.08719.

[30] JACOB, P. E., O'LEARY, J. and ATCHADÉ, Y. F. (2020). Unbiased Markov chain Monte Carlo methods with couplings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 543–600. https://doi.org/10.1111/rssb.12336

[31] JOHNSON, D. S., BROST, B. M. and HOOTEN, M. B. (2020). Greater than the sum of its parts: Computationally flexible Bayesian hierarchical modeling. Available at arXiv:2010.12568.

[32] THE JULIA PROGRAMMING LANGUAGE. Available at https://julialang.org/.

[33] LEE, D. and SARRAN, C. (2015). Controlling for unmeasured confounding and spatial misalignment in long-term air pollution and health studies. *Environmetrics* **26** 477–487. MR3415567 https://doi.org/10.1002/env.2348

[34] LINDSTEN, F., JOHANSEN, A. M., NAESSETH, C. A., KIRKPATRICK, B., SCHÖN, T. B., ASTON, J. A. D. and BOUCHARD-CÔTÉ, A. (2017). Divide-and-conquer with sequential Monte Carlo. *J. Comput. Graph. Statist.* **26** 445–458. MR3640200 https://doi.org/10.1080/10618600.2016.1237363

[35] LIU, F., BAYARRI, M. J. and BERGER, J. O. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Anal.* **4** 119–150. MR2486241 https://doi.org/10.1214/09-BA404

[36] LIU, Y. and GOUDIE, R. J. B. (2021). Generalized geographically weighted regression model within a modularized Bayesian framework. Available at arXiv:2106.00996.

[37] LIU, Y. and GOUDIE, R. J. B. (2022). Stochastic approximation cut algorithm for inference in modularized Bayesian models. *Stat. Comput.* **32** Paper No. 7. MR4350200 https://doi.org/10.1007/s11222-021-10070-2

[38] LUNN, D., BARRETT, J., SWEETING, M. and THOMPSON, S. (2013). Fully Bayesian hierarchical modelling in two stages, with application to meta-analysis. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **62** 551–572. MR3083911 https://doi.org/10.1111/rssc.12007

[39] LUNN, D., BEST, N., SPIEGELHALTER, D., GRAHAM, G. and NEUENSCHWANDER, B. (2009). Combining MCMC with 'sequential' PKPD modelling. *J. Pharmacokinet. Pharmacodyn.* **36** 19–38. https://doi.org/10.1007/s10928-008-9109-1

[40] MAISHMAN, T., SCHAAP, S., SILK, D. S., NEVITT, S. J., WOODS, D. C. and BOWMAN, V. E. (2021). Statistical methods used to combine the effective reproduction number, $R(t)$, and other related measures of COVID-19 in the UK. Available at arXiv:2103.01742.

[41] MANDERSON, A. A. and GOUDIE, R. J. B. (2021). Combining chains of Bayesian models with Markov melding. Available at arXiv:2111.11566.

[42] MASSA, M. S. and LAURITZEN, S. L. (2010). Combining statistical models. In *Algebraic Methods in Statistics and Probability II* (M. A. G. Viana and H. P. Wynn, eds.). *Contemp. Math.* **516** 239–259. Amer. Math. Soc., Providence, RI. MR2730753 https://doi.org/10.1090/conm/516/10179

[43] MATHUR, R., RENTSCH, C., MORTON, C., HULME, W., SCHULTZE, A., MACKENNA, B., EGGO, R., BHASKARAN, K., WONG, A. et al. (2021). Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: An observational cohort study using the OpenSAFELY platform. *Lancet* **397** 1711–1724.

[44] MAUCORT-BOULCH, D., FRANCESCHI, S., PLUMMER, M. and IARC HPV PREVALENCE SURVEYS STUDY GROUP (2008). International correlation between human papillomavirus prevalence and cervical cancer incidence. *Cancer Epidemiol. Biomark. Prev.* **17** 717–720. https://doi.org/10.1158/1055-9965.EPI-07-2691

[45] MAUFF, K., STEYERBERG, E., KARDYS, I., BOERSMA, E. and RIZOPOULOS, D. (2020). Joint models with multiple longitudinal outcomes and a time-to-event outcome: A corrected two-stage approach. *Stat. Comput.* **30** 999–1014. MR4108688 https://doi.org/10.1007/s11222-020-09927-9

[46] MINISTRY OF HOUSING, COMMUNITIES & LOCAL GOVERNMENT. English indices of deprivation 2019. Available at https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019.

[47] MORALES, D. R. and ALI, S. N. (2021). COVID-19 and disparities affecting ethnic minorities. *Lancet* **397** 1684–1685. https://doi.org/10.1016/S0140-6736(21)00949-1

[48] MUGGLIN, A. S., CARLIN, B. P. and GELFAND, A. E. (2000). Fully model-based approaches for spatially misaligned data. *J. Amer. Statist. Assoc.* **95** 877–887. https://doi.org/10.1080/01621459.2000.10474279

[49] NICHOLLS, G. K., LEE, J. E., WU, C.-H. and CARMONA, C. U. (2022). Valid belief updates for prequentially additive loss functions arising in semi-modular inference. Available at arXiv:2201.09706.

[50] NICHOLSON, G., BLANGIARDO, M., BRIERS, M., DIGGLE, P. J., FJELDE, T. E., GE, H., GOUDIE, R. J. B., JERSAKOVA, R., KING, R. E. et al. (2022). Supplement to "Interoperability of statistical models in pandemic preparedness: Principles and reality." https://doi.org/10.1214/22-STS854SUPP

[51] NICHOLSON, G., LEHMANN, B., PADELLINI, T., POUWELS, K. B., JERSAKOVA, R., LOMAX, J., KING, R. E., MALLON, A.-M., DIGGLE, P. J. et al. (2022). Improving local prevalence estimates of SARS-CoV-2 infections using a causal debiasing framework. *Nat. Microbiol.* **7** 97–107. https://doi.org/10.1038/s41564-021-01029-0

[52] PADELLINI, T., JERSAKOVA, R., DIGGLE, P. J., HOLMES, C., KING, R. E., LEHMANN, B. C. L., MALLON, A.-M., NICHOLSON, G., RICHARDSON, S. et al. (2022). Time varying association between deprivation, ethnicity and SARS-CoV-2 infections in England: A population-based ecological study. *The Lancet Regional Health—Europe* **15** 100322. https://doi.org/10.1016/j.lanepe.2022.100322

[53] PIRANI, M., MASON, A. J., HANSELL, A. L., RICHARDSON, S. and BLANGIARDO, M. (2020). A flexible hierarchical framework for improving inference in area-referenced environmental health studies. *Biom. J.* **62** 1650–1669. MR4184022 https://doi.org/10.1002/bimj.201900241

[54] PLUMMER, M. (2015). Cuts in Bayesian graphical models. *Stat. Comput.* **25** 37–43. MR3304902 https://doi.org/10.1007/s11222-014-9503-z

[55] POMPE, E. and JACOB, P. E. (2021). Asymptotics of cut distributions and robust modular inference using Posterior Bootstrap. Available at arXiv:2110.11149.

[56] POOLE, D. and RAFTERY, A. E. (2000). Inference for deterministic simulation models: The Bayesian melding approach. *J. Amer. Statist. Assoc.* **95** 1244–1255. MR1804247 https://doi.org/10.2307/2669764

[57] POUWELS, K. B., HOUSE, T., PRITCHARD, E., ROBOTHAM, J. V., BIRRELL, P. J., GELMAN, A., VIHTA, K.-D., BOWERS, N., BOREHAM, I. et al. (2021). Community prevalence of SARS-CoV-2 in England from April to November, 2020: Results from the ONS Coronavirus Infection Survey. *The Lancet Public Health* **6** e30–e38.

[58] PRESANIS, A. M., OHLSSEN, D., SPIEGELHALTER, D. J. and DE ANGELIS, D. (2013). Conflict diagnostics in directed acyclic graphs, with applications in Bayesian evidence synthesis. *Statist. Sci.* **28** 376–397. MR3135538 https://doi.org/10.1214/13-STS426

[59] RIEBLER, A., SØRBYE, S. H., SIMPSON, D. and RUE, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Stat. Methods Med. Res.* **25** 1145–1165. MR3541089 https://doi.org/10.1177/0962280216660421

[60] RILEY, S., AINSLIE, K. E., EALES, O., JEFFREY, B., WALTERS, C. E., ATCHISON, C. J., DIGGLE, P. J., ASHBY, D., DONNELLY, C. A. et al. (2020). Community prevalence of SARS-CoV-2 virus in England during May 2020: REACT study. *MedRxiv*.

[61] ROSE, T. C., MASON, K., PENNINGTON, A., MCHALE, P., TAYLOR-ROBINSON, D. C. and BARR, B. (2020). Inequalities in COVID19 mortality related to ethnicity and socioeconomic deprivation. *MedRxiv*.

[62] RUE, H., MARTINO, S. and CHOPIN, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 319–392. MR2649602 https://doi.org/10.1111/j.1467-9868.2008.00700.x

[63] SCOTT, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *J. Amer. Statist. Assoc.* **97** 337–351. MR1963393 https://doi.org/10.1198/016214502753479464

[64] SIMPSON, D., RUE, H., RIEBLER, A., MARTINS, T. G. and SØRBYE, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.* **32** 1–28. MR3634300 https://doi.org/10.1214/16-STS576

[65] SØRBYE, S. H. and RUE, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spat. Stat.* **8** 39–51. MR3326820 https://doi.org/10.1016/j.spasta.2013.06.004

[66] TEH, Y. W., BHOOPCHAND, A., DIGGLE, P., ELESEDY, B., HE, B., HUTCHINSON, M., PAQUET, U., READ, J., TOMASEV, N. et al. (2021). Efficient Bayesian inference of instantaneous re-production numbers at fine spatial scales, with an application to mapping and nowcasting the Covid-19 epidemic in british local authorities. Technical report. To be published in Journal of the Royal Statistical Society, Series A (Statistics in Society). Available at https://localcovid.info/assets/docs/localcovid-writeup.pdf.

[67] UK DATA SERVICE CENSUS. Available at https://ukdataservice.ac.uk/learning-hub/census/.

[68] WELTON, N. J., SUTTON, A. J., COOPER, N. J., ABRAMS, K. R. and ADES, A. E. (2012). *Evidence Synthesis for Decision Making in Healthcare*. Wiley, Chichester.

[69] YU, B. and KUMBIER, K. (2020). Veridical data science. *Proc. Natl. Acad. Sci. USA* **117** 3920–3929. MR4075122 https://doi.org/10.1073/pnas.1901326117

[70] YU, X., NOTT, D. J. and SMITH, M. S. (2021). Variational inference for cutting feedback in misspecified models. Available at arXiv:2108.11066.

[71] ZHANG, L., BEAL, S. L. and SHEINER, L. B. (2003). Simultaneous vs. sequential analysis for population PK/PD data I: Best-case performance. *J. Pharmacokinet. Pharmacodyn.* **30** 387–404. https://doi.org/10.1023/b:jopa.0000012998.04442.1f