A Workflow for Infectious Disease Modelling

Sam Abbott[1], Xiahui Li[†2], Punya Alahakoon[†3], Dhorasso Temfack[†4], Sabine Van Elsland[5], Johannes Bracher[6], Felix Günther[7], Adrian Lison[8], James Hay[3], Oliver Eales[10, 14], Eben Kenah[11], James M. McCaw[10, 14], Mircea T. Sofonea[12], Pierre Nouvellet[7, 13], Freya Shearer[14], Sebastian Funk[1], Daniela De Angelis[15], Michael J. Plank[*16], Anne Cori[*5], and Anne M. Presanis[*15]

[1]London School of Hygiene & Tropical Medicine
[2]School of Mathematics and Statistics, University of St Andrews
[3]Pandemic Sciences Institute, Nuffield Department of Medicine, University of Oxford
[4]School of Computer Science and Statistics, Trinity College Dublin
[5]MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London
[6]Karlsruhe Institute of Technology
[7]Robert Koch Institute
[8]Computational Evolution, ETH Zurich
[9]Nuffield Department of Medicine, University of Oxford
[10]School of Mathematics and Statistics, University of Melbourne
[11]College of Public Health, Ohio State University
[12]University of Montpellier
[13]School of Life Sciences, University of Sussex
[14]School of Population and Global Health, University of Melbourne
[15]MRC Biostatistics Unit, University of Cambridge
[16]School of Mathematics and Statistics, University of Canterbury

March 30, 2026

[*†]Equal contribution
[†*]Equal contribution

Abstract

Infectious disease models can be used to inform critical public health decisions, yet often do not follow systematic development and validation practices. The infectious disease modelling community has been slow to adopt rigorous model development and criticism cycles such as the Bayesian workflow, even as these approaches have become increasingly used in other domains. Recent outbreaks have demonstrated some domain-specific challenges that infectious disease modelling faces, including evolving research questions, new or poorly characterised data sources, and adapting surveillance systems. Here, we suggest the adoption of a workflow for developing and evaluating infectious disease models, which builds on existing generic Bayesian workflows but focuses on domain-specific challenges. This workflow is designed for anyone developing an infectious disease model, and for users of model outputs who need to be able to evaluate modelling studies. At each stage, we provide recommendations based on our experience. We begin by outlining an approach for characterising epidemiological data source properties through a structured checklist. We then present an iterative workflow that applies the Bayesian workflow to the infectious disease domain, with the checklist informing decisions throughout each workflow stage. The workflow proposed here includes defining the research question, development of Directed Acyclic Graph representations of process and observation models in a state-space framework, model modularisation, inference and computation choices, model specification and validation, integration method selection, and real-world considerations. Throughout, we identify feedback loops where later decisions impact earlier choices. We also give guidance on using the workflow in evolving settings, such as outbreaks, and on how to report its use. To demonstrate this workflow, we provide a schematic case study that progressively integrates data sources for estimating transmission intensity. At each stage, we give examples of navigating real-world trade-offs between model complexity, computational feasibility, and inferential goals. This case study highlights how different data types can provide complementary information but may also impact workflow choices. Our suggested framework emphasises parsimony, modularity, interpretability, and model criticism. By proposing domain-specific workflow practices, we aim to provide a foundation for improving the quality and transparency of infectious disease modelling, particularly during outbreaks where flexible, but principled, approaches are essential.

Contents

1 Introduction **4**

2 Data Sources and Characteristics **5**

3 Workflow **6**
- 3.1 Research Question and Target Estimands . . . 9
- 3.2 Process DAG Development . . . 9
- 3.3 Data Source Selection . . . 10
- 3.4 Observation DAG Construction . . . 10
- 3.5 Refining the model DAGs . . . 11
- 3.6 Modularising DAGs . . . 12
- 3.7 Inference and Computation Choices . . . 12
  - 3.7.1 Model Complexity Assessment . . . 12
  - 3.7.2 Inference Method Selection . . . 14
  - 3.7.3 Implementation considerations . . . 17
  - 3.7.4 Diagnostics . . . 17
- 3.8 Model Implementation . . . 18
- 3.9 Model Specification and Validation . . . 18
  - 3.9.1 Choosing a model family . . . 18
  - 3.9.2 Prior Specification and Prior Predictive Checks . . . 19
  - 3.9.3 Model Fitting and Computational Validation . . . 19
  - 3.9.4 Evaluate and use the model . . . 19
  - 3.9.5 Model Comparison . . . 20
- 3.10 Data Integration Choices . . . 21
  - 3.10.1 Full joint modelling . . . 21
  - 3.10.2 Ensembling independent models . . . 22

4 Workflow Management **22**
- 4.1 Outbreak Evolution and Workflow Iteration . . . 22
- 4.2 Reporting . . . 24

5 Case Study **24**
- 5.1 Data Sources and Characteristics . . . 25
- 5.2 Case Study Model 1: Single-Source Baseline (Cases Only) . . . 28
- 5.3 Case Study Model 2: Two-Source Integration (Cases and Deaths) . . . 31
- 5.4 Case Study Model 3: Three-Source Integration (Cases, Deaths and Wastewater) . . . 34
- 5.5 Case Study Model 4: Individual-Level Data (Cases and Transmission Pairs) . . . 36

6 Discussion **37**

7 Acknowledgements **41**

1 Introduction

Infectious disease models can inform critical public health decisions, yet often don't follow systematic development and validation practices (for recent examples, see Ward et al. [2024], Fyles et al. [2024], Abbott et al. [2021], Abbott and Funk [2022]). Responses to the COVID-19 pandemic, the 2022 global mpox clade II outbreak, and the 2014-2016 West Africa Ebola outbreak [e.g. Knock et al., 2021, Rø et al., 2025, Abbott et al., 2021, Abbott and Funk, 2022, Ward et al., 2024, Birrell et al., 2025] demonstrated both the value of infectious disease models during emergencies and the limitations of rapid model development. These outbreak settings share domain-specific challenges that infectious disease modelling faces: new data streams emerge over time, surveillance systems evolve to meet changing needs, and models must be developed under time constraints with limited understanding of novel data sources [McCaw and Plank, 2023]. These challenges can be addressed through an iterative process of model development, systematic criticism, and principled uncertainty quantification. However, the infectious disease modelling community has not widely adopted rigorous model development and criticism practices [Box, 1979], such as those underlying existing Bayesian workflow [Green et al., 2003, Gelman et al., 2020, Nicholson et al., 2022]. This lack of adoption is particularly problematic when integrating multiple data sources, which has been found to improve parameter estimation, reduce uncertainty, and provide more robust evidence for public health decision making, but often requires more structurally complicated models [De Angelis and Presanis, 2018, Sherratt et al., 2021].

Current approaches linking infectious disease models to data broadly falls into two categories: pipeline approaches, which consist of fitting separate models sequentially for different processes (e.g. data processing, missing data imputation, and parameter estimation), and passing the outputs from one model to the next as an input Mitchell et al. [2022]; and joint modelling approaches that embed these processes in a unified statistical framework [De Angelis and Presanis, 2018]. Pipeline approaches offer computational efficiency and modular development, but may propagate errors, and fail to capture dependencies or to fully account for uncertainty [Lison et al., 2024a, Ward et al., 2024]. Joint modelling can provide more principled uncertainty quantification and better parameter identifiability, but typically leads to complex models and, consequently, requires substantial computational resources[De Angelis et al., 2015, Russell et al., 2024, Lison et al., 2024a]. Aside from the key choice of an appropriate approach, there are numerous challenges when integrating multiple data sources, including: detecting and resolving conflicts between data sources, e.g. when different data streams suggest different insights; combining data sources with different spatial or temporal resolutions; and navigating multiple modelling choices where multi-data source integration choices impose model structure constraints [De Angelis and Presanis, 2018, Nicholson et al., 2022]. Challenges in fitting and calibration, such as computational intractability, parameter non-identifiability, and the need to approximate ideal model structures for practical inference, further complicate model implementation [Corbella et al., 2025a, Ward et al., 2024]. These considerations can impact model design, yet their implications are rarely made explicit in published analyses. Whilst some work has systematically compared modelling choices for specific case studies [Bouman et al., 2024] or provided workflow guidance focused on specific computational frameworks [Grinsztajn et al., 2021], structured guidance remains limited [Behrend et al., 2020, Nicholson et al., 2022, De Angelis and Presanis, 2018].

This paper provides a workflow for infectious disease modelling that applies broadly across model types (e.g. deterministic, stochastic, continuous-time, discrete-time) and model aims, including scenario analysis, disease monitoring, and outbreak response. We provide specific guidance for outbreak settings and multiple data source integration, as our experience has shown that these contexts present particular workflow challenges. The Bayesian paradigm, combining prior information with data, is a natural framework for infectious disease modelling, since such models require at a minimum prior assumptions, usually both prior information and at least one data stream, and potentially, multiple data streams. The approach that we discuss here, therefore, builds on established Bayesian workflows [Green et al., 2003, Gelman et al., 2020], extending them for the specific challenges of infectious disease modelling, including evolving research questions, emerging data sources, and time-constrained development. We advocate for Bayesian thinking throughout the model development process, including model specification, model criticism, and predictive checks.

We start by outlining an approach for characterising epidemiological data source properties through a structured checklist (Section 2), then present an iterative workflow (Section 3), with the checklist informing decisions throughout each workflow stage. The proposed workflow includes defining the research question, developing Directed Acyclic Graph (DAG) representations of disease and observation processes, decomposing

the model into sub-modules, inference and computation choices, model specification and validation, data integration method selection, and real-world considerations. Throughout, we identify feedback loops where insights obtained later on prompt earlier choices to be revisited. We also provide guidance for using the workflow in evolving settings, such as long-lasting outbreaks, and for clearly reporting its implementation.

We provide a case study on reproduction number estimation to demonstrate the workflow, starting with a single data source, then integrating multiple data sources with similar characteristics, and finally including data of different types and scales. By focusing on a common estimation task, we aim to demonstrate broad principles for improving current workflow practice that are applicable across a range of infectious disease modelling contexts. We also provide an example of our suggested data characterisation approach, which we then use throughout our case study to inform our modelling choices. This paper is designed for diverse audiences and readers may wish to engage with the text in different ways. We suggest that most readers start by reading the introduction, data sources and characteristics (Section 2), the overview of the workflow (start of Section 3, including the schematics), and the discussion (Section 6). Depending on the use case, readers may then wish to focus primarily on either the full workflow (Sections 3) or the case study (Section 5). We provide cross-references between corresponding areas of these sections.

2 Data Sources and Characteristics

Using multiple data streams can enable or improve parameter estimation and may reduce uncertainty or bias in modelling studies for monitoring infectious diseases [De Angelis and Presanis, 2018, Sherratt et al., 2021]. While each data source may offer unique advantages, such as timeliness, population coverage, and direct information on some quantities, they also come with distinct limitations, including reporting delays, measurement noise, or lack of representativeness. There is currently no standardised approach to assess and compare data sources in terms of their strengths and weaknesses, hindering our ability to compare, integrate, and prioritise data streams effectively.

The WHO guidelines for integrated surveillance of respiratory pathogens with pandemic potential emphasise the importance of standardised, multi-purpose surveillance systems that also support public health responses and modelling [World Health Organization, 2024]. Inspired by these, we propose six overarching characteristics with which to evaluate different data sources in a common framework (Figure 1). These characteristics are: (1) *metadata* that encapsulates foundational information about each data stream; (2) the *scope* of a data source, i.e. its representative breadth and epidemiological relevance; (3) *resolution* that describes the granularity and level of detail captured by a data source; (4) *quality* that encompasses the reliability, accuracy and completeness of the measurements within a data stream; (5) *utility* that refers to the practical applicability of a data source in informing epidemiological metrics and supporting public-health decision making; and (6)*practical considerations* that address the feasibility of using a data stream in modelling and surveillance contexts. We define more detailed potential criteria for each of these characteristics as part of the case study and in the Supplementary Materials.

We propose that these six categories are useful as a guide for the systematic assessment of a data stream's strengths and weaknesses. While the precise structure, relevance, and sub-attributes of each component may vary depending on the specific surveillance and modelling objectives, the key aim of these categories is to promote structured, transparent thinking. We suggest first assessing our checklist for its relevance to a specific research question, then modifying it to better align if necessary. Once happy with the checklist, we recommend reviewing each dataset available in turn, ideally independently, and collating the responses. We then suggest summarising these responses, as demonstrated in our case study, and using them to inform the remainder of the workflow.

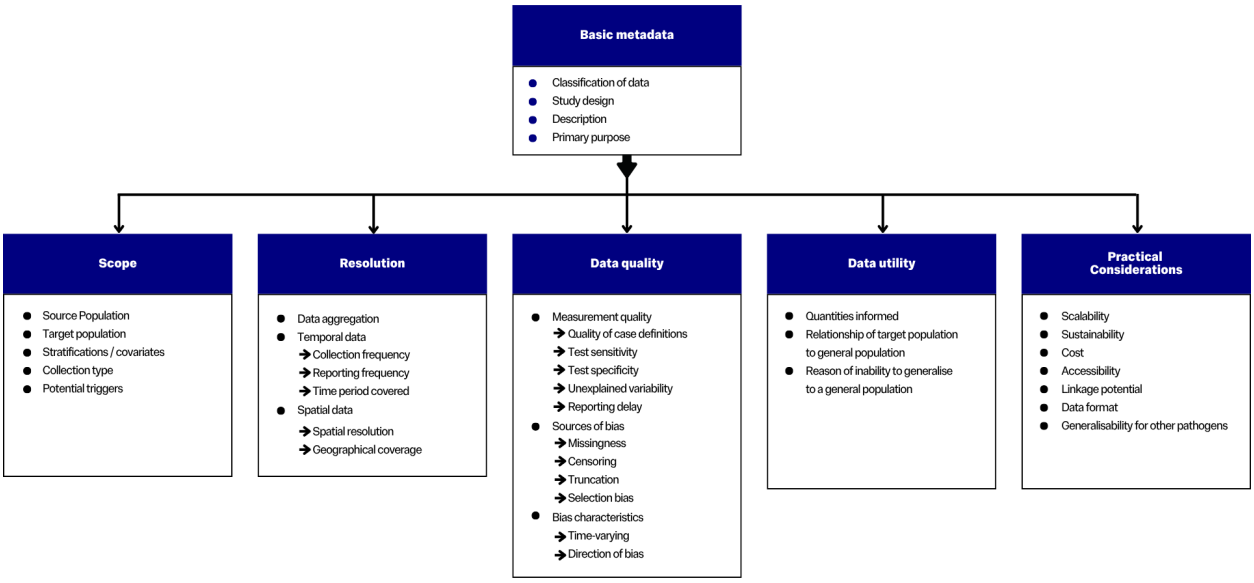

Figure 1: The proposed checklist for characterising a data stream used in infectious disease surveillance and modelling. The six core characteristics—metadata, scope, resolution, data quality, data utility and practical considerations—are illustrated alongside representative attributes that can be used to assess the strengths and limitations of each data source.

3 Workflow

The workflow should be implemented through multiple iterations, starting with high-level considerations and becoming more detailed with each pass, as some steps require decisions in later workflow components. Users may approach this process differently, with some doing fewer, more detailed iterations, and others rapidly working through the workflow initially and then iterating over a greater number of iterations.

We start with clearly defining a **research question** and **target estimands** (Section 3.1, e.g. time-varying reproduction number, variance in the number of secondary infections per index infection). Next, we suggest using a **DAG** to identify dependencies between variables, thereby defining our model structure. We also suggest using a state space formulation [Corbella, 2019] to delineate the definition of the **process** and **observation** models. Starting with the latent process model (Section 3.2), the corresponding DAG represents the underlying epidemiological process, i.e. the transmission and related processes, and the processes required to answer the research question. Even if we do not need to explicitly model the transmission or infection process to answer the research question, an initial process DAG that includes the transmission process is useful to communicate any simplifying assumptions that are being made. Our next step is to use the data source characterisation from Section 2 to **select available data sources** (Section 3.3) that inform our process model. Data sources should be selected based on their expected contributions to the research question. For each data source, we then suggest developing an observation DAG (Section 3.4) that links the process model to data, accounting for measurement systems and model misspecification. A key part of our workflow is iteratively refining these representations (Section 3.5), e.g. in light of preliminary results, as an outbreak evolves, or as the research question is updated. Before continuing with the full model a useful step is to develop a **modularised** version of our joint model (Section 3.6) where the model DAG is decomposed as much as possible into smaller submodels that can each then be independently verified.

Next, we make **inference and computation choices** (Section 3.7): we select an inference method

based on the model structure, and theoretical and practical considerations for each module. We then decide how to **implement** our model in the inferential and computational framework chosen (Section 3.8), ideally following software development best practices in a probabilistic programming language. Finally, we apply **model specification and validation** (Section 3.9) to each module independently. At this stage, we choose the specific model family to implement our model DAG. Potential model families include deterministic or stochastic differential equations, discrete-time models, and agent-based models. We argue that this choice is made in this part of the workflow as it is in effect part of the prior specification. Some workflow users may find it easier to think about this choice earlier in the workflow as it may make specifying the model structure easier to conceptualise. Whilst we think that approach is valid we also recommend always revisiting that choice at this stage of the workflow. Our choice is influenced by available expertise, the model structure and our inference and implementation choices. In some settings, we recommend using a hybrid approach where, for example, the process model is implemented using differential equations and the observation models are discrete-time statistical models. Once the model has been specified, we move on to prior specification, parameter identifiability assessment (including whether or not non-identifiability impacts our ability to answer our research question), and prior and posterior predictive checks.

After validating individual modules, we then face key decisions on **data integration**, namely which modules to combine and how (Section 3.10). If combining multiple modules is not beneficial or feasible, we can proceed to the combination of estimates, e.g. where multiple models are used to estimate the effective reproduction number from different data sources, and we combine the different estimates through ensemble approaches. If combining multiple modules is warranted, we select a data integration method and return to inference and computation choices (Section 3.7) for the newly combined model.

Throughout this workflow, several feedback loops are possible between steps, so that model development is rarely a single forward pass (see Figure 2). The intended use of the model influences decisions throughout the workflow. Data characteristics can alter process DAG structure: for example, individual-level data requires different representations than population-level data. Inference and computation constraints may require approximating or restructuring the process DAG. Practical constraints shape integration choices in multiple ways. Teams using incompatible programming languages may need to use non-joint approaches. Similarly, computational constraints may favour approximate over exact joint approaches or ensembling independent models. Alternatively, they can suggest refining the underlying model DAGs. Our choice of model family may also impact our model DAGs, as well as the inference approach and implementation method. Parameter identifiability issues discovered during model validation may, depending on how they impact our ability to address the research question Cobelli and DiStefano [1980], Yan [2017], prompt us to consider simplifying either the process or observation models. The complexity of integration may also require simplifying models to maintain computational feasibility. Finally, conflicts between data sources during integration often reveal model misspecification, requiring revision of earlier assumptions.

In the following sections, we discuss each of the workflow steps in more detail and include references to relevant resources for further exploration.

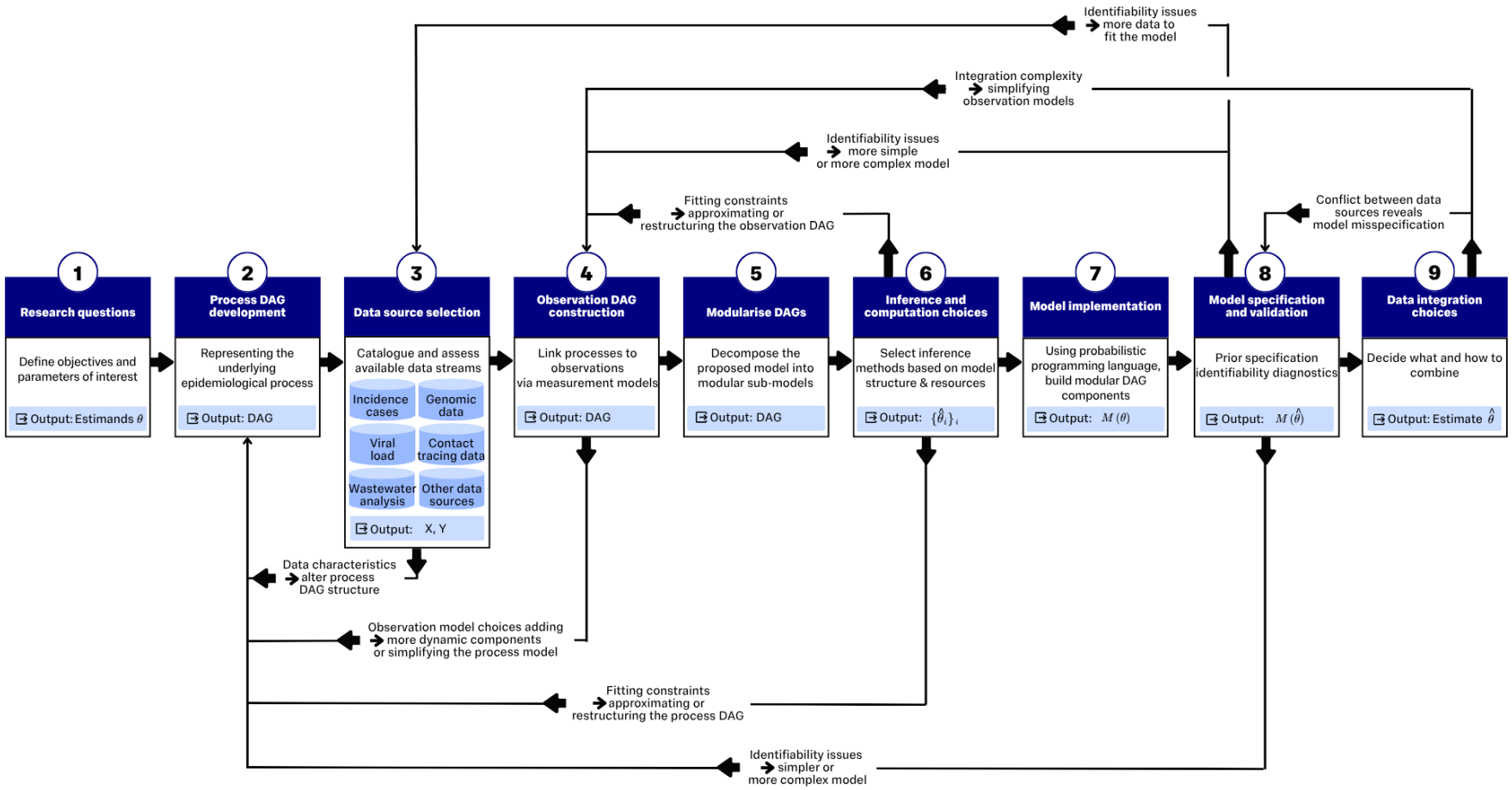

Figure 2: **Workflow for integrating multiple data sources in infectious disease modelling.** Key feedback loops from downstream parts of the workflow that impact earlier choices are represented with dashed arrows and boxes.

3.1 Research Question and Target Estimands

Clearly defining the research question and the epidemiological parameters $\boldsymbol{\theta}$ to be estimated or predicted shapes the entire modelling workflow. Our definitions will result from the main modelling aims, e.g. forecasting hospital admissions or estimating the potential impact of interventions, and are best decided in collaboration with a range of stakeholders [Marshall et al., 2024]. Often, different public health objectives need knowledge of multiple quantities, that require balancing trade-offs. For example, we need both short-term forecasts for operational planning and effective reproduction number (R_t) estimates for public communication and interpretability. Our aim is to translate these policy needs into specific target estimands, by understanding how parameters inform decisions [Nicholson et al., 2022, World Health Organization, 2023]. Policymakers might need to know if transmission is increasing rather than precise values, or whether superspreading drives transmission, to choose between population-wide versus targeted interventions. We recommend defining the scope of an analysis explicitly, including what it cannot address: e.g. a national R_t estimate cannot easily reveal local outbreak dynamics, aggregate case data cannot identify transmission chains, and symptom-based surveillance may not detect mild infections. These limitations help shape stakeholder expectations and prevent misuse of results.

Research questions are likely to evolve throughout outbreaks. Early outbreak questions often focus on growth rates and severity using limited case data. As surveillance expands, questions might shift, for example, to understand variant dynamics, requiring genomic data integration; then to quantifying population immunity, needing serological data [Bhatia et al., 2023, Shearer et al., 2024]. See Section 4.1 for more on dealing with evolving outbreak settings.

3.2 Process DAG Development

Epidemic processes can often be described using stochastic state-space models [Birrell et al., 2018], where latent states $\boldsymbol{X}_t$, e.g. the number of infectious individuals I_t, evolve over (discrete) time according to dynamics

$$\begin{aligned}\boldsymbol{X}_0 \mid \boldsymbol{\theta} &\sim p(\boldsymbol{x}_0 \mid \boldsymbol{\theta}) \\ \boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}, \boldsymbol{\theta} &\sim p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta})\end{aligned}$$

governed by (a subset of) parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The state-space representation is completed by the observation process (Section 3.4), where data are observed conditional on the current state of the system, governed by observation model parameters (a subset of $\boldsymbol{\theta}$, e.g. a reporting rate):

$$\boldsymbol{Y}_t \mid \boldsymbol{X_t}, \boldsymbol{\theta} \sim p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta})$$

so that the full joint probability model in a Bayesian framework of the random variables $(\boldsymbol{\Theta}, \boldsymbol{X_{0:T}}, \boldsymbol{Y_{1:T}})$ is

$$p(\boldsymbol{x}_{0:T}, \boldsymbol{y}_{1:T}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}) p(\boldsymbol{x}_0 \mid \boldsymbol{\theta}) \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{\theta}) p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta}).$$

The state-space framework links the process model to observed data. Although in general state-space models represent stochastic processes over time, models that include deterministic relationships or that are static can be seen as special cases [Birrell et al., 2018].

State-space models can be visualised using DAGs, which represent the joint distribution of all variables (both observed $\boldsymbol{Y}_t$ and latent $\{\boldsymbol{X}_t, \boldsymbol{\theta}\}$) in a model, visualised as nodes. This joint distribution is factorised into the distribution of each variable conditional on its parent nodes, with directed edges representing these dependences. This formal representation makes conditional independence assumptions explicit and easily readable from the DAG, with the basic assumption being that any node is conditionally independent of its non-descendants given its parents [Lauritzen, 1996]. We recommend using DAGs in a state-space framework for structuring infectious disease models because they separate the components which are key to our research question and that influence transmission from how we observe them. For example, in wastewater modelling, the shedding process is usually part of the observation DAG; however, if we think that viral material shed into the environment could also cause infections, it would become part of the process model. This

separation enables modular model development where process components can be reused across different surveillance contexts [Nicholson et al., 2022]. Some common elements of process DAGs might be population structure, contact patterns, infection progression, and immunity dynamics [De Angelis and Presanis, 2018]. Our aim when constructing a process DAG is to translate our research question and target estimands into structures that can be evaluated, communicated, and refined. Rather than starting from scratch, we recommend aiming to adapt established epidemiological models for similar pathogens or transmission routes, modifying components as outbreak-specific data emerges [Gelman et al., 2020]. We can either begin model development simply and add complexity, or start with a comprehensive model and simplify to essential components [Gelman et al., 2020]. Generally, we recommend starting with a simple representation of core transmission dynamics, then iteratively refining as understanding improves or as needed by other parts of the workflow. For example, early outbreak DAGs might represent homogeneous mixing, while later versions could incorporate age structure, spatial heterogeneity, or variant dynamics. However, this workflow is not always possible (e.g. when starting with an already developed model), and so a more complex initial process model may need to be specified, which can then be simplified in later steps, e.g. if the data don't allow identification of all parameters. Biological mechanisms often shape the process DAG structure. For example, incubation periods create latent states between the susceptible and infectious states, with the numbers in each state represented by a node in the DAG; generation time distributions determine numbers of newly infected individuals with this dependence represented by a directed edge; and waning immunity adds nodes representing transition rates from recovered back to susceptible compartments. These structural choices directly affect transmission dynamics and parameter identifiability (Section 3.9).

3.3 Data Source Selection

With the process DAG and target estimands defined, we next select appropriate data sources, drawing from the data review (Section 2), based on their characteristics and the specific epidemiological quantities they inform. For example, case data can be a proxy for infection incidence with reporting delays and ascertainment bias, hospitalisations capture severe outcomes, and wastewater indicates population-level pathogen shedding, independent of healthcare-seeking behaviour. In principle, integrating multiple data sources should enhance the identifiability and precision of parameter estimates [De Angelis and Presanis, 2018, Lison et al., 2024a, Russell et al., 2024, Birrell et al., 2025]. In practice, however, data integration may pose challenges. These include inconsistencies due to unaccounted biases [Presanis et al., 2013, Knock et al., 2021, Ward et al., 2024, Corbella et al., 2025a]; computational complexity [Corbella et al., 2025a]; and operational constraints during an outbreak emergency [McCaw and Plank, 2023]. We therefore recommend starting with minimum data requirements for target estimands and then expanding systematically. We also suggest picking initial data sources that are expected, based on their characteristics, to be the most straightforward to model. We then recommend evaluating complementarity among selected data sources, by assessing whether they provide independent signals on different parameters. For example, wastewater surveillance may provide early warning of outbreaks, while clinical surveillance provides individual-level severity data. Multiple data sources may alternatively inform the same parameter, reinforcing evidence on that parameter and introducing some redundancy. While this redundancy can enhance precision, consistent with meta-analytic principles [De Angelis and Presanis, 2018, Borenstein et al., 2021], it can also increase model complexity.

Finally, we recommend prioritising data sources based on their expected information gain relative to implementation cost and feasibility. "Value of information" methods [Jackson et al., 2019, Heath et al., 2024] can formally guide whether additional data would meaningfully improve precision. However, for complex models especially, parameter identifiability may only become evident after fitting an initial version, noting that assessment of the implications of non-identifiability for addressing the research question requires additional considerations Yan [2017].

3.4 Observation DAG Construction

We now construct the observation process component of the state-space model, again using DAGs to map how latent epidemiological processes generate the observed data $\boldsymbol{Y}_t \mid \boldsymbol{X}_t, \boldsymbol{\theta}$, with complexity shaped by the target estimands and selected data sources [De Angelis and Presanis, 2018, Birrell et al., 2018]. We propose building these DAGs by thinking through each intermediate step that leads from the process model to ob-

servation and then proposing the simplest model that can plausibly represent these observation processes. In future feedback loops, this representation can then be refined as needed to capture the observed data better. For example, for clinical surveillance data, steps may include symptom-based healthcare seeking, obtaining a sample to be tested, testing in a laboratory, and reporting; whilst for wastewater surveillance they may describe pathogen shedding, sewage transport, lab processing, and quantification. Where a step is expected to introduce delays, biases, or induce missing data that affect inference, our observation model should explicitly account for them. For example, we need to adjust for reporting and other delays that create temporal misalignment between transmission events, disease sequelae and observations [Seaman et al., 2022]. It is also important to distinguish between different missing data mechanisms since random missingness and systematic under-ascertainment require different handling [Sherratt et al., 2021]. Often, we may also need to account for time-varying observation probabilities, e.g. driven by changes in testing capacity, healthcare seeking or surveillance intensity. Observation components should ideally be modular and replaceable [Gelman et al., 2020]. For example, initial Poisson noise models may need to become negative binomial to account for overdispersion in reported counts. This flexibility of approach reduces initial specification burden whilst enabling systematic refinement. As for the process DAG, we recommend starting with a simplified observation DAG and iteratively increasing complexity as required by the research question and later stages of the workflow.

Whilst it is hard to give general guidance as to the appropriate observation model for a given setting it is useful to consider some examples. Estimating population-level transmission requires simpler observation models than understanding variant-specific dynamics or age-stratified patterns. Individual-level data enables different, more realistic, observation models compared to aggregated reports. An important challenge when constructing the observation model using multiple data sources is how we handle dependencies between them De Angelis et al. [2015], Corbella et al. [2025b]. For example, data streams may relate to the same underlying epidemiological process; or surveillance systems may have dependencies i.e., individuals may be captured by two or more systems. Developing a DAG will help us identify such dependencies. Finally, we must account for hierarchical structure and population heterogeneity when demographic subgroups are present in the data or are central to the research question. We can validate our understanding of observation processes when we have multiple observation DAGs for the same underlying process by checking whether different data streams imply consistent transmission dynamics when properly integrated.

3.5 Refining the model DAGs

After mapping data sources and constructing our observations DAGs, we revisit our process and observation DAGs to ensure alignment between what we want to model and what our data can support. This step of the workflow aims to create a single joint DAG of our state-space model composed of our process and observation DAGs. Data availability can drive both increases and decreases in process model complexity. For example, contact tracing data with identified transmission pairs allows us to shift from population-level compartmental models to individual-based representations that capture heterogeneous mixing patterns. Conversely, discovering that surveillance only provides aggregate counts may require collapsing a planned age-structured model into a simpler model. Strong prior information can sometimes be used to compensate for data limitations, providing a further source of evidence, and so supporting otherwise difficult to identify model specifications. For example, well-informed priors about transmission parameters from previous outbreaks might support maintaining a risk stratified transmission model even when we have limited data; and detailed knowledge about age-specific contact patterns from previous studies could support using age structure, even if we only have total case counts. Note that assuming fixed parameter values, common in mechanistic modelling literature, is effectively placing infinitely strong priors: we advocate avoiding such fixed values where possible in favour of appropriately uncertain prior distributions. We discuss these considerations in more detail in Section 3.9. Temporal resolution affects what variation can be identified rather than what can be modelled: we can always build a daily-scale model, but weekly surveillance data will usually inform weekly or longer-term patterns, not daily fluctuations. Additional DAG iteration typically occurs at key workflow stages (Figure 2), such as during model specification when parameter identifiability issues emerge, during validation and/or data integration, when misalignment between model structure and data patterns is exposed, or due to practical or theoretical considerations related to model fitting [Corbella et al., 2025a]. Each iteration should address specific problems rather than making arbitrary changes.

In some cases, we may be unable to use domain knowledge or reasoning to produce a single valid DAG and instead have several candidates for either or both the process and observation DAGs. For example, independent teams working on a similar problem might generate different DAGs, particularly when mechanisms are not well-known or the data-generating process is poorly reported. If so, we proceed through the remainder of the workflow with each candidate DAG, comparing them in 3.9 and potentially integrating them in 3.10.

3.6 Modularising DAGs

After selecting data sources and developing and iterating on our DAGs, we decompose our proposed model into modular sub-models (we may already have these modules from generating our observation process DAGs). Each sub-model should be as simple as possible, ideally including the process DAG and a single observation DAG, therefore depending on a single data source. We may need to make additional simplifying assumptions, which will be relaxed during the data integration step. For particularly complex models, we recommend starting with an even more modular approach, with the process and observation DAGs being initially treated separately from each other or even modularising them still further into lower-level components (likely needing to use simulated data and inputs for this step and, as with other modularisation steps, making simplifying assumptions).

We start with simple sub-models as it is easier to diagnose issues such as poor fit, model misspecification or convergence issues for modules than for complex models. We then add complexity incrementally, only as far as needed. This modularisation facilitates detection of inconsistencies or conflicts between data sources [Presanis et al., 2013, Manderson and Goudie, 2023] and offers computational advantages over joint modelling [De Angelis and Presanis, 2018, Goudie et al., 2019, Gelman et al., 2020, Nicholson et al., 2022].

We apply the workflow stages (Inference and Computation Choices, Implementation Considerations, Model Specification and Validation) to each module independently, then, if validated, we proceed to Data Integration Choices (Section 3.10) to determine how to combine them. The integrated model may then require additional cycles through the workflow.

3.7 Inference and Computation Choices

Inference for infectious disease models involves estimating parameters $\boldsymbol{\theta}$ from observed data $\boldsymbol{y_{1:T}}$. We assume these data arise from a probabilistic model with a likelihood function $p(\boldsymbol{y_{1:T}}|\boldsymbol{\theta})$ that links parameters to data as specified in our process and observation DAGs (Sections 3.2 and 3.4). Bayesian inference, our recommended approach, updates prior beliefs $p(\boldsymbol{\theta})$ with the likelihood to obtain posterior distributions $p(\boldsymbol{\theta}|\boldsymbol{y_{1:T}}) \propto p(\boldsymbol{y_{1:T}}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, providing principled uncertainty quantification. This inference task requires balancing model complexity, computational feasibility, likelihood tractability, and available expertise. For this reason, we structure inference and computation decisions through four stages: model complexity, selecting inference methods, implementation, and diagnostics (Figure 3). These choices create feedback loops within the workflow where computational constraints may require approximating or restructuring earlier DAG specifications. This section provides an overview of these considerations.

3.7.1 Model Complexity Assessment

Before determining the inference approach, it is important to assess the complexity of the model. Parameter dimensionality and structure are key to method selection as low-dimensional problems with relatively few parameters require different approaches than problems with hundreds or thousands of parameters (see "Handles high-dim. params.?" in Table 1). Similarly, strongly correlated parameters, discrete parameter spaces, or mixed continuous-discrete problems each impact which methods we can use or how they are implemented. We also need to consider the number of latent variables, such as unobserved infection times, individual-level disease states, or spatially varying transmission rates. When there are many latent variables relative to observed data, computational feasibility becomes a constraint that rules out certain approaches. The tractability of the likelihood also determines which methods can be applied. Key questions we typically ask ourselves include whether we can write the likelihood $p(\boldsymbol{y_{1:T}}|\boldsymbol{\theta})$ analytically, evaluate it numerically at reasonable cost, or differentiate it with respect to parameters. As an example, models with stochastic

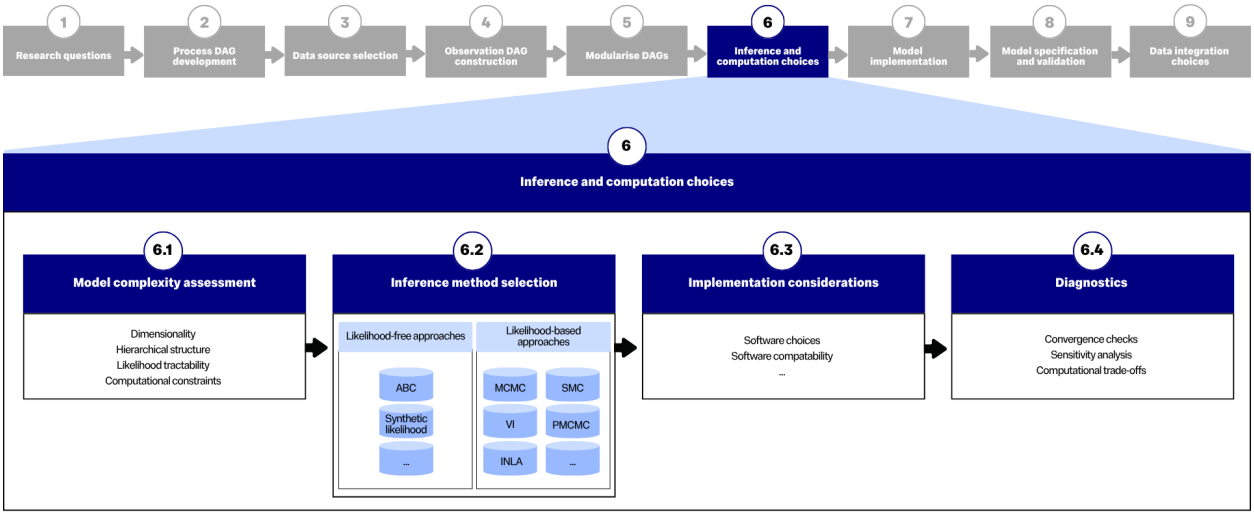

Figure 3: **Inference and computation choices workflow for integrating multiple data sources in infectious disease modelling.** The workflow contains four components: Model Complexity Assessment (6.1) evaluates dimensionality, hierarchical structure, likelihood tractability, and computational constraints; Inference Method Selection (6.2) chooses between likelihood-free approaches and likelihood-based methods; Implementation Considerations (6.3) address software choices and compatibility requirements; and Diagnostics (6.4) involve convergence checks, sensitivity analysis, and computational trade-offs. For simplicity, feedback loops are omitted in this schematic. For a complete representation of feedback mechanisms, see Figure 2.

differential equations or discrete event processes often lack tractable likelihoods, ruling out methods that require likelihood evaluation or gradients.

3.7.2 Inference Method Selection

Numerous methods and algorithmic variants exist within each broad class of inference approach, with distinct computational characteristics and practical trade-offs. Rather than adopting a universal approach, our choice should be tailored to the specific modelling context, inferential goals, and available expertise. The following gives guidance to selecting a method, but does not cover all settings.

Gradient-based Markov Chain Monte Carlo (MCMC) approaches represent our preferred methods when applicable due to their combination of efficiency, reliability, and ease of use [Gilks et al., 1995, Lekone and Finkenstädt, 2006]. MCMC methods generate samples from the posterior distribution by constructing a Markov chain that converges to the target distribution, providing a flexible framework for Bayesian inference when analytical solutions are unavailable. Hamiltonian Monte Carlo (HMC) leverages gradient information for efficient MCMC exploration of differentiable likelihoods, with its adaptive variant No-U-Turn Sampler (NUTS) representing our default recommendation for models with continuous parameters, or where discrete parameters can be marginalised out [Duane et al., 1987, Hoffman et al., 2014, Andrade and Duggan, 2020]. When implementing gradient-based methods, the automatic differentiation method can have a substantial impact: forward-mode differentiation suits models with few parameters (typically fewer than 10), whilst reverse-mode differentiation scales better for high-dimensional parameter spaces [Baydin et al., 2018, Andelfinger, 2023]. In our experience, when it is possible to do so testing different automatic differentiation options can greatly speed up model development. When full uncertainty quantification is less critical, Variational Inference (VI) approximates the posterior by optimising tractable surrogate distributions, offering fast approximation with modern variants like Pathfinder [Blei et al., 2017, Chatzilena et al., 2019]. These approaches can often be useful for prototyping, but we have found that they need to be used with care as they can mask issues with models that less approximate approaches identify.

Sampling-based approaches become necessary when gradients are unavailable. Standard sampling-based MCMC methods including Metropolis-Hastings and Gibbs sampling handle discrete or mixed parameter spaces [Hastings, 1970, Geman and Geman, 1984, Gilks et al., 1995]. For discrete parameters, we recommend considering whether marginalisation is possible before resorting to sampling, as analytical integration can significantly improve efficiency and mixing. Parallel tempering extends these methods to multimodal posteriors by running multiple chains at different temperatures, with modern implementations available in packages like Pigeons.jl [Surjanovic et al., 2023]. For models in which the latent state dimension grows over time, as in state-space models, MCMC becomes increasingly inefficient.

Sequential Monte Carlo (SMC) methods, or particle filters, offer a natural alternative in models with latent states. These methods approximate distributions, e.g. the posterior distribution of latent states, over time using many parallel simulations (referred to as particles), each weighted according to how well it matches the observed data. A key property of SMC is that it provides an unbiased estimate of the likelihood, which makes it particularly useful when embedded in parameter inference schemes [Doucet et al., 2001]. Particle Markov Chain Monte Carlo (PMCMC) exploits this by combining SMC with MCMC for joint state–parameter inference [Andrieu et al., 2010, Endo et al., 2019], while Sequential Monte Carlo Squared (SMC^2) extends the idea to fully sequential Bayesian updating as observations accumulate [Chopin et al., 2013, Temfack and Wyse, 2025]. A significant advantage of SMC methods is that they enable online Bayesian updating as new data arrive, making them well-suited for real-time epidemic tracking [Birrell et al., 2020, Storvik et al., 2023]. PMCMC is widely used for infectious disease modelling including in the Partially Observed Markov Process (POMP) and ODIN ecosystems [King et al., 2016, FitzJohn et al., 2021]. Despite these strengths and being widely used, particle-based methods face challenges such as particle degeneracy (where few particles carry most weight) and memory demands with long time series. Table 1 summarizes their main advantages and limitations. Approximate Bayesian inference methods provide efficient alternatives for specific model structures. Integrated Nested Laplace Approximation (INLA) offers a computationally efficient method for approximating posterior marginals in latent Gaussian models, making it particularly effective for spatio-temporal disease models with spatial correlation structures [Rue et al., 2017]. This method excels when models can be reformulated with latent Gaussian components but struggles with non-Gaussian latent structures.

Beyond Bayesian approaches, Maximum Likelihood Estimation (MLE) provides both point estimates for model parameters and uncertainty quantification in the form of standard errors and confidence intervals. The MLE converges to the true parameter as the sample size increases and is asymptotically efficient, achieving the lowest possible variance among unbiased estimators [Myung, 2003, Baltazar-Larios et al., 2024]. However, uncertainty quantification for complex models can be challenging in a frequentist framework without resorting to computationally expensive bootstrapping. Profile likelihood methods extend MLE by examining the likelihood surface for individual parameters whilst optimising over others, providing confidence intervals without full posterior sampling [Tönsing et al., 2018, Plank and Simpson, 2024]. All of these approaches struggle to incorporate domain knowledge to the extent that priors allow for Bayesian methods, and so we in general, don't recommend them outside of prototyping and where other approaches are not available.

We think that traditional likelihood-free approaches are methods of last resort due to their computational demands and limited diagnostic capabilities. These simulation-based methods bypass likelihood evaluation entirely, instead using model simulations to approximate the posterior distribution. Approximate Bayesian Computation (ABC) approximates the posterior distribution by accepting parameter values that generate simulated data close to the observed data under a predefined distance metric, whilst Synthetic Likelihood (SL) methods approximate the likelihood by assuming Gaussian distributions for summary statistics [Beaumont et al., 2002, Wood, 2010, Price et al., 2018].

Recent advances in machine learning-augmented inference show promise, with neural network-based summary statistic construction [Raynal et al., 2019] and deep learning phylodynamic methods [Voznica et al., 2022] providing potential alternatives to traditional likelihood evaluation. Alongside these advances, recent developments such as generalized adaptive HMC [Bou-Rabee et al., 2025a], gradient-free MCMC algorithms [Bou-Rabee et al., 2025b], microcanonical HMC [Robnik et al., 2023], and normalizing flows [Papamakarios et al., 2021] are making high-performance tools increasingly accessible for applied researchers. As these next generation techniques continue to develop, it is important to revisit and refine current methodological recommendations to ensure they remain aligned with emerging computational capabilities.

Table 1: **Comparison of Likelihood-Based and Likelihood-Free Fitting Methods.** This table focuses on foundational algorithms. More recent methodological advancements that build upon these classical approaches are discussed in Section 3.7. "PPL" stands for Probabilistic Programming Languages. For this review, PPL examples include Stan, PyMC, JAGS, NIMBLE, Turing.jl, etc."PPC" stands for posterior predictive checks.

Feature	**Likelihood-Based**						**Likelihood-Free**	
	MLE	MCMC	SMC*	PMCMC	INLA	VI	ABC	SL
Theoretical Considerations								
Requires likelihood?	Yes	Yes	Yes (estimated)	Yes (estimated)	Yes	Yes	No	No
Handles high-dim. params.?	Poor	Poor	Moderate	Moderate	Good	Good	Moderate	Moderate
Convergence guarantees	Asymptotic	Asymptotic	Asymptotic	Asymptotic	Approx.	Approx.	Approx.	Approx.
Distributional flexibility	Low	High	High	High	Medium	Medium	High	Medium
Approximation error	Exact (asymp.)	Exact (asymp.)	Exact (asymp.)	Exact (asymp.)	Deterministic	Variational	Simulation	Simulation
Practical Considerations								
Computational cost	Low	High	Low–Med	Very high	Low	Low–Med	High	High
Scalability (big data)	Good	Poor	Moderate	Poor	Good	Good	Poor	Moderate
Example software	PPL	PPL	LibBi	POMP, NIMBLE	R-INLA	PPL	abctools, EasyABC, ELFI	ELFI, synlik, BSL
Tuning required?	Minimal	Step size, priors	Number of particles	Complex	Minimal	ELBO opt.	Sum. stats., distance, threshold	Sum. stats.
Capable of real-time inference?	Yes	No	Yes	No	Yes	Yes	No	No
Parallelization potential	High	Limited, chain-level, GPU possible but hard	High, particle-level, GPU possible	Limited, particle-level, GPU possible for simple models	Low, matrix operation hard to parallelise	Medium, gradient parallelisation; GPU possible	High, simulations parallelisable; GPU possible	Medium, GPU possible
Diagnostics	Likelihood ratio	$\hat{R}$, ESS, Trace	ESS	$\hat{R}$, ESS, Trace	Residuals	ELBO, PPC	Acc. rate, PPC	Acc. rate, PPC
Best use case	Simple models	General Bayesian inference	Real-time inference	State-space models	Latent Gaussian models	Fast approximation	Intractable likelihood	Intractable likelihood + summary stats

* Parameter estimation with SMC usually requires modifications or combination with other approaches such as MLE or MCMC.

Hybrid and Nested Approaches Modern Probabilistic programming language (PPL)s, which are programming languages that allow programming using probability distributions and often have inference capabilities, enable sophisticated combinations of inference methods. For example, INLA an be nested within NUTS, using efficient spatial approximations while sampling other parameters. Some PPLs, like Turing.jl, support mixed samplers, applying NUTS to continuous parameters whilst using Gibbs for discrete components Fjelde et al. [2025]. SMC methods can incorporate HMC kernels for parameter moves. This allows them to leverage particle filtering for states and gradient-based sampling for parameters Buchholz et al. [2021], Devlin et al. [2024], Rosato et al. [2024]. These hybrid approaches capitalise on the strengths of multiple methods with the downside of increasing the complexity of the overall inference approach.

When methods fail to initialise properly, we suggest first trying to initialise within prior bounds. If this fails, simpler methods can provide starting values for more sophisticated approaches. We have found that VI often proves effective for initialisation due to its speed and robustness, whilst MLE can provide reasonable point estimates as initial values. However, it is important to be aware that initialisation difficulties may indicate unrealistic prior specifications. This should be considered during model validation (Section 3.9) and result in needing to loop back to refining the model DAG (Section 3.5).

3.7.3 Implementation considerations

It is advisable to balance computational cost against inferential quality when implementing a chosen inference method, considering both model complexity and available computational resources (Table 1) [Funk and King, 2020]. We also recommend prioritising methods with automatic tuning to reduce implementation burden and improve reliability. HMC and NUTS require minimal manual intervention through automatic step size adaptation and mass matrix estimation, making them relatively straightforward to use. Standard MCMC and particle methods demand careful tuning of proposals, particle counts, and resampling thresholds through pilot runs. ABC approaches require the most expertise for selecting summary statistics, distance metrics, and tolerance thresholds. INLA and VI may require upfront model reformulation to fit their frameworks.

Parallelisation can determine whether methods can scale to population-level inference or handle real-time analysis during outbreaks. Methods differ substantially in their parallelisation capabilities (Table 1). MCMC methods face inherent parallelisation challenges due to their sequential chain structure. Multi-chain parallelisation provides the primary scaling mechanism. Parallel tempering exploits multi-core architectures through temperature-parallel chains [Surjanovic et al., 2023]. Some implementations support within-chain parallelisation for gradient calculations (Stan and Turing.jl) or matrix operations (Graphics Processing Unit (GPU) acceleration). VI scales well through mini-batch parallelisation and routinely uses GPU acceleration for large datasets [Hoffman et al., 2013, Abbott et al., 2021]. Sequential and particle methods excel at parallelisation through particle-level parallelism. Modern GPUs enable thousands of particles to be processed simultaneously. Memory requirements scale with particle count [Henriksen et al., 2012]. Approximate methods show mixed potential for parallelisation. INLA has limited opportunities due to sequential matrix operations, though some implementations exploit parallel linear algebra libraries. Nested approximations can parallelise across spatial or temporal components when model structure permits. Likelihood-free methods provide parallelisation opportunities for rejection-based ABC. Rejection ABC parallelises trivially across Central Processing Unit (CPU) clusters and GPU architectures. ABC-MCMC retains the sequential constraints of MCMC methods. Memory requirements can become substantial for large simulated datasets [Kulkarni et al., 2022]. Deterministic approaches parallelise most effectively. MLE is highly parallelisable across data points on both CPUs and GPUs. Profile likelihood calculations naturally parallelise across parameter values.

3.7.4 Diagnostics

The quality and availability of diagnostic tools for checking the performance of inferential algorithms vary substantially across inference methods, representing a key factor in method selection (see Table 1 for method-specific diagnostics). Implementations of HMC and NUTS provide the gold standard with divergence warnings, Bayesian Fraction of Missing Information (BFMI), energy plots, and traditional MCMC diagnostics like $\hat{R}$ and Effective Sample Size (ESS), making them easier to use reliably [Betancourt, 2017, Carpenter et al., 2017]. Standard MCMC and parallel tempering offer moderate diagnostic capabilities through traditional convergence statistics and trace plots, with parallel tempering adding temperature swap rates to monitor multimodal exploration [Honkela, 2020]. SMC and PMCMC rely on particle-specific diagnostics. SMC tracks

ESS and weight degeneracy to detect particle impoverishment, while PMCMC combines these with standard MCMC diagnostics and SMC-based likelihood variance to ensure proper mixing and acceptance [Michaud et al., 2021, Andrieu et al., 2010]. In contrast, VI provides only Evidence lower bound (ELBO) convergence monitoring (which does not guarantee accurate posterior approximation) [Blei et al., 2017, Chatzilena et al., 2019], whilst ABC methods offer minimal diagnostics through acceptance rates alone. Diagnostic failures across any method indicate potential model specification issues requiring validation (Section 3.9), with specific diagnostics like divergent transitions in NUTS suggesting particular problems, such as an issue with the posterior geometry in this case.

3.8 Model Implementation

For implementing our models, we recommend using a PPL where possible. PPLs and their ecosystems make many tasks in model specification and validation more straightforward (Section 3.9). For models with complex stochastic simulation components or those lacking tractable likelihoods (Section 3.7), a PPL may provide less benefit. However, using the same implementation approach for all model components will usually require the least amount of effort. This means that even in settings where the combined model derives relatively little benefit from using a PPL, it can still be a good choice due to streamlining the development of submodels.

Generative PPLs, where we can sample from and compute gradients for the same model specification, enable simulation-based validation approaches without separate data generators Štrumbelj et al. [2024]. Turing.jl, PyMC, and NumPyro, amongst others provide this capability [Ge et al., 2018, Fjelde et al., 2025, Abril-Pla et al., 2023, Phan et al., 2019]. Stan offers the most mature diagnostic and validation tools [Carpenter et al., 2017], but requires separate simulation code. JAGS and NIMBLE [Plummer et al., 2003, de Valpine et al., 2017], inspired by the original BUGS language [Lunn et al., 2013], offer flexibility for model structures that may not be well-suited to gradient-based methods. Turing.jl is likely the most composable option [Ge et al., 2018], which proves useful when building models from modules as we recommend (Section 3.6). Speed is difficult to assess and is often model and context-specific, though generally Stan and NumPyro are the most efficient options when applicable [Carpenter et al., 2017, Phan et al., 2019]. Selecting a flexible implementation approach is important for evolving settings (Section 4.1).

Making use of standard software development practices is likely to improve all aspects of our workflow [Gelman et al., 2020]. These include: unit testing to verify model components, continuous integration to catch errors early, version control to track model evolution, and comprehensive test coverage for data processing and diagnostics.

Once we have chosen our software implementation, we translate our modularised DAGs (Section 3.6) into separate code modules. We aim to structure observation models as interchangeable components connecting to the process modelthrough defined interfaces (e.g., both case and wastewater modules interface through infection incidence). Similarly, within each observation model and within the process model, we try to structure these as independent components as much as possible. We also recommend documenting interfaces explicitly to enable component reuse across teams and maintaining consistent parameter naming between mathematical formulation and code when possible.

3.9 Model Specification and Validation

This part of our suggested workflow closely mirrors the workflow suggested in Gelman et al. [2020]. The main difference is that we first need to pick the kind of modelling approach to use, and we explicitly suggest starting by validating models in small subcomponents first, then exploring combinations to identify integration challenges early, and finally repeating this step with the full model.

3.9.1 Choosing a model family

We begin by selecting a model family appropriate for the research question, available data, inference method, and team expertise. Potential model families include compartmental models (population-level transitions between disease states) and agent-based models (individual-level representations) [Skrip and Townsend, 2019]. Compartmental models can be formulated as continuous-time (ordinary or partial differential equations) or discrete-time (difference equations), and as deterministic or stochastic. We find that expertise in a given

framework often outweighs other considerations. Specifications are also commonly hybrid: for example, a continuous-time compartmental process model paired with discrete-time statistical observation models [Bouranis et al., 2025]. This means that whilst the choice of model family can be made considering the complete model it is also valid to make this choice for each module. Once the model family has been chosen, earlier workflow decisions, particularly DAG complexity, inference approach, and implementation method, may need to be revisited. We can now implement our model in the chosen model family making sure to follow the recommendations from Section 3.8.

3.9.2 Prior Specification and Prior Predictive Checks

We specify priors using domain knowledge where available through informative priors: for example, if we suspect that the reproduction number R_0 for a new pathogen is similar to a previous pathogen, but want to allow for the possibility some difference, we might use a previous estimate but with a wider prior interval than the original estimate's uncertainty; or we might use prior elicitation techniques [O'Hagan et al., 2006] to formalise expert opinion on a biological quantity. When substantive prior information is not available, we use weakly informative priors that regularise without dominating the likelihood. We recommend avoiding uniform priors, as often in multiple-parameter models these lead to very unlikely prior predictive outcomes.

Prior-predictive checks simulate data from the prior to ensure model behaviour is epidemiologically plausible before seeing data and to detect prior-data conflict [Box, 1980, Yang et al., 2025]. These checks can reveal several issues requiring model modification: unrealistic parameter ranges may indicate the need for different priors or model constraints; implausible epidemic dynamics may suggest process DAG structural changes; and computational issues during simulation may indicate the need for model simplifications or reparameterisation. Results of prior-predictive checks may indicate a need to either: revisit prior specification, e.g. to obtain additional data sources, previous estimates or expert consultation to inform appropriate prior distributions; or return to DAG development (Sections 3.2 and 3.4) to simplify models, add model components, increase complexity through hierarchical structure, or implement approximations for computational tractability.

3.9.3 Model Fitting and Computational Validation

After resolving issues raised by prior-predictive checks, we fit the model using our chosen method from Section 3.7. We validate computational aspects, such as convergence and algorithm performance, through diagnostic checks, synthetic data simulation, and simulation-based calibration where feasible. Using simulated scenarios to validate parameter recovery provides valuable insights into model performance [Bouman et al., 2024], whilst simulation-based calibration for each submodel provides stronger validation by verifying that inference algorithms can recover true parameters from a range of simulated data [Talts et al., 2018]. We suggest addressing computational issues systematically: simplify model structure or increase prior informativeness if convergence fails; examine submodels in detail if joint fitting proves intractable; reparameterise models if mixing is poor; plot intermediate quantities like effective reproduction numbers to identify problematic components; run on data subsets or reduce iterations to speed up iteration; and check for multimodality using multiple chains or inference approaches designed for handling multiple modes, such as parallel tempering [Gelman et al., 2020]. Our goal is to make predictions or learn about epidemiological processes, not just achieve convergence of an arbitrary model. If computational validation succeeds, we provisionally accept the model for evaluation and use. Persistent computational issues require returning to model specification or considering alternative integration approaches (Section 3.10).

3.9.4 Evaluate and use the model

We use posterior- and mixed-predictive checks [Rubin, 1984, Gelman et al., 1995] to assess whether fitted models reproduce key features across all integrated sources. These checks form the foundation of model validation by comparing model-generated predictions against observed patterns. To avoid overconfidence and conservative checks due to double use of the data and to respect temporal data structures, we suggest applying cross-validation strategies, particularly leave-one-out and time-aware validation [Dautel et al., 2023]. The influence of individual data points and prior specifications should be examined through sensitivity analysis.

To assess parameter identifiability, we examine whether data sources provide sufficient information to uniquely estimate target parameters: joint modelling often resolves identifiability issues that plague pipeline approaches [Lison et al., 2024a, Russell et al., 2024]. Value of information approaches [Jackson et al., 2019, Heath et al., 2024] can indicate where additional data or prior information is needed. While ideally all parameters would be uniquely identifiable, whether this is *necessary* to address the research question(s) requires additional assessment. It may be that non-identifiability of certain subsets of model parameters does not invalidate the modelling approach, provided that it is appropriately included in the uncertainty quantification (see Cobelli and DiStefano [1980], Yan [2017] and the case study in Section 5).

Conflict in the context of data integration refers to when sources of information provide inconsistent or incompatible evidence on model parameters. Such conflicts may arise between prior assumptions and data ("prior-data conflict"), between different data sources, or even within a data source but between different units of data [Presanis et al., 2013, Thomas et al., 2015, Yang et al., 2025]. They typically manifest as a lack of identifiability (with algorithm convergence problems as a symptom) or a poor fit to one or more data sources [Presanis et al., 2013, De Angelis and Presanis, 2018]. Detection methods based on generalising cross-validatory posterior-, mixed- or prior-predictive checks to any latent node in a DAG [Presanis et al., 2013, Yang et al., 2025] usually also measure the extent of conflict. We can use complementary prior sensitivity analysis to quantify how much priors must change to accommodate observed data, revealing potential model misspecification or data quality issues [Roos et al., 2015, Kallioinen et al., 2024, Yang et al., 2025]. We resolve conflicts through careful consideration of unaccounted observational biases or identifying overly restrictive model assumptions. We adjust model structure as needed, including loosening prior assumptions, implementing weighting approaches to account for selection biases, and other bias adjustment approaches [De Angelis and Presanis, 2018].

If model validation reveals issues, we return to DAG development (Sections 3.2 and 3.4) to modify model structure, add complexity, or implement different approximations.

Once we have validated the model, we can use it to answer our research question. Although we designed the model with this question in mind (Section 3.2), it is often only when we come to use the model that we identify limitations. The model may fit observed data well and include the necessary components on paper, but lack sufficient detail or nuance to address the question we want to ask. Stakeholders may also identify misalignment between model assumptions and their domain knowledge that was not apparent during specification. We have found that adjustment may be needed at this stage, as seeing model output often helps communicate assumptions to non-modellers more effectively than written specifications. These considerations may create a feedback loop to process DAG development (Section 3.2).

Depending on the research question, there may be additional considerations when using the model. For example, for counterfactual scenarios (e.g. evaluating interventions), we must make sure to propagate uncertainty from data fitting and integration through to each scenario. We also need to ensure that random events unaffected by the counterfactual scenario remain identical between simulations, for example by using the same draws from the joint posterior distribution over parameters. Similarly, for forecasting applications, we have found that using the model can highlight important performance considerations related to the intended forecast horizon. Short-term forecasting often prioritises recent data fit and computational speed for rapid updates, whilst long-term forecasting may require representation of mechanisms that persist beyond the calibration period, such as behavioural responses or waning immunity. We must validate performance at the relevant timescale to ensure the model maintains realistic behaviour over the intended horizon.

3.9.5 Model Comparison

Model comparison is the final validation step. It aims to identify which of multiple candidate models are the most plausible. When multiple plausible model specifications exist, we recommend using leave-one-out cross-validation (LOO-CV), Watanabe-Akaike information criterion (WAIC), out-of-sample continuous ranked probability score (CRPS), or other methods to compare predictive accuracy [Yao et al., 2018, Gneiting and Raftery, 2007] combined with visual and/or quantitative posterior-predictive checks. Typically, when modelling infectious diseases, it is more suitable to use time-aware forms of these metrics. These metrics, such as high Pareto-k values from LOO-CV, can also often be used to identify influential observations requiring careful modelling. For infectious disease models, we consider scoring transformations such as the log when using CRPS [Bosse et al., 2023], as this approach is likely to better capture the data generating process.

The general aim is to maximise sharpness subject to calibration, i.e. models should produce statistically consistent predictions whilst minimising required uncertainty. [Gneiting and Raftery, 2007].

3.10 Data Integration Choices

Decisions about data integration are inherently context-dependent, but the decision tree in Figure 4 gives an example of how we might think through such decisions. We recommend combining the modularised models sequentially and repeating the validation processes outlined in 3.9 for each newly combined module, until the full model has been implemented and validated. We further recommend using joint modelling of these modules where possible, taking advantage of Markov melding [Goudie et al., 2019] or its approximations, again where possible, to achieve the joint model in a computationally efficient way (Section 3.10.1). We can also make use of cut posterior distributions [Plummer, 2015] to downweight the influence of less trusted modules if needed, although a general framework to carefully achieve cutting feedback is still in its infancy [Liu and Goudie, 2025]. This joint modelling approach offers a coherent framework for integrating data sources, in the spirit of Bayesian hierarchical models [Gelman et al., 2020, De Angelis and Presanis, 2018], with suitable uncertainty quantification and avoiding the definitional ambiguities that affect output-level combinations [Manley et al., 2024, Brockhaus et al., 2023].

Ensemble methods (Section 3.10.2) are valuable when joint modelling proves impossible due to computational constraints, data access limitations, or institutional boundaries between research teams. When employing ensembles, we recommend paying careful attention to ensuring all models target the same estimand, resolving temporal alignment, and documenting methodological assumptions. The choice of ensembling or meta-analytic approach should align with the research objective (e.g. estimation or prediction) and the overlap or otherwise of data used in each model, influencing how estimates or predictions are weighted (e.g. by precision or prediction accuracy; or downweighted to compensate for overlapping use of the same data). In contrast, when integrating sub-models, the approach will depend on whether existing sub-models have already been fitted or if the models are being developed de novo.

In addition to prior-data conflict and between-data source conflict, different modules may also conflict. These types of conflict should be assessed during the model validation and specification process, as discussed in Section 3.9 [Presanis et al., 2013, Sherratt et al., 2021, Yang et al., 2025], once the data integration process has been decided upon.

3.10.1 Full joint modelling

A fully joint modelling approach is especially suited when there are complex dependencies between data sources [Corbella et al., 2025a], perhaps conditional on the parameters of the underlying process model, such that identification of all parameters is compromised when some data are left out. However, joint models can be computationally intensive and challenging to fit, especially in high-dimensional or real-time settings. Care is needed to ensure complex dependencies are appropriately modelled and that observational biases, such as selection biases, are adequately accounted for in the model, to avoid conflict between data sources (see Section 3.9) [Presanis et al., 2013, Corbella et al., 2025a].

Markov melding [Goudie et al., 2019] allows a full joint model to be fitted in a computationally efficient approach, by combining sub-models in such a way that the melded posterior is the full joint posterior. This posterior is achieved by using posterior samples from one sub-model to serve as a proposal distribution for the next, potentially in a sequential chain of models [Manderson and Goudie, 2023]. Similar sequential strategies, using the posterior from one sub-model as the prior (rather than proposal) for another, have a long history [West and Harrison, 1997], closely related to SMC algorithms [Doucet et al., 2001] (see Section 3.7).

While the theory and some applications of Markov melding are well-established [Goudie et al., 2019, Nicholson et al., 2022, Manderson and Goudie, 2023], practical implementation remains challenging. Limitations include the availability of user-friendly Bayesian software and methods for assessing sub-model compatibility and combining them [Goudie et al., 2019, Yang et al., 2025]. As a result, approximate methods are often used. A common approach is the one used in standard meta-analysis [Borenstein et al., 2021] to combine previous estimates using a normal approximation. A Stage 1 point estimate, $\hat{y}_1$, and its corresponding standard error or posterior standard deviation $\hat{\sigma}_1^2$, can be incorporated in a Stage 2 model as a

likelihood term, on an appropriate scale:

$$\hat{y}_1 \sim N(\theta, \hat{\sigma}_1^2)$$

for some parameter θ in the Stage 2 model informed by the Stage 1 estimate. This approach has been shown to be an approximation to Markov melding with "product of experts" pooling [Goudie et al., 2019] and we recommend it unless other less approximate approaches are available.

Modular approaches also support judgements of selective trust in sub-models, as indicated by the name "modularisation" in the first appearance of a "cut" posterior in the literature [Liu et al., 2009]. This approach 'cuts' feedback from one module to another, i.e. preventing less reliable sub-models from influencing more trusted ones while allowing the reverse, providing robustness against sub-model misspecification [Plummer, 2015, Carmona and Nicholls, 2022, Yu et al., 2023]. Liu and Goudie [2025] provide a general framework for deciding how to partition and restrict information flow between sub-models. As with Markov melding, however, practical implementation remains a challenge, since it involves a potentially computationally expensive multiple imputation approach for each MCMC sample [Plummer, 2015].

3.10.2 Ensembling independent models

When multiple models target the same estimand, ensemble methods such as Bayesian stacking [Yao et al., 2018], model averaging [Hoeting et al., 1999], or meta-analytical pooling [Jackson et al., 2011] can be used to combine estimates. These approaches are related to model comparison, as discussed in Section 3.9. They differ from approaches like Markov melding in that they combine independent model outputs rather than partitioning a joint model into connected sub-models. Our choice of ensemble strategy should reflect the research objective and the relationship between the models and the underlying data. When models are fitted on the same dataset, ensemble weights can be based on predictive performance, posterior precision, or model credibility, with Bayesian stacking or model averaging providing principled frameworks for combining the shared estimand [Yao et al., 2018]. In contrast, when models are fitted on different datasets targeting the same estimand, traditional model averaging is less formally justified, as marginal likelihoods cannot be compared across disjoint data. In such cases, posterior pooling or stacking on a common predictive task (e.g., predicting future case counts or other observable outcomes implied by the estimand) can be used, or weights can be assigned based on the reliability or relevance of each data source. Ensemble methods offer practical advantages when joint modelling proves infeasible, as they can enable collaboration between research teams and potentially improve robustness over single models. For example, the European COVID-19 Forecast Hub demonstrated performance gains from simple median aggregation [Sherratt et al., 2021].

However, important limitations constrain ensemble approaches for infectious disease modelling. When models make different assumptions about latent quantities, apparent heterogeneity may reflect definitional differences rather than true uncertainty about the estimand [Brockhaus et al., 2023]. Temporal alignment becomes particularly problematic for latent targets where the underlying definition remains ambiguous, with models using different reference points, generation time distributions, or reporting delay assumptions [Brockhaus et al., 2023]. These limitations can lead to estimates that are less timely and/or less precise, with lower utility for decision makers than the component models. An example of this is the UK's Scientific Pandemic Infections group on Modelling (SPI-M) R_t consensus, which used random-effects meta-analysis to pool R_t estimates from multiple academic groups [Manley et al., 2024]. These estimates shared data sources, made different assumptions, and had differing temporal alignment. Combining outputs at the prediction level also discards information about data conflicts and model disagreements that would be preserved in joint modelling approaches. Weighted ensemble methods often show limited improvement over simple aggregation in multi-team forecasting applications [Sherratt et al., 2021]

4 Workflow Management

4.1 Outbreak Evolution and Workflow Iteration

Two types of changes during outbreaks require revisiting our suggested workflow. First, new information emerges, for example, about pathogen characteristics, surveillance capabilities, or population dynamics that may necessitate updating model assumptions and data integration choices [McCaw and Plank, 2023]. Substantial new data sources may also become available, such as genomic surveillance or serological surveys;

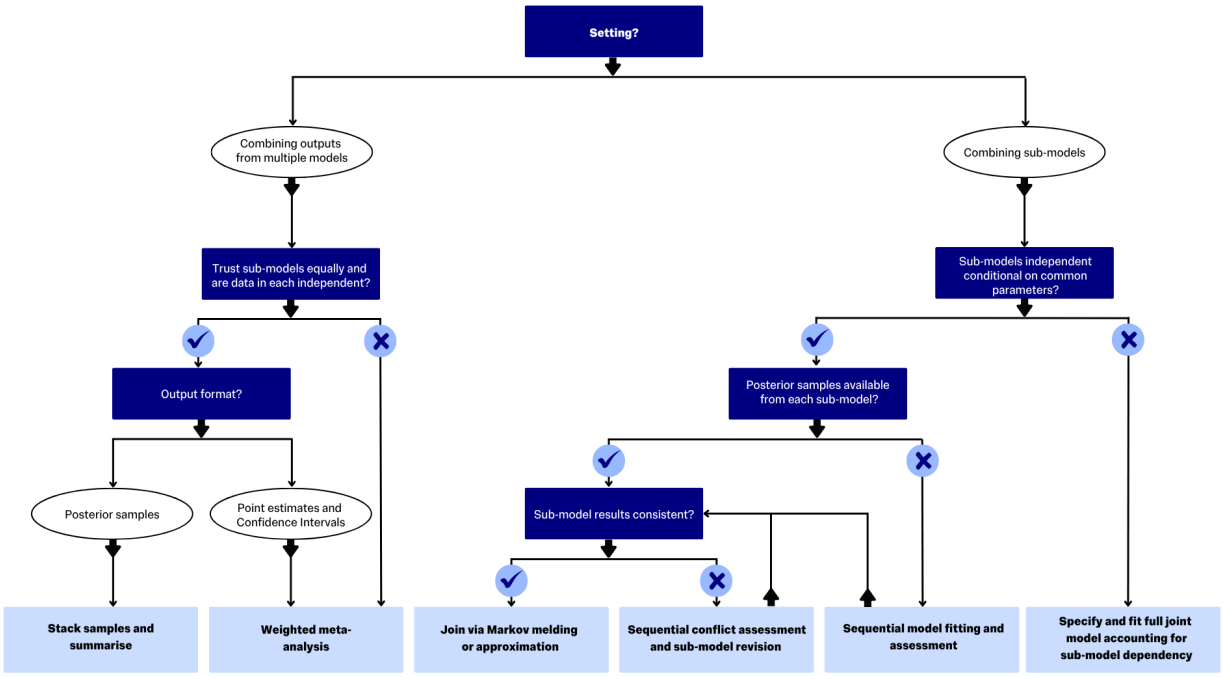

Figure 4: **Example decision tree for integration choices when integrating multiple data sources in infectious disease model inference.** Decisions depend on whether: we are combining outputs or sub-models; we trust models equally or sub-model results are consistent; format of outputs; and dependence or not of data underlying each model/sub-model. For example, if we are combining outputs and the data in each model are not independent, standard meta-analysis would overstate the certainty of estimates due to multiple use of the same data, so we might want to downweight individual models to compensate for the overconfidence.

or our understanding of pathogen biology may change our process model assumptions [Knock et al., 2021]. Second, research questions commonly evolve as outbreak priorities shift from initial detection and characterisation towards intervention evaluation and long-term planning [World Health Organization, 2023], altering target estimands and, therefore, our process model structure. When these changes occur, we recommend iterating through the workflow, revisiting earlier decisions, rather than developing entirely new models.

The modular approach we advocate enables this approach, as we can update specific workflow components whilst preserving elements from previous iterations. For example, observation models developed for case surveillance can often be reused when adding death reporting. On the other hand, process models for transmission dynamics may require updating when new variants emerge, whereas we may be able to reuse observation processes as they are. This incremental approach reduces development time and maintains continuity across evolving outbreak contexts.

These steps are often carried out on an ad-hoc basis by teams during outbreaks. We think that a more formal approach that tracks decisions is likely to improve the quality of models produced and be ultimately less resource-intensive. We recommend establishing regular review points to assess whether workflow iteration is needed, with more regular reviews during acute outbreak phases.

4.2 Reporting

We have found that complex multi-source models require clear documentation to enable evaluation and reproduction. Without structured reporting, even well-designed workflows are challenging to assess or improve. Reporting decisions made at each stage of the workflow is essential, including rationales for data source selection, integration method choices, model structure assumptions, and validation procedures undertaken. Making analyses public at all stages builds trust, though confidential data requires careful management [Abbott et al., 2021, Abbott and Funk, 2022]. We recommend sharing code and data alongside results because it enables verification, accelerates progress, and is just as much a part of the analysis as the manuscript. Not sharing code, sharing it on request, or only once a paper is in a peer-reviewed journal all create barriers to assessing model development and criticism workflows. Where barriers to transparent reporting exist, it is always better to report as much as possible. Mathematical specifications, validation code, and synthetic examples all contribute to transparency even when data cannot be shared, e.g. helping readers understand model behaviour [Mellor et al., 2025]. We also recommend visual schematics of the workflow and model architecture are developed and referred to throughout the reporting process.

5 Case Study

We demonstrate our iterative workflow through a case study schematic with four parts, each increasing in complexity and showing different integration choices and methodological considerations. The main estimand is the time-varying reproduction number (R_t), while in the fourth part we add the overdispersion parameter (k) as an additional estimand.

To support this case study, we first provide examples of our suggested data review approach (Section 2), focusing on the data streams we use in the case studies, based on an example structured questionnaire. The examples illustrate how the strengths and limitations of each stream are reflected in practice, including variability in factors such as resolution, data utility, and practical considerations. This characterisation shows how these features shape the suitability of each stream for integration, modelling, and inference.

The case study illustrates progressive integration of data sources by developing a series of four models. Case study model 1 uses a single data source, reported case counts. Model 2 combines case data with death data, overcoming some limitations of each single source. Model 3 adds wastewater data, which are independent of surveillance biases but have their own drawbacks and biases, as a third source. Models 1–3 all use aggregated data, which means they are limited in their ability to capture individual variability. Model 4 combines case counts with individual-level transmission pair data to estimate variability in the number of secondary infections per index infection.

Whilst this case study focuses on discrete-time branching process and renewal equation models, there is a wide range of other infectious disease modelling approaches. Our workflow applies in these more general settings.

5.1 Data Sources and Characteristics

During an in-person workshop, we designed a questionnaire for evaluating data characteristics across data sources to support both qualitative and quantitative assessment (Supplementary Material S2). As a group, we then completed the questionnaire, drawing on our collective experience modelling infectious diseases during the COVID-19 pandemic. We characterised confirmed case notifications, deaths, transmission pairs, and wastewater data, as well as additional data sources presented in the Supplementary Material (S2). Our approach was ad hoc and is intended to be representative of the review process rather than a definitive assessment.

As suggested in our workflow, before starting to design a model in response to a research question, it is often useful to characterise the kinds of data sources that are available. Here we see that our responses varied both by individual and data sources, reflecting differences in interpretation, disciplinary background, and experience (Figure 5, Table 2). We saw confirmed cases and deaths as mostly serving clinical purposes (75% and 80% of responses), while we viewed transmission pairs and wastewater as non-clinical (100% responses). We thought that scope differed with 88% of us seeing confirmed cases as targeting the whole population based on the data, compared to 80% for deaths, 20% for transmission pairs, and 50% for wastewater. We also thought that stratification was common in transmission pairs (80%) and confirmed cases (75%), but absent in wastewater. In these example datasets, we considered data formats to range from fully individual-level (100% for transmission pairs) to fully aggregated (100% for wastewater). We felt that temporal data were present in all sources, but spatial resolution varied, with wastewater rated highest (100%). For prevalence, we thought that confirmed cases were split (25% direct, 38% indirect, 38% not informed), and wastewater was consistently indirect (100%). We also thought that scalability was linear for cases (63%), exponential for transmission pairs (60%), and independent for wastewater (83%). Our quantitative ratings reinforced these patterns, with deaths scoring the highest test sensitivity (mean: 4.8 [range: 4-5]), and wastewater and deaths rated lowest for linkage potential (2 [1,5] and 2[1,3], respectively). Our example responses highlight how expert assessments can diverge across contexts, even within a shared framework. The variation in our views highlights the importance of acknowledging context-specific practices when evaluating data sources.

Table 2: Categorical survey responses for confirmed time series (8 responses), deaths (5 responses), transmission pairs (5 responses), and wastewater (6 responses).

Core characteristic	**Characteristic**	**Confirmed cases time series** N = 8[1]	**Deaths** N = 5[1]	**Transmission pairs** N = 5[1]	**Wastewater** N = 6[1]
Basic meta-data	**Study design**				
	Contact tracing	0%	0%	100%	0%
	Routine surveillance	88%	100%	0%	50%
	Sentinel surveillance	13%	0%	0%	50%
	Primary purpose				
	Clinical management	75%	80%	0%	0%
	Other	25%	20%	100%	100%
Scope	**Target population**				
	Convenience	0%	0%	20%	0%
	Geographic structure	13%	0%	0%	50%
	Geographic structure, Whole population	0%	20%	0%	0%
	Healthcare workers	0%	0%	20%	0%
	Risk-specific	0%	0%	20%	0%
	Specific age groups, Risk-specific, Convenience	0%	0%	20%	0%
	Whole population	88%	80%	20%	50%
	Stratification				
	Clinical	0%	0%	40%	0%
	Demographic	38%	40%	20%	0%

Continued on next page

Table 2: Categorical survey responses for confirmed time series (8 responses), deaths (5 responses), transmission pairs (5 responses), and wastewater (6 responses). (Continued)

	Demographic, Clinical	38%	20%	40%	0%
	None	25%	40%	0%	100%
	Collection type				
	One-off	0%	0%	20%	0%
	Routine	100%	100%	0%	100%
	Triggered	0%	0%	80%	0%
	If triggered, potential triggers				
	Greater than a threshold	0%	100%	0%	NA
	New variant/pathogen detected	100%	0%	100%	NA
	Unknown	7	4	1	6
Resolution	**Data aggregation**				
	Aggregated	50%	60%	0%	100%
	Individual	50%	40%	100%	0%
	Temporal data				
	Yes	100%	100%	100%	100%
	Collection frequency				
	Continuously	25%	40%	60%	17%
	Daily	63%	60%	20%	0%
	Multiple times a week	0%	0%	20%	67%
	Weekly	13%	0%	0%	17%
	Reporting frequency				
	Continuously	0%	0%	20%	0%
	Daily	50%	40%	40%	0%
	Less frequently than monthly	0%	0%	20%	0%
	Multiple times a week	0%	0%	20%	33%
	Weekly	50%	60%	0%	67%
	Time period covered				
	Continuous	88%	100%	0%	83%
	Early outbreak	0%	0%	100%	0%
	Endemic	13%	0%	0%	17%
	Spatial data	63%	80%	80%	100%
	Spatial resolution				
	Hyper-local	0%	0%	40%	0%
	Local	13%	40%	40%	100%
	National	25%	20%	0%	0%
	Regional	63%	40%	20%	0%
Data utility	**Quantities informed [Basic reproduction number]**				
	Direct	0%	0%	0%	17%
	Direct, Indirect	0%	20%	20%	0%
	Indirect	88%	80%	80%	17%
	Not informed	13%	0%	0%	67%
	Quantities informed [Time-varying reproduction number]				
	Direct	50%	0%	0%	17%
	Indirect	50%	100%	100%	83%
	Quantities informed [Time-varying / basic reproduction number ratio]				
	Direct	0%	0%	0%	17%

Continued on next page

Table 2: Categorical survey responses for confirmed time series (8 responses), deaths (5 responses), transmission pairs (5 responses), and wastewater (6 responses). (Continued)

	Indirect	88%	100%	100%	17%
	Not informed	13%	0%	0%	67%
	Quantities informed [Incidence]				
	Direct	50%	0%	0%	0%
	Indirect	50%	100%	80%	67%
	Not informed	0%	0%	20%	33%
	Quantities informed [Prevalence]				
	Direct	25%	0%	0%	0%
	Indirect	38%	80%	80%	100%
	Not informed	38%	20%	20%	0%
	Quantities informed [Heterogeneity in transmission (e.g. superspreading)]				
	Indirect	50%	40%	100%	0%
	Not informed	50%	60%	0%	100%
	Quantities informed [Drivers of transmission]				
	Indirect	38%	40%	100%	0%
	Not informed	63%	60%	0%	100%
Practical considerations	**Scalability**				
	Exponentially	0%	20%	60%	0%
	independent	0%	20%	0%	83%
	Linearly	63%	60%	40%	0%
	Sub-exponentially	38%	0%	0%	17%

[1] %

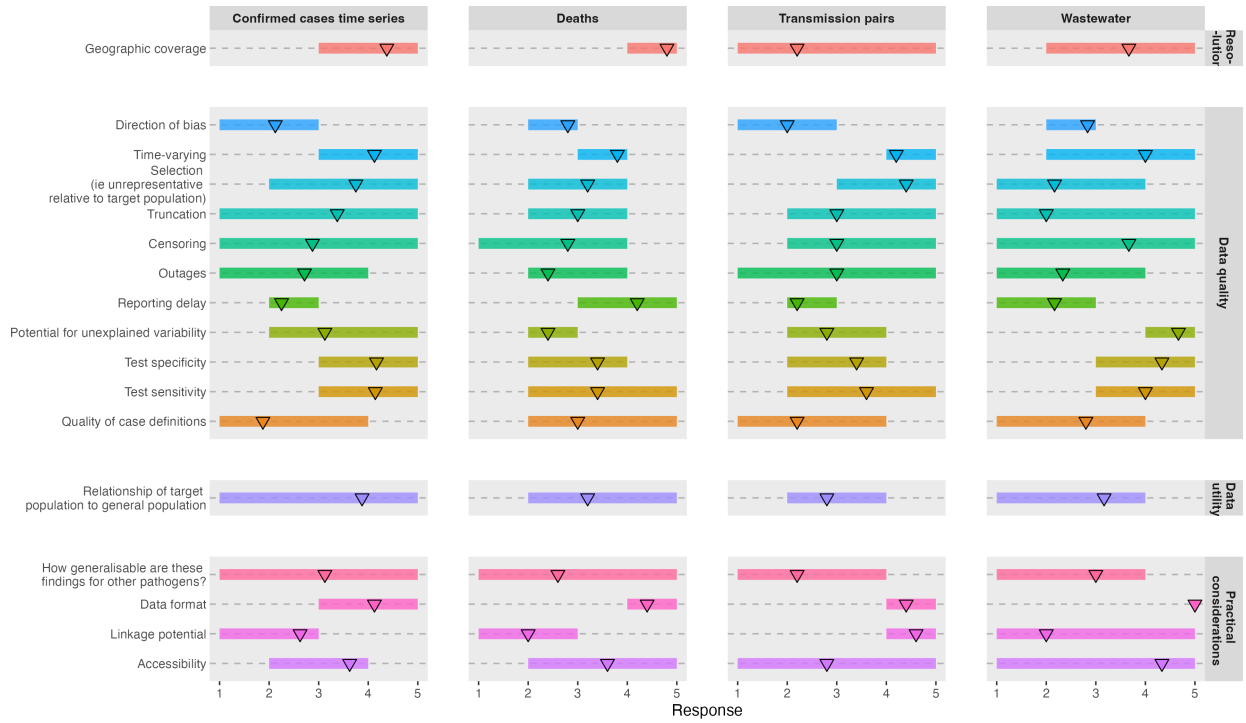

Figure 5: Selected data characteristics from the survey, showing responses related to confirmed case time series (8 responses), deaths (5 responses), transmission pairs (5 responses), and wastewater surveillance (6 responses). The y-axis represents specific data characteristics, while the x-axis indicates the response range on a 5-point scale, where 1 corresponds to a low level and 5 to a high level. The lower triangles present the average response for each data characteristic.

5.2 Case Study Model 1: Single-Source Baseline (Cases Only)

In this first part of the case study, we showcase our workflow by exploring some of the assumptions and modelling and fitting choices underlying the estimation of R_t from a single time series of case incidence.

Workflow Demonstration:

1. **Research question and target estimands:** The research question for this part of the case study (and case study models 2 and 3) is: What is the time-varying reproduction number R_t during this outbreak period?

2. **Process DAG development:** We choose to use a discrete-time model with a daily time step. This is convenient as it matches the frequency of reported data in many real-world surveillance systems. However, we note that other choices are possible, including continuous-time models or different choices of time step. Also, it is not necessary for the time step in the process model to be the same as the data reporting frequency and it is possible to explicitly model the temporal aggregation of transmission events observed with interval censoring as part of the observation model. We begin with a flexible branching process model for the number of people $N_{i,t}$ newly infected by individual i on day t:

$$N_{i,t} \sim \text{Poiss}\left(Y_i R_t g_{t-T_i}\right) \tag{1}$$

Here, T_i is the infection time of individual i, Y_i is the relative transmission rate of individual i (which may capture a combination of biological and social factors), and g_s is the relative infectiousness s days after infection normalised such that $\sum_{s=1}^{\infty} g_s = 1$. We start with this approach as it explicitly models the full transmission tree of the epidemic whilst remaining relatively parsimonious. We further

assume that the infectiousness profile g_s does not change over time, but note that we would need to revisit this assumption if there was evidence this had changed, for example due to biological changes in the pathogen, changes in contact patterns, or case isolation measures. The daily infection incidence $I_t = \sum_i N_{i,t}$ satisfies

$$I_t \sim \text{Poiss}\left(R_t \sum_i Y_i g_{t-T_i}\right) \tag{2}$$

3. **Data source selection:** Case incidence data is typically readily available with very frequent updates and relatively good timeliness (good scope and resolution), though this can depend on lab processing and reporting timelines. These factors make case data an obvious choice for trying to model time-varying transmission parameters. However, case counts are an imperfect and often biased proxy for infection incidence and are sensitive to testing patterns, which can vary over time (data quality issues) – see Table 2 and Figure 5.

4. **Observation DAG construction:** The observation model relates the observed data (here daily case incidence) with underlying latent variables (here, daily infection incidence). This requires us to make assumptions about case ascertainment (what fraction of infections are detected cases), reporting delays (time between infection and a case being detected and reported), and random noise (typically used to capture other things not explicitly specified in the model).

   There are at least $2^3 = 8$ possible observation models to consider, which either ignore or include each of these three aspects. We refer to these models as O_{000} (for the simplest observation model not modelling any of the above) to O_{111} (for the most comprehensive accounting for all three observation features). We represent the O_{111} model by the following equation for the observed variable C_t, representing the number of reported cases on day t:

$$C_t \sim \text{Poiss}\left(\sum_{s=1}^{t} M_{t-s,s}\right) \tag{3}$$

   where $M_{t,s}$ is the number of people infected on day t and reported as a case s days later, given by

$$M_{t,\cdot} \sim \text{TruncatedMultinomial}\left(I_t, p_c \boldsymbol{\alpha}\right) \tag{4}$$

   Here p_c is the probability an infection is reported as a case (i.e. case ascertainment ratio), which we assume to be constant over time, and $\boldsymbol{\alpha}$ is a vector containing the probability mass function of the distribution of time from infection to reporting. We choose to use the most comprehensive observation model O_{111} as it is generally important to account for all three factors.

5. **Refine the model DAGs:** Since we only have aggregate-level data on the number of reported cases and no individual-level data about the transmission tree, and no explicit domain knowledge to inform a prior, we recognise that individual heterogeneity in transmission rates (represented by the variable Y_i) will not be identifiable. We therefore simplify the model by assuming that $Y_i = 1$ for all individuals. With this simplification, Eq. (2) for daily infection incidence I_t reduces to

$$I_t \sim \text{Poiss}\left(R_t \sum_{s=1}^{\infty} I_{t-s} g_s\right) \tag{5}$$

   This is an example of the well-known renewal process model Fraser [2007]. This is a popular model for R_t estimation as it captures fundamental features of the transmission process (newly infected individuals at time t transmit the pathogen to other individuals some time later), without making additional assumptions, for example, about the size of the susceptible population or the strength and duration of infection-acquired immunity. This provides a simple but flexible model to which additional complexity can be added in subsequent iterations of the workflow. There are other valid assumptions we might wish to make depending on other workflow considerations. Here, we consider three variants of the renewal process model, which, conditional on previous infections, all have the same mean daily infection incidence $E(I_t)$, but assume different models for the distribution of I_t:

P_1. Deterministic model for daily infection incidence (no variance).

$$I_t = R_t \sum_{s=1}^{\infty} I_{t-s} g_s \tag{6}$$

P_2. Poisson-distributed daily infection incidence.

$$I_t \sim \text{Poiss}\left(R_t \sum_{s=1}^{\infty} I_{t-s} g_s \right) \tag{7}$$

P_3. Negative binomially distributed daily infection incidence (larger variance than Poisson).

$$I_t \sim \text{NegBin}\left(R_t \sum_{s=1}^{\infty} I_{t-s} g_s, k \right) \tag{8}$$

where k is the dispersion parameter.

Figure 6 shows the process DAGs and observation DAGs for all potential models. Here, P_1 is unlikely to be realistic as there will almost always be some source of random variability in daily infection incidence. The choice between P_2 and P_3 could depend on whether superspreading is known to be a major factor in the transmission dynamics. Since P_3 contains P_2 in the limiting case $k \to \infty$, we choose to use P_3 in this example (see Figure 7). Importantly, the negative binomial distribution in P_3 captures both potential biological overdispersion (superspreading) and unmodelled variance in the data. We could include additional sources of noise in the observation model (e.g. day-of-the-week reporting effects [Abbott et al., 2020]) if the data showed evidence for these either on first inspection or during model specification and validation.

6. **Modularise the DAGs:** Here, we do not need to do any modularisation as we are only dealing with a single data source, and the model is simple enough that we don't need to investigate the process and observation models as separate modules.

7. **Inference and implementation choices:** Implementation now requires selection of one process model, one observation model and one fitting method. The choice of fitting method depends on the analytical tractability of the combined process and observation model, computational complexity and speed of computation relative to the time available, and parameter identifiability issues (see Section 3.7). For our chosen combination of models (P_3-O_{111}), we recommend using NUTS, but only if we can approximate the discrete observation models (e.g. EpiNow2 [Abbott et al., 2020]). If approximation is not feasible or warranted, we would need to use an alternative method such as PMCMC. If we needed a simpler, more computationally efficient fitting method, we could return to the "Refine models" stage and select a simpler model. For the simplest model combination with no underreporting, delays or observation noise (P_1-O_{000}), we can calculate R_t directly for a given choice of generation time distribution by rearranging Eq. (6). This is likely to give biased and highly noisy estimates in practice and does not provide any uncertainty quantification. For the next-simplest model (P_2-O_{000}), we can calculate an analytical posterior distribution for R_t from a conjugate gamma prior, which is fast and computationally efficient. This is implemented in the widely used EpiEstim method [Cori et al., 2013].

8. **Model specification and validation:** Model specification requires a prior for R_t as well as specification of the profile of infectiousness over time g_t, the reporting probability p_c, the distribution of time from infection to reporting $\boldsymbol{\alpha}$ and (for process model P_3) the dispersion coefficient k. The simplest approach is to take these as fixed parameters, which corresponds to an infinite-strength prior. This is unlikely to ever be appropriate but may be required for computational tractability. However, a better approach in the presence of uncertainty is to define a prior over these parameters based on domain expertise. In the case of probability distributions $\mathbf{g}$ and $\boldsymbol{\alpha}$, we could use a parametric distribution assumption, e.g. a discretised gamma [Charniga et al., 2024, Park et al., 2024] or other non-negative-integer valued distribution with specified priors for the mean and variance.

Different modelling approaches have different strengths and limitations. For example, simple approaches are fast to implement, but ignore important factors in the data observation process. These may be useful for a quick initial estimate, but should be revisited as needed. In this model, R_t estimates may be insensitive to underreporting provided the reporting fraction is constant over the period of interest. Furthermore, any change in reporting fraction will not be identifiable from a case time series alone (if R_t varies over the same time scale), so it would not be sensible to attempt to estimate a time-varying reporting fraction in this example. However, reporting delays can be substantial and will affect both the magnitude and timing of R_t estimates. We can account for this effect by selecting an appropriate observation model, at the cost of needing to move to a more computationally costly fitting method. Ultimately, the most appropriate option will depend on the balance between need for timely estimates (e.g. for modelling an outbreak in real-time) and the realism of the associated assumptions. Finally, a common weakness of all the models outlined above is that they assume the profile of infectiousness over time is constant. Some approaches (e.g. some versions of EpiEstim, and EpiNow2 [Abbott et al., 2020]) extend the approach to relax this assumption.

9. **Data integration choices:** As there is only one data source in this part of the case study, we do not need to make any decisions about what and how to integrate multiple sources. In a scenario where multiple groups use different methods to produce estimates of the target estimand (in this case R_t) based on the same data, we could integrate the results by ensembling them into a combined estimate [Maishman et al., 2022, Manley et al., 2024]. In this context, it is desirable to align the assumptions made by different models, for example, about the generation time distribution or the temporal smoothness in R_t, to ensure that estimates from the different models are comparable [Brockhaus et al., 2023]. Potentially, estimates may also need to be downweighted when drawn from the same data sources.

5.3 Case Study Model 2: Two-Source Integration (Cases and Deaths)

More data sources often become available later in an outbreak. For example, early in an outbreak, surveillance data may be restricted to reported cases due to the delay from infection to development of clinically severe disease. However, once the outbreak has progressed and/or reporting systems are brought online, other events such as hospitalisation, Intensive Care Unit (ICU) admission or death notification become available. In this part of the case study, we model time-varying transmission rates using two data sources: reported cases and deaths. We illustrate the process and challenges of adding an additional data stream and explore how this can improve R_t estimation and reduce assumption dependence.

1. **Research question and target estimands:** This part of the case study targets the same estimand as in case study model 1 (R_t).

2. **Process DAG development:** We consider the same process DAG as in case study model 1 (see Figure 7), as we assume that the addition of data on deaths does not impact the transmission process.

3. **Data source selection:** Data on deaths typically have longer delays (delay from infection to death and potentially from date of death to date of registration) and are less representative of the overall population, but have lower noise (e.g. due to "day-of-the-week" reporting effects), higher ascertainment, and less dependence on testing patterns than case incidence data. Combining death data with case data can therefore potentially overcome some of the limitations present in each single data source. Based on the data survey (Table 5), we expect these data sources to have differing time-varying biases and so to be useful for unpicking the dynamics of the underlying infection time series. As noted, the major issue we need to be aware of is the potential for conflict as these data sets represent different target populations.

4. **Observation DAG construction:** Similar to the number of reported cases defined by Eq. (3), the number of deaths D_t that occur on day t depends on the time series of infections up to day t, the infection-fatality ratio p_d, and the distribution v_s of time (s) from infection to death. We represent this in the observation DAG (Figure 7) and by the following convolution equation [Bhatt et al., 2023]:

$$D_t \sim \text{Poiss}\left(p_d \sum_{s=1}^{\infty} I_{t-s} v_s\right) \tag{9}$$

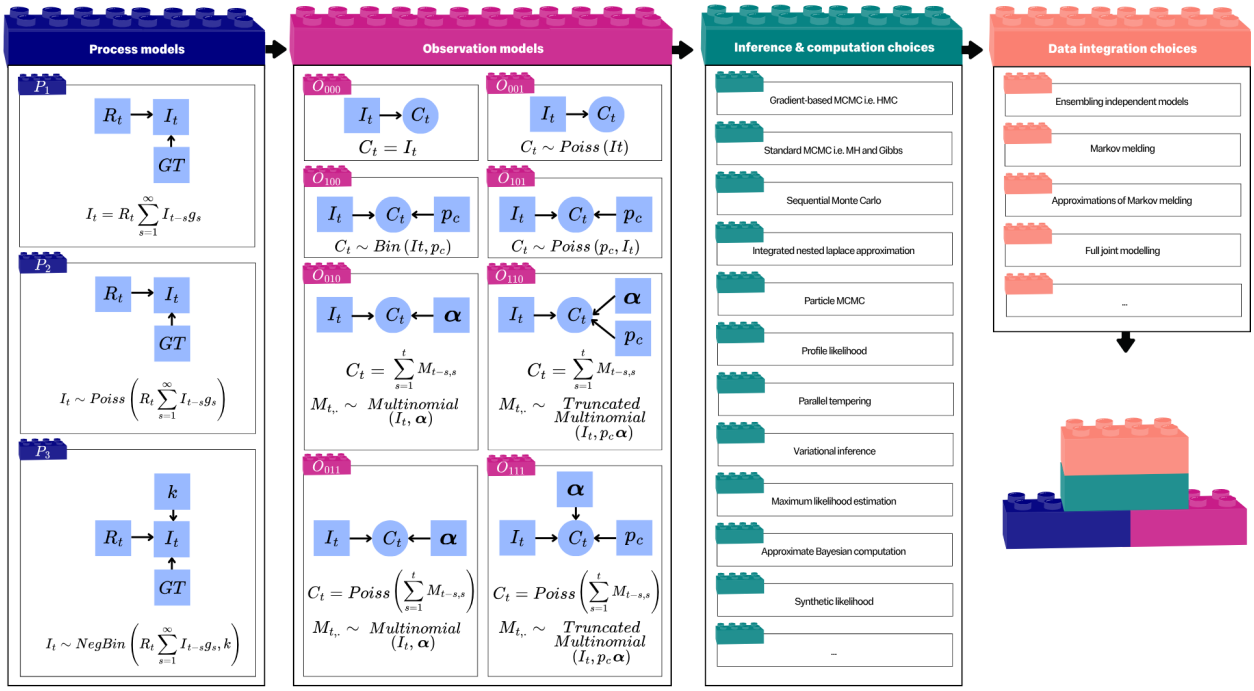

Figure 6: **Modular representation for infectious disease modelling showing the interchangeable components of process models, observation models, integration approaches, and fitting methods.** Process models include deterministic (P_1), stochastic Poisson (P_2), and negative binomial (P_3) formulations relating reproduction number (R_t) to number of new infections (I_t) with generation time (GT) distributions. P_3 incorporates overdispersion parameter k to caption extra-Poisson variation. Observation models range from direct observation (O_{000}) to complex hierarchical structures (O_{111}) that model the relationship between I_t and (observed) case incidence through underreporting, reporting delays and measurement uncertainty. The jigsaw puzzle metaphor illustrates how different modelling components can be combined flexibly, with inference method including MCMC, VI, ABC, etc. Integration approaches span from full joint modelling to ensembling independent models, enabling researchers to select optimal combinations based on data availability and research objectives.

Other models for deaths are possible, for example, using a multinomial model for the number of deaths occurring on day t due to infections on day $t - s$ ($s = 1, 2, \ldots$). However, the Poisson model is a good approximation provided $p_d \ll 1$. Eq. (9) provides a model for the time series of daily deaths D_t conditional on the time series of daily infections I_t and would be coupled with one of Eqs. (6)–(8), which describe the dynamics of infections, to produce a joint model for infections and deaths. We begin with the simplest observation model, which assumes that all deaths are recorded immediately, so the observed variable is simply $D_t^{\text{obs}} = D_t$.

5. **Refine the model DAGs:** Similar to the different choices of observation model for cases, the observation model for deaths could include underreporting, reporting delays, and random noise. The observation model described above corresponds to O_{000}. However, in revisiting the data on deaths, we discover that there is sometimes a substantial delay from date of death to date of registration [Seaman et al., 2022]. If we were conducting the analysis retrospectively, we could use model O_{000} for data by date of death (i.e. define D_t^{obs} to be the number of deaths that occurred on day t). However, for a real-time analysis, this data would be incomplete and it would be important to account for the delay. Hence, we use data by date of registration (i.e. define D_t^{obs} to be the number of deaths registered on day t) and use model O_{010}. This still assumes that all deaths are eventually reported with no additional observation noise (which is reasonable for jurisdictions with comprehensive death records), but accounts for the observation delay via the following equation for the number of deaths registered on day t:

$$D_t^{\text{obs}} = \sum_{s=0}^{\infty} Q_{t-s,s} \tag{10}$$

where $Q_{t,s}$ is the number of deaths that occurred on day t and were registered s days later, given by

$$Q_{t,\cdot} \sim \text{Multinomial}\left(D_t, \boldsymbol{\beta}\right) \tag{11}$$

where $\boldsymbol{\beta}$ is a vector containing the probability mass function for the number of days from date of death to date of registration.

6. **Modularise the DAGs:** The module for reported cases is the same as in case study model 1. Here, we add a second module for deaths (see Figure 7).

7. **Inference and implementation choices:** As each submodel and the joint model have a differentiable likelihood, we recommend NUTS for fitting. If some of the submodels had non-differentiable likelihoods, we would recommend using NUTS on the submodels where this was possible, and PMCMC on the remaining submodels and the joint model.

8. **Model specification and validation:** In addition to the parameters from case study model 1, this model requires us to specify the infection fatality ratio p_d, the distribution of time from infection to death $\mathbf{v}$ and from date of death to date to date of registration $\boldsymbol{\beta}$. Again, we recommend using appropriately informative priors based on domain knowledge where possible to allow for appropriate uncertainty.

9. **Data integration choice:** Two different approaches are possible for integration: (1) a joint model including both cases and deaths with a shared latent state R_t [Scott et al., 2020]; (2) separate models for cases and deaths, resulting in two estimates for R_t to be combined via a weighted ensemble or Markov melding. As outlined above, we recommend fitting separate models to the two time series initially, to understand their behaviour and reveal whether they lead to consistent or conflicting estimates of R_t [Sherratt et al., 2021]. Inconsistent results emerging from the two data sources could prompt us to refine the model. For example, if R_t estimates show similar trends but shifted in time, we would need to revisit assumptions about delays (return to model specification and validation (Section 3.9)). If R_t estimates are largely consistent, but the case-based estimate shows a transient increase in R_t that does not occur in the deaths-based estimate, an increase in case ascertainment might be an explanation (return to refine the model DAGs (Section 3.5) to include time-varying ascertainment). In case study model 1, using data on cases alone, the case ascertainment ratio was non-identifiable and had to be assumed. Including data on deaths allows estimation of trends in case ascertainment, provided the

infection-fatality ratio can be assumed constant over time. If the infection-fatality ratio additionally has a relatively informed prior distribution, the absolute value of case ascertainment can be estimated. Both approaches require assumptions about the distribution of time from infection to death, demonstrating how additional data sources can shift rather than eliminate the need for modelling assumptions. As both of these models are relatively simple Markov melding is likely not needed when fitting the joint model.

5.4 Case Study Model 3: Three-Source Integration (Cases, Deaths and Wastewater)

Wastewater-based epidemiology has emerged as a complementary data source to traditional disease surveillance methods [Keshaviah et al., 2023]. Wastewater surveillance has been used for a range of viral pathogens including norovirus [Zheng et al., 2024], poliovirus [Whitehouse et al., 2024], and influenza viruses [Zheng et al., 2023]. During the Covid-19 pandemic, wastewater surveillance supported outbreak detection [Hewitt et al., 2022], situational awareness and forecasting [Varkila et al., 2023, Jin et al., 2024], and multi-strain modelling [Dreifuss et al., 2025].

1. **Research question and target estimands:** In this part of the case study, we include measurements of viral Ribonucleic Acid (RNA)/Deoxyribonucleic Acid (DNA) concentration in wastewater samples as a third data source alongside reported cases and deaths. We explore ways of handling conflicting signals between data sources and incorporating complex observation processes. The target estimand is R_t, as in case study models 1 and 2.

2. **Process DAG development:** We start with the same process DAG as in case study models 1 and 2 (see Figure 7).

3. **Data source selection:** Wastewater data offers population-level information (broad scope) independent of testing biases and changes in testing rates, with moderate timeliness, but requires environmental expertise and is subject to high levels of unexplained variability. It could also be sensitive to unknown changes in the average shedding rate over time and, often, has relatively poor resolution (limited spatial resolution and no information about age or other demographic variables).

4. **Observation DAG construction:** Suppose that the average viral RNA/DNA load W_t in wastewater on day t, typically quantified in genome copies per person per day, depends on the time series of infections up to day t and the average amount of viral RNA/DNA that an individual sheds into wastewater s days after being infected ($s = 1, 2, \ldots$). We represent this in the process DAG in Figure 7 and the following equation for W_t:

$$W_t = \frac{c}{N_{\text{pop}}} \sum_{s=1}^{\infty} I_{t-s} u_s \tag{12}$$

where $c > 0$ is a constant representing the average total amount of viral RNA/DNA that an infected individual sheds over the course of their infection, N_{pop} is the total population size, and u_s is the average shedding rate s days after infection, normalised such that $\sum_{s=1}^{\infty} u_s = 1$. We would couple Eq. (12) with one of Eqs. (6)–(8) for infections.

Wastewater samples are typically collected at some cadence from one or more sampling sites, and the concentration of viral RNA/DNA in the samples is quantified via Polymerase Chain Reaction (PCR) testing. The existence of multiple sites with different catchment populations and non-contemporaneous sampling frequencies complicates the interpretation of quantitative wastewater data, which can be modelled at varying levels of complexity. Suppose that measurements of the wastewater concentration $W_{j,t}^{\text{obs}}$ are taken from sampling sites $j = 1, \ldots, J$ on some subset of days t. Similarly to [Watson et al., 2024], we assume that these observations are conditionally independent gamma random variables with the same mean W_t and variance bW_t^2/N_j, where N_j is the population size in the catchment for sampling site j and b is a variance parameter:

$$W_{j,t}^{\text{obs}} \sim \Gamma\left(\text{mean} = W_t, \text{var} = \frac{bW_t^2}{N_j}\right) \tag{13}$$

5. **Refine the model DAGs:** The formulation above assumes that the different catchment areas have identical epidemic dynamics and the only difference between catchments is that those with smaller populations have higher measurement variance. This is a reasonable starting assumption if the outbreak is evenly distributed geographically. However, it assumes that the variance in observed data is primarily due to environmental variation and individual heterogeneity, and hence scales inversely with population size. This ignores the contribution to variability of noise in PCR measurements, which depends on the concentration and therefore on sewage system-specific factors like total flow volume [Lison et al., 2024b]. The relationship between mean and variance in wastewater measurements is likely to be much less direct than for case count data. Furthermore, different sampling sites revealing different temporal trends would indicate the need to refine the model to account for different epidemic dynamics in different regions, assuming that the differing trends aren't driven by some unknown site-specific observation process differences. This finding could require revisiting the data source selection (Section 3.3) for an assessment of whether the catchment areas for wastewater sampling are aligned with the geographical stratification (if any) available for reported cases and deaths. More complex observation models could also incorporate other factors such as individual-level and site-level variation, catchment population dynamics, spatial heterogeneity, different sampling methods, and environmental degradation. Adding these to the model would need to be supported by additional data and by the results of the model specification and validation step (e.g. if the model is fitting poorly to individual sites, then site-level variation in observation processes may be needed).

6. **Modularise the DAGs:** The modules for cases and deaths are the same as in case study models 1 and 2, respectively. The module for wastewater data follows the same pattern, with wastewater representing a third process downstream from the daily infection incidence.

7. **Inference and implementation choices:** As before, we recommend using a gradient-based sampling method such as NUTS where possible. For real-time estimation during an outbreak, this may be impractical, in which case, we might need to consider a method such as SMC that supports real-time learning (e.g. sequential data assimilation), provided that model parameters are reasonably well known (informed by expert opinion or previous epidemics).

8. **Model specification and validation:** In addition to the parameters from the case study models 1 and 2, we need to specify the average shedding rate c, the total population size N_{pop}, the population in each sampled catchment N_j, and the shedding profile over time u_t. Wastewater data enables R_t estimation independent of shifts in testing patterns, but requires assumptions about the shedding, environmental and sampling processes. Similar to the unknown case ascertainment ratio in case study model 1, estimates of R_t will, at sufficiently large incidences, be insensitive to the value of c provided it does not change over time [Dreifuss et al., 2025]. However, changes in the average shedding rate due to, for example, pathogenic evolution or changes in population immunity, would invalidate this simple model and require a more nuanced approach.

9. **Data integration choice:** As in case study model 2, we recommend a modular approach with sequential consistency assessment to detect and resolve data source conflicts. For example, a time lag in estimates of R_t from wastewater data alone relative to estimates of R_t from cases alone could suggest misspecification of the shedding rate distribution u_s or of the delay from infection to case report. It is also useful to compare estimates of latent states between separate models. For example, if estimates of I_t from the cases model and the wastewater model are initially consistent but start to diverge after a certain time, this could indicate that either the case ascertainment ratio p_c or the average shedding rate c has changed. An assessment of which of these explanations is more likely might be informed by the context. For example, a decrease in testing would suggest a drop in case ascertainment is likely, while a vaccine rollout might suggest that a drop in average shedding is likely due to increased immunity.

   Once we have identified and resolved any conflicts, we could ensemble independent estimates of R_t from these models into a combined estimate, or fit a joint model with a shared R_t parameter. Fitting a joint model has the advantage that information from each data source is pooled. This enables inference of the joint distribution of the parameters (p_c, p_d and c) relating the observed variables to the latent variable I_t, which would not be provided by separate models. If the steps above revealed some likely

time-dependence in these quantities, we might refine the model, for example by allowing the case ascertainment ratio p_c or the shedding rate c to be time-dependent latent states as opposed to fixed parameters [Watson et al., 2024]. For this part of the case study, especially as we have already fit case study models 1 and 2, Markov melding is likely to be useful to speed up the fitting process (see Section 3.10).

5.5 Case Study Model 4: Individual-Level Data (Cases and Transmission Pairs)

In this final part of the case study, we illustrate how we can combine different types of data (time series count data and individual-level data from contact tracing records) to enable estimation of additional quantities.

1. **Research question and target estimands:** The research question for this part of the case study is again what is the time-varying reproduction number R_t with the addition of how much heterogeneity is there in individual transmission rates? The latter is captured by the additional estimand k, representing the dispersion in the distribution of the number of secondary cases per index case [Lloyd-Smith et al., 2005].

2. **Process DAG development:** Now that we have some information on individual-level variables, as opposed to just aggregate daily counts, we can return to our original process model (Section 3.2) for the full epidemic transmission tree from case study model 1, defined by Eq. (1) (see Figure 7). Building on the model of Lloyd-Smith et al. [2005], if $Y \sim \Gamma(1, k)$ for some dispersion parameter k, then the number of secondary cases caused by an individual who was infected at a given time $T_i = t$ is

$$Z_i \mid (T_i = t) \sim \text{NegBin}\left(\tilde{R}_t, k\right) \tag{14}$$

   where $\tilde{R}_t = \sum_s R_{t+s} g_s$ is the effective reproduction number averaged over the infectious period of an individual who was infected on day t.

3. **Data source selection:** Case study models 1–3 only consider aggregate-level data on daily counts or wastewater measurements. This type of data typically only allows estimation of average transmission, i.e. R_t. In contrast, contact tracing records identify transmission pairs, which contain information about heterogeneity in transmission patterns, contingent on contact tracing system quality. The smaller the dispersion parameter k, the more variance would be expected in the distribution of the number of secondary infections per index case.

4. **Observation DAG construction:** Suppose that, for some subset of reported cases i, the number of observed secondary cases Z_i^{obs} linked to index case i is recorded. For simplicity, we assume that the index cases for whom these data are available are a randomly chosen subset of all reported cases, and each secondary infection independently has a fixed probability p_l of being linked to the index case. We assume that all linked secondary cases are associated with the correct index case (i.e. we ignore any misclassification and uncertainty about who infected whom).

   Then we can calculate the probability that an index case reported on day τ_i has Z_i^{obs} linked secondary cases by conditioning on the unobserved time of infection T_i:

$$P(Z_i^{\text{obs}} = z \mid \tau_i = t) = \sum_{s=1}^{t} F_{NB}(z; p_l \tilde{R}_s, k) P(T_i = s | \tau_i = t) \tag{15}$$

   where $F_{NB}(.; \mu, k)$ is the probability mass function for a negative binomial distribution with mean μ and dispersion k, and τ_i is the reporting time for case i. We can calculate the conditional probability on the right-hand side of this equation via Bayes' theorem to give

$$P(Z_i^{\text{obs}} = z \mid \tau_i = t) = \frac{\sum_{s=1}^{t} F_{NB}(z; p_l \tilde{R}_s, k) \alpha_{t-s} I_s}{\sum_{s=1}^{t} \alpha_{t-s} I_s} \tag{16}$$

   where α_t is the distribution of time from infection to reporting.

5. **Refine the model DAGs:** We start by assuming that the relative infectiousness profile over time g_t and the distribution of time from infection to reporting α_t are the same for all individuals. This assumption is a reasonable starting point, but it may be important to relax it in some situations, for example, to model the impact of quarantine and isolation measures on individuals identified by contact tracing.

6. **Modularise DAGs:** We modularise the observation DAG into a sub-DAG for daily case incidence C_t and a sub-DAG for observed secondary case counts Z_i for each index case i (see Figure 7).

7. **Inference and implementation choices:** We again suggest using NUTS with data augmentation (latent variables) for unobserved transmissions, chosen for efficient handling of latent variables. Alternatively, if we needed real-time data assimilation, we could use a SMC method.

8. **Model specification and validation:** Individual-level data enables direct overdispersion estimation without distributional assumptions but requires assumptions about the completeness of contact tracing data. This part of the case study shows how data granularity can change model structure and inference requirements. Note that the observed variable Z_i^{obs} in Eq. (16) may not be available in real-time as information about linked secondary cases takes time to collect. If the model was found to be impractical for these reasons, it may be necessary to simplify the model by making some additional assumptions. For example, if there was a period of time in which R_t was estimated to be relatively steady (an example of a well-specified process model for this setting), we could simplify the model by assuming a constant reproduction number and estimating k from contact tracing data relating to that period.

9. **Data integration choices:** Since the two datasets in this part of the case study each inform different quantities (case counts inform R_t, data on transmission pairs inform k), it would not make sense to fit separate models in this example, and a joint model is needed.

6 Discussion

We have presented a suggested workflow for infectious disease modelling that addresses domain-specific challenges, extends established Bayesian workflow principles [Green et al., 2003, Gelman et al., 2020], and has a particular focus on integrating multiple data sources. This workflow aims to serve as guidance for modellers when they are developing infectious disease models and as a resource for readers evaluating modelling papers, whether they work in academic research, public health agencies, or policy contexts that rely on model outputs. We used a case study to demonstrate this workflow using progressively more data sources for reproduction number estimation, showing how workflow practices can enable principled model development from single-source models to complex multi-stream integrations. Each progression reinforces that additional data sources often alter both process and observation DAGs and can require new integration, inference and implementation decisions. Throughout, we highlight that considering each submodel in isolation is a useful starting point for model development and validation.

The structure of our workflow addresses a critical gap in infectious disease modelling practice, where ad hoc approaches have often led to inadequate validation and limited transparency in methodological choices. A key strength of our framework is the explicit separation of process and observation DAGs in the spirit of state-space models, which provides clarity for model development and validation by recognising that epidemiological processes and observation mechanisms draw from different domains. A significant limitation is that we have not implemented the workflow in practice, providing only a schematic case study that illustrates conceptual progression rather than actual code or examples. Despite this lack of implementation, the framework's modular structure enables users to adopt specific components like our suggested data review process or DAG-based planning independently of other steps in their workflow. In some sense, this lack of a specific tool example is a strength as it makes it easier to view this workflow as a guide, which can then be combined with tools of their choice. Our structured data characterisation checklist provides an important benefit by establishing a common language for discussing data source trade-offs within modelling teams. Yet this checklist is necessarily context-dependent, with criteria and their relative importance varying across pathogens, settings, and research questions. This means that we cannot give an overview of all data sources,

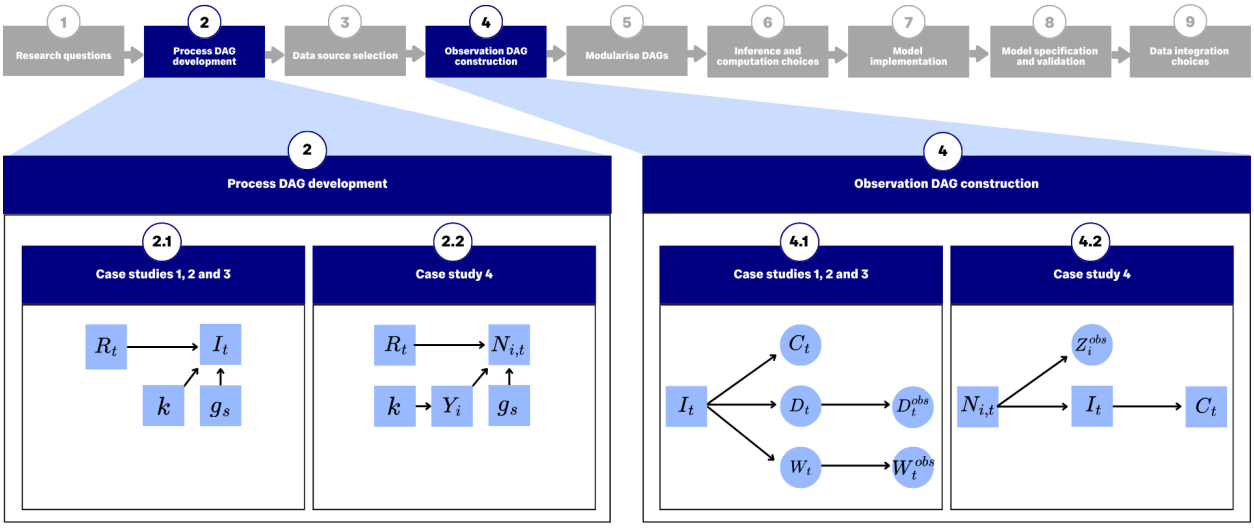

Figure 7: **A Schematic for Constructing the DAGs.** This schematic shows how data source selection and model choices can lead to different construction of the observation DAG from the underlying process DAG for the different case study models. For simplicity, feedback loops are omitted in this schematic. For a complete representation of feedback mechanisms, see Figure 2.

or even many common ones. A limitation of our more specific implementation recommendations, such as which inference algorithm or PPL to use, is that they will inevitably become outdated. However, the general principles underlying these recommendations, such as using gradient-based methods where applicable or implementing models in probabilistic programming frameworks, should remain useful, as should our broader recommendations for using a structured modelling workflow. A final strength of our proposed framework is its applicability across diverse contexts, from early outbreak analysis to routine surveillance, offering a consistent structure regardless of data availability and research questions.

Our workflow extends Gelman et al.'s general Bayesian workflow framework [Gelman et al., 2020], which itself builds on wider statistical modelling workflows [Box, 1979, 1980, Green et al., 2003], by addressing workflow adoption challenges specific to infectious disease modelling. Gelman et al. give a workflow for model validation and criticism within a single modelling cycle with extensive guidance and practical examples. We embed their modelling loop within a larger workflow that aims to address domain-specific challenges: data source characterisation, integration strategies, and workflow iteration as outbreaks evolve. As noted, whilst we don't give fully implemented examples, we do provide additional domain-specific examples for surveillance data quality, reporting biases, and epidemiological process representation not covered in the general statistical literature.

Recent work has made progress towards structured infectious disease modelling workflows. For example, Bouman et al. [2024] systematically compared modelling choices for Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) transmission and provided an open-source implementation. However, they do not provide structured workflow guidance and key workflow elements such as prior predictive checks, systematic model criticism, and clear decision frameworks are absent. Their work highlights the gap between implementing models and following robust and reproducible workflow practices. In contrast, our framework provides explicit structured guidance through each workflow stage, with a case study demonstrating our decision process itself. Grinsztajn et al. [2021] provides systematic Bayesian workflow guidance for infectious disease transmission models in Stan, demonstrating prior predictive checks, inference criticism, and posterior predictive checks. Their workflow successfully integrates multiple data sources, including deaths, hospitalisations, and seroprevalence. However, their primary focus is on Stan implementation, and they mostly focus on implementing the Gelman et al. workflow without the infectious disease modelling modifications we have proposed.

Broader efforts have sought to standardise infectious disease modelling practice through consensus principles. Behrend et al. [2020] developed five key principles emphasising stakeholder engagement, reproducibility, data documentation, uncertainty communication, and testable model outcomes. Our workflow complements these principles by providing specific methodological guidance for technical decisions, particularly multi-source data integration. Our structured approach, reporting recommendations and focus on modularity reflect their recommendations.

Future progress requires infrastructure that makes it easier to implement methodological best practices. Whilst this paper provides a schematic workflow, practical tutorials demonstrating actual implementation would make it easier to follow and likely identify weaknesses in our recommendations. However, such tutorials and tools are likely to require advances in software composability, where self-contained components can be combined to build complex models, to make them feasible. This would enable domain experts to contribute specialised components without rebuilding core functionality and make all steps of our proposed workflow easier to implement. Improved tooling for Markov melding [Goudie et al., 2019] would facilitate modular integration, which would allow teams to more easily combine submodels without full joint specification or specialist knowledge. Further methodological work on cut posteriors [Liu and Goudie, 2025] is needed before easy-to-use tools to implement them can be developed. Computational scalability remains an area where more work is needed, both for formal conflict detection [Yang et al., 2025] and particularly for joint analysis of individual-level data with population surveillance. Recent advances in inference algorithms, including gradient-free methods, adaptive sampling techniques, neural approaches for likelihood-free inference, and GPU-accelerated approaches, show promise for infectious disease modelling. However, more work is needed to evaluate existing advances and to develop new methods for scalable joint analysis of individual-level and population surveillance data, and methods that meet the real-time computational constraints of outbreak response.

Our proposed workflow aims to address the slow adoption of rigorous practices that would benefit infectious disease modelling, by providing structured guidance for model development, validation, and multi-

source integration that can adapt to evolving outbreak contexts. The modular approach we advocate enables adding complexity incrementally whilst maintaining transparency about modelling assumptions and integration choices. We recommend that the infectious disease modelling community adopt a structured workflow, such as the one presented here, for model development, evaluation, and reporting. This is particularly important during rapidly evolving outbreaks where research questions can change and new data sources emerge over time.

List of Abbreviations

DAG Directed Acyclic Graph

MCMC Markov Chain Monte Carlo

HMC Hamiltonian Monte Carlo

NUTS No-U-Turn Sampler

VI Variational Inference

SMC Sequential Monte Carlo

SMC^2 Sequential Monte Carlo Squared

POMP Partially Observed Markov Process

PMCMC Particle Markov Chain Monte Carlo

INLA Integrated Nested Laplace Approximation

MLE Maximum Likelihood Estimation

ABC Approximate Bayesian Computation

SL Synthetic Likelihood

BFMI Bayesian Fraction of Missing Information

ELBO Evidence lower bound

ESS Effective Sample Size

PPL Probabilistic programming language

LOO-CV leave-one-out cross-validation

WAIC Watanabe-Akaike information criterion

CRPS continuous ranked probability score

GPU Graphics Processing Unit

CPU Central Processing Unit

RNA Ribonucleic Acid

DNA Deoxyribonucleic Acid

PCR Polymerase Chain Reaction

ICU Intensive Care Unit

SARS-CoV-2 Severe acute respiratory syndrome coronavirus 2

7 Acknowledgements

We thank the organisers and participants of the workshop "Analysis and modelling for the design of future epidemic surveillance systems" (CIRM, Marseille, 28 April–2 May 2025) for valuable discussions and feedback that helped shape this work. All workshop participants provided useful discussion and feedback. We thank Poppy the dog for making sure to ask the important questions.

References

Sam Abbott and Sebastian Funk. Estimating epidemiological quantities from repeated cross-sectional prevalence measurements. *MedRxiv*, pages 2022–03, 2022.

Sam Abbott, Joel Hellewell, Robin N Thompson, Katharine Sherratt, Hamish P Gibbs, Nikos I Bosse, James D Munday, Sophie Meakin, Emma L Doughty, June Young Chun, et al. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. *Wellcome Open Research*, 5(112):112, 2020.

Sam Abbott, CMMID COVID-19 Working Group, Adam J Kucharski, and Sebastian Funk. Estimating the increase in reproduction number associated with the Delta variant using local area dynamics in England. *medRxiv*, pages 2021–11, 2021.

Oriol Abril-Pla, Virgile Andreani, Colin Carroll, Larry Dong, Christopher J Fonnesbeck, Maxim Kochurov, Ravin Kumar, Junpeng Lao, Christian C Luhmann, Osvaldo A Martin, et al. PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, 9:e1516, 2023.

Philipp Andelfinger. Towards differentiable agent-based simulation. *ACM Transactions on Modeling and Computer Simulation*, 32(4):1–26, 2023.

Jair Andrade and Jim Duggan. An evaluation of Hamiltonian Monte Carlo performance to calibrate age-structured compartmental SEIR models to incidence data. *Epidemics*, 33:100415, 2020.

Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. Particle Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(3):269–342, 2010.

Fernando Baltazar-Larios, Francisco Delgado-Vences, and Saul Diaz-Infante. Maximum likelihood estimation for a stochastic SEIR system with a COVID-19 application. *International Journal of Computer Mathematics*, 101(12):1356–1378, 2024.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153):1–43, 2018.

Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

Matthew R Behrend, María-Gloria Basáñez, Jonathan ID Hamley, Travis C Porco, Wilma A Stolk, Martin Walker, Sake J de Vlas, and NTD Modelling Consortium. Modelling for policy: the five principles of the Neglected Tropical Diseases Modelling Consortium. *PLoS neglected tropical diseases*, 14(4):e0008033, 2020.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Sangeeta Bhatia, Natsuko Imai, Oliver J Watson, Auss Abbood, Philip Abdelmalik, Thijs Cornelissen, Stéphane Ghozzi, Britta Lassmann, Radhika Nagesh, Manon L Ragonnet-Cronin, et al. Lessons from COVID-19 for rescalable data collection. *The Lancet Infectious Diseases*, 23(9):e383–e388, 2023.

Samir Bhatt, Neil Ferguson, Seth Flaxman, Axel Gandy, Swapnil Mishra, and James A Scott. Semi-mechanistic Bayesian modelling of COVID-19 with renewal processes. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(4):601–615, 2023.

Paul J Birrell, Daniela De Angelis, and Anne M Presanis. Evidence synthesis for stochastic epidemic models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 33(1):34, 2018.

Paul J Birrell, Lorenz Wernisch, Brian DM Tom, Leonhard Held, Gareth O Roberts, Richard G Pebody, and Daniela De Angelis. Efficient real-time monitoring of an emerging influenza pandemic: How feasible? *The annals of applied statistics*, 14(1):74, 2020.

Paul J Birrell, Joshua Blake, Joel Kandiah, Angelos Alexopoulos, Edwin van Leeuwen, Koen B Pouwels, Sanmitra Ghosh, Colin Starr, Ann Sarah Walker, Thomas A House, et al. Real-time modelling of the SARS-CoV-2 pandemic in England 2020–2023: a challenging data integration. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnaf030, 2025.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John wiley & sons, 2021.

Nikos I Bosse, Sam Abbott, Anne Cori, Edwin Van Leeuwen, Johannes Bracher, and Sebastian Funk. Scoring epidemiological forecasts on transformed scales. *PLoS computational biology*, 19(8):e1011393, 2023.

Nawaf Bou-Rabee, Bob Carpenter, Tore Selland Kleppe, and Sifan Liu. The Within-Orbit Adaptive Leapfrog No-U-Turn Sampler. *arXiv preprint arXiv:2506.18746*, 2025a.

Nawaf Bou-Rabee, Bob Carpenter, Sifan Liu, and Stefan Oberdörster. The no-underrun sampler: A locally-adaptive, gradient-free MCMC method. *arXiv preprint arXiv:2501.18548*, 2025b.

Judith A Bouman, Anthony Hauser, Simon L Grimm, Martin Wohlfender, Samir Bhatt, Elizaveta Semenova, Andrew Gelman, Christian L Althaus, and Julien Riou. Bayesian workflow for time-varying transmission in stratified compartmental infectious disease transmission models. *PLOS Computational Biology*, 20(4): e1011575, 2024. doi: 10.1371/journal.pcbi.1011575. URL `https://doi.org/10.1371/journal.pcbi.1011575`.

Lampros Bouranis, Nikolaos Demiris, Konstantinos Kalogeropoulos, and Ioannis Ntzoufras. Bayesian analysis of diffusion-driven multi-type epidemic models with application to covid-19. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnaf130, 2025.

George E. P. Box. Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General)*, 143(4):383–430, 1980. doi: 10.2307/2982063.

George EP Box. Robustness in the strategy of scientific model building. In *Robustness in statistics*, pages 201–236. Elsevier, 1979.

Elisabeth K Brockhaus, Daniel Wolffram, Tanja Stadler, Michael Osthege, Tanmay Mitra, Jonas M Littek, Ekaterina Krymova, Anna J Klesen, Jana S Huisman, Stefan Heyder, Matthias an der Heiden, Sebastian Funk, Sam Abbott, and Johannes Bracher. Why are different estimates of the effective reproductive number so different? A case study on COVID-19 in Germany. *PLOS Computational Biology*, 19(11): e1011653, 2023.

Alexander Buchholz, Nicolas Chopin, and Pierre E Jacob. Adaptive tuning of Hamiltonian Monte Carlo within Sequential Monte Carlo. *Bayesian Analysis*, 16(3):745–771, 2021.

Chris U Carmona and Geoff K Nicholls. Scalable semi-modular inference with variational meta-posteriors. *arXiv preprint arXiv:2204.00296*, 2022.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 2017.

Kelly Charniga, Sang Woo Park, Andrei R Akhmetzhanov, Anne Cori, Jonathan Dushoff, Sebastian Funk, Katelyn M Gostic, Natalie M Linton, Adrian Lison, Christopher E Overton, et al. Best practices for estimating and reporting epidemiological delay distributions of infectious diseases. *PLoS Computational Biology*, 20(10):e1012520, 2024.

Anastasia Chatzilena, Edwin van Leeuwen, Oliver Ratmann, Marc Baguelin, and Nikolaos Demiris. Contemporary statistical inference for infectious disease models using Stan. *Epidemics*, 29:100367, 2019.

Nicolas Chopin, Pierre E Jacob, and Omiros Papaspiliopoulos. SMC^2: an efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(3):397–426, 2013.

C. Cobelli and J. J. DiStefano. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 239(1):R7–R24, 1980. doi: 10.1152/ajpregu.1980.239.1.R7. PMID: 7396041.

Alice Corbella. Statistical inference in stochastic/deterministic epidemic models to jointly estimate transmission and severity, 2019. URL `https://doi.org/10.17863/CAM.41539`.

Alice Corbella, Anne M Presanis, Paul J Birrell, and Daniela De Angelis. Inferring Epidemics from Multiple Dependent Data via Pseudo-Marginal Methods. *Annals of Applied Statistics, in press*, 2025a. URL `https://arxiv.org/abs/2204.08901`.

Alice Corbella, Anne M Presanis, Paul J Birrell, and Daniela De Angelis. Inferring epidemics from multiple dependent data via pseudo-marginal methods. *Annals of Applied Statistics*, 19(4):3221–3243, 2025b.

Anne Cori, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez. A new framework and software to estimate time-varying reproduction numbers during epidemics. *American journal of epidemiology*, 178(9): 1505–1512, 2013.

Kimberly A Dautel, Ephraim Agyingi, and Pras Pathmanathan. Validation framework for epidemiological models with application to COVID-19 models. *PLOS Computational Biology*, 19(3):e1010968, 2023.

Daniela De Angelis and Anne M Presanis. Analysing Multiple Epidemic Data Sources. *arXiv preprint arXiv:1808.04312*, 2018. URL `https://arxiv.org/abs/1808.04312`.

Daniela De Angelis, Anne M Presanis, Paul J Birrell, Gianpaolo Scalia Tomba, and Thomas House. Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*, 10:83–87, 2015.

Perry de Valpine, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik. Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017.

Lee Devlin, Matthew Carter, Paul Horridge, Peter L Green, and Simon Maskell. The no-u-turn sampler as a proposal distribution in a sequential Monte Carlo sampler without accept/reject. *IEEE Signal Processing Letters*, 31:1089–1093, 2024.

Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential Monte Carlo methods. pages 3–14, 2001.

David Dreifuss, Jana S Huisman, Johannes C Rusch, Lea Caduff, Pravin Ganesanandamoorthy, Alexander J Devaux, Charles Gan, Tanja Stadler, Tamar Kohn, Christoph Ort, et al. Estimated transmission dynamics of SARS-CoV-2 variants from wastewater are unbiased and robust to differential shedding. *Nature Communications*, 16(1):7456, 2025.

Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

Akira Endo, Edwin Van Leeuwen, and Marc Baguelin. Introduction to particle markov-chain monte carlo for disease dynamics modellers. *Epidemics*, 29:100363, 2019.

Richard G FitzJohn, Edward S Knock, Lilith K Whittles, Pablo N Perez-Guzman, Sangeeta Bhatia, Fernando Guntoro, Oliver J Watson, Charles Whittaker, Neil M Ferguson, Anne Cori, et al. Reproducible parallel inference and simulation of stochastic state space models using odin, dust, and mcstate. *Wellcome open research*, 5:288, 2021.

Tor Erlend Fjelde, Kai Xu, David Widmann, Mohamed Tarek, Cameron Pfiffer, Martin Trapp, Seth D Axen, Xianda Sun, Markus Hauru, Penelope Yong, et al. Turing. jl: a general-purpose probabilistic programming language. *ACM Transactions on Probabilistic Machine Learning*, 2025.

Christophe Fraser. Estimating individual and household reproduction numbers in an emerging epidemic. *PloS One*, 2(8):e758, 2007.

Sebastian Funk and Aaron A King. Choices and trade-offs in inference with infectious disease models. *Epidemics*, 30:100383, 2020.

Martyn Fyles, Jonathon Mellor, Robert Paton, Christopher E Overton, Alexander M Phillips, Alex Glaser, and Thomas Ward. The incidence and prevalence of SARS-CoV-2 in the UK population from the UKHSA winter COVID infection study. *medRxiv*, pages 2024–10, 2024.

Hong Ge, Kai Xu, and Zoubin Ghahramani. Turing: a language for flexible probabilistic inference. In *International conference on artificial intelligence and statistics*, pages 1682–1690. PMLR, 2018.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020. URL `https://arxiv.org/abs/2011.01808`.

Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Robert JB Goudie, Anne M Presanis, David Lunn, Daniela De Angelis, and Lorenz Wernisch. Joining and splitting models with Markov melding. *Bayesian analysis*, 14(1):81, 2019.

Peter J Green, Nils Lid Hjort, and Sylvia Richardson, editors. *Highly structured stochastic systems*, volume 27. Oxford University Press, 2003.

Léo Grinsztajn, Elizaveta Semenova, Charles C Margossian, and Julien Riou. Bayesian workflow for disease transmission modeling in Stan. *Statistics in medicine*, 40(27):6209–6234, 2021.

W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97–109, 1970. doi: 10.1093/biomet/57.1.97.

Anna Heath, Natalia Kunst, and Christopher Jackson. *Value of information for healthcare decision-making*. CRC Press, 2024.

Soren Henriksen, Adrian Wills, Thomas B Schön, and Brett Ninness. Parallel implementation of particle MCMC methods on a GPU. *IFAC Proceedings Volumes*, 45(16):1143–1148, 2012.

Joanne Hewitt, Sam Trowsdale, Bridget A Armstrong, Joanne R Chapman, Kirsten M Carter, Dawn M Croucher, Cassandra R Trent, Rosemary E Sim, and Brent J Gilpin. Sensitivity of wastewater-based epidemiology for detection of SARS-CoV-2 RNA in a low prevalence setting. *Water Research*, 211:118032, 2022.

Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401, 1999.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Matthew D Hoffman, Andrew Gelman, et al. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Antti Honkela. Computational Statistics I, 2020.

Christopher Jackson, Anne Presanis, Stefano Conti, and Daniela De Angelis. Value of information: Sensitivity analysis and research design in Bayesian evidence synthesis. *Journal of the American Statistical Association*, 2019.

Dan Jackson, Richard Riley, and Ian R White. Multivariate meta-analysis: potential and promise. *Statistics in Medicine*, 30(20):2481–2498, 2011.

Shihui Jin, Martin Tay, Lee Ching Ng, Judith Chui Ching Wong, and Alex R Cook. Combining wastewater surveillance and case data in estimating the time-varying effective reproduction number. *Science of the Total Environment*, 928:172469, 2024.

Noa Kallioinen, Topi Paananen, Paul-Christian Bürkner, and Aki Vehtari. Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *Statistics and Computing*, 34:57, 2024. doi: 10.1007/s11222-023-10366-5.

Aparna Keshaviah, Megan B Diamond, Matthew J Wade, Samuel V Scarpino, Warish Ahmed, Fabian Amman, Olusola Aruna, Andrei Badilla-Aguilar, Itay Bar-Or, Andreas Bergthaler, et al. Wastewater monitoring can anchor global disease surveillance systems. *Lancet Global Health*, 11(6):e976–e981, 2023.

Aaron A King, Dao Nguyen, and Edward L Ionides. Statistical inference for partially observed Markov processes via the R package pomp. *Journal of Statistical Software*, 69:1–43, 2016.

Edward S Knock, Lilith K Whittles, John A Lees, et al. Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. *Science Translational Medicine*, 13(589), 2021. doi: 10.1126/scitranslmed.abg4262.

Sourabh Kulkarni, Mario Michael Krell, Seth Nabarro, and Csaba Andras Moritz. Hardware-accelerated simulation-based inference of stochastic epidemiology models for COVID-19. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 18(2):1–24, 2022.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Phenyo E Lekone and Bärbel F Finkenstädt. Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177, 2006.

Adrian Lison, Sam Abbott, Jana Huisman, and Tanja Stadler. Generative Bayesian modeling to nowcast the effective reproduction number from line list data with missing symptom onset dates. *PLOS Computational Biology*, 20(4):e1012021, 2024a.

Adrian Lison, Timothy R Julian, and Tanja Stadler. Improving inference in wastewater-based epidemiology by modelling the statistical features of digital pcr. *bioRxiv*, page 10.1101/2024.10.14.618307, 2024b.

F Liu, M J Bayarri, and J O Berger. Modularization in Bayesian Analysis, with Emphasis on Analysis of Computer Models. *Bayesian Analysis*, 4:119–150, 2009. ISSN 1931-6690. doi: 10.1214/09-ba404. URL `http://dx.doi.org/10.1214/09-ba404`.

Yang Liu and Robert JB Goudie. A general framework for cutting feedback within modularized Bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf012, 2025.

James O Lloyd-Smith, Sebastian J Schreiber, P Ekkehard Kopp, and Wayne M Getz. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359, 2005.

David Lunn, Christopher Jackson, Nicky Best, Andrew Thomas, and David Spiegelhalter. The BUGS book. *A practical introduction to Bayesian analysis, Chapman Hall, London*, 2013.

Thomas Maishman, Stephanie Schaap, Daniel S Silk, Sarah J Nevitt, David C Woods, and Veronica E Bowman. Statistical methods used to combine the effective reproduction number, R (t), and other related measures of COVID-19 in the UK. *Statistical Methods in Medical Research*, 31(9):1757–1777, 2022.

Andrew A Manderson and Robert JB Goudie. Combining chains of Bayesian models with Markov melding. *Bayesian analysis*, 18(3):807, 2023.

Harrison Manley, Josie Park, Luke Bevan, Alberto Sanchez-Marroquin, Gabriel Danelian, Thomas Bayley, Veronica Bowman, Thomas Maishman, Thomas Finnie, André Charlett, Nicholas A Watkins, Johanna Hutchinson, Graham Medley, Steven Riley, Jasmina Panovska-Griffiths, and Nowcasts Model Contributing Group. Combining models to generate a consensus effective reproduction number R for the COVID-19 epidemic status in England. *Epidemiology and Infection*, 152:e59, 2024.

Colm Marshall, Paul L Delamater, and Joseph P Messina. When are predictions useful? A new method for evaluating epidemic forecasts. *medRxiv preprint*, 2024. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC11622944/`.

James M McCaw and Michael J Plank. The role of the mathematical sciences in supporting the COVID-19 response in Australia and New Zealand. *arXiv preprint arXiv:2305.04897*, 2023. URL `https://arxiv.org/abs/2305.04897`.

Jonathon Mellor, Maria L Tang, Emilie Finch, Rachel Christie, Oliver Polhill, Christopher E Overton, Ann Hoban, Amy Douglas, Sarah R Deeny, and Thomas Ward. An application of nowcasting methods: Cases of norovirus during the winter 2023/2024 in England. *PLOS Computational Biology*, 21(2):e1012849, 2025.

Nicholas Michaud, Perry de Valpine, Daniel Turek, Christopher J Paciorek, and Dao Nguyen. Sequential Monte Carlo methods in the nimble and nimbleSMC R packages. *Journal of Statistical Software*, 100: 1–39, 2021.

Sonia Natalie Mitchell, Andrew Lahiff, Nathan Cummings, Jonathan Hollocombe, Bram Boskamp, Ryan Field, Dennis Reddyhoff, Kristian Zarebski, Antony Wilson, Bruno Viola, et al. Fair data pipeline: provenance-driven data management for traceable scientific workflows. *Philosophical Transactions of the Royal Society A*, 380(2233):20210300, 2022.

In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1): 90–100, 2003.

George Nicholson, Marta Blangiardo, Sylvia Richardson, Chris Holmes, et al. Interoperability of Statistical Models in Pandemic Preparedness: Principles and Reality. *Statistical Science*, 37(2):183–206, 2022. doi: 10.1214/22-sts854.

Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Sang Woo Park, Andrei R Akhmetzhanov, Kelly Charniga, Anne Cori, Nicholas G Davies, Jonathan Dushoff, Sebastian Funk, Katie Gostic, Bryan Grenfell, N M Linton, et al. Estimating Epidemiological Delay Distributions for Infectious Diseases. *medRxiv*, 2024. doi: https://doi.org/10.1101/2024.01.12.24301247.

Du Phan, Neeraj Pradhan, and Martin Jankowiak. Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*, 2019.

Michael J Plank and Matthew J Simpson. Structured methods for parameter inference and uncertainty quantification for mechanistic models in the life sciences. *Royal Society Open Science*, 11(8):240733, 2024.

Martyn Plummer. Cuts in Bayesian graphical models. *Statistics and Computing*, 25(1):37–43, 2015.

Martyn Plummer et al. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, number 10 in 125, pages 1–10. Vienna, Austria, 2003.

Anne M Presanis, David Ohlssen, David J Spiegelhalter, and Daniela De Angelis. Conflict Diagnostics in Directed Acyclic Graphs, with Applications in Bayesian Evidence Synthesis. *Statistical Science*, 28(3): 376–397, 2013.

Leah F Price, Christopher C Drovandi, Anthony Lee, and David J Nott. Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11, 2018.

Louis Raynal, Jean-Michel Marin, Pierre Pudlo, Mathieu Ribatet, Christian P Robert, and Arnaud Estoup. ABC random forests for Bayesian parameter inference. *Bioinformatics*, 35(10):1720–1728, 2019.

Jakob Robnik, G Bruno De Luca, Eva Silverstein, and Uroš Seljak. Microcanonical Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 24(311):1–34, 2023.

Małgorzata Roos, Thiago G. Martins, Leonhard Held, and Håvard Rue. Sensitivity Analysis for Bayesian Hierarchical Models. *Bayesian Analysis*, 10(2):321 – 349, 2015. doi: 10.1214/14-BA909. URL `https://doi.org/10.1214/14-BA909`.

Conor Rosato, Joshua Murphy, Alessandro Varsi, Paul Horridge, and Simon Maskell. Enhanced SMC^2: Leveraging Gradient Information from Differentiable Particle Filters Within Langevin Proposals. In *2024 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 1–8. IEEE, 2024.

Donald B Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, pages 1151–1172, 1984.

Håvard Rue, Andrea Riebler, Sigrunn H Sørbye, Janine B Illian, Daniel P Simpson, and Finn K Lindgren. Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4(1):395–421, 2017.

Timothy W Russell, Hermaleigh Townsley, Joel Hellewell, Sam Abbott, et al. Combined analyses of within-host SARS-CoV-2 viral kinetics and information on past exposures to the virus in a human cohort identifies intrinsic differences of Omicron and Delta variants. *PLOS Biology*, 22(1), 2024. doi: 10.1371/journal.pbio.3002463.

Gunnar Øyvind Isaksson Rø, Steinar Engen, Karin Nygård, and Gunnar Øyvind Isaksson Rø. Estimating the trend of COVID-19 in Norway by combining multiple surveillance indicators. *PLOS ONE*, 20(1), 2025. doi: 10.1371/journal.pone.0317105.

James A. Scott, Axel Gandy, Swapnil Mishra, Juliette Unwin, Seth Flaxman, and Samir Bhatt. epidemia: Modeling of Epidemics using Hierarchical Bayesian Models, 2020. URL `https://imperialcollegelondon.github.io/epidemia/`. R package version 1.0.0.

Shaun R Seaman, Pantelis Samartsidis, Meaghan Kall, and Daniela De Angelis. Nowcasting COVID-19 deaths in England by age and region. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(5):1266–1281, 2022.

Freya M Shearer, Laura Edwards, Martyn Kirk, Oliver Eales, Nick Golding, Jenna Hassall, Bette Liu, Michael Lydeamore, Caroline Miller, Robert Moss, et al. Opportunities to strengthen respiratory virus surveillance systems in Australia: lessons learned from the COVID-19 response. *Communicable Diseases Intelligence*, 48:10.33321/cdi.2024.48.47, 2024.

Katharine Sherratt, Sam Abbott, Sophie R Meakin, et al. Exploring surveillance data biases when estimating the reproduction number: with insights into subpopulation transmission of COVID-19 in England. *Philosophical Transactions of the Royal Society B*, 376(1829), 2021. doi: 10.1098/rstb.2020.0283.

Laura A Skrip and Jeffrey P Townsend. Modeling approaches toward understanding infectious disease transmission. In *Immunoepidemiology*, pages 227–243. Springer, 2019.

Geir Storvik, Alfonso Diz-Lois Palomares, Solveig Engebretsen, Gunnar Øyvind Isaksson Rø, Kenth Engø-Monsen, Anja Bråthen Kristoffersen, Birgitte Freiesleben de Blasio, and Arnoldo Frigessi. A sequential Monte Carlo approach to estimate a time-varying reproduction number in infectious disease models: the Covid-19 case. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(4):616–632, 2023.

Erik Štrumbelj, Alexandre Bouchard-Côté, Jukka Corander, Andrew Gelman, Håvard Rue, Lawrence Murray, Henri Pesonen, Martyn Plummer, and Aki Vehtari. Past, present and future of software for Bayesian inference. *Statistical Science*, 39(1):46–61, 2024.

Nikola Surjanovic, Miguel Biron-Lattes, Paul Tiede, Saifuddin Syed, Trevor Campbell, and Alexandre Bouchard-Côté. Pigeons.jl: Distributed sampling from intractable distributions. *arXiv:2308.09769*, 2023.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv*, 2018.

Dhorasso Temfack and Jason Wyse. Sequential Monte Carlo Squared for online inference in stochastic epidemic models. *Epidemics*, 52:100847, 2025. ISSN 1755-4365. doi: https://doi.org/10.1016/j.epidem.2025.100847. URL `https://www.sciencedirect.com/science/article/pii/S1755436525000350`.

Emma G Thomas, James M McCaw, Heath A Kelly, Kristina A Grant, and Jodie McVernon. Quantifying differences in the epidemic curves from three influenza surveillance systems: a nonlinear regression analysis. *Epidemiology and Infection*, 143(2):427–439, 2015.

Christian Tönsing, Jens Timmer, and Clemens Kreutz. Profile likelihood-based analyses of infectious disease models. *Statistical methods in medical research*, 27(7):1979–1998, 2018.

Meri RJ Varkila, Maria E Montez-Rath, Joshua A Salomon, Xue Yu, Geoffrey A Block, Douglas K Owens, Glenn M Chertow, Julie Parsonnet, and Shuchi Anand. Use of wastewater metrics to track COVID-19 in the US. *JAMA Network Open*, 6(7):e2325591–e2325591, 2023.

Jakub Voznica, Anna Zhukova, Veronika Boskova, E Saulnier, F Lemoine, Mathieu Moslonka-Lefebvre, and Olivier Gascuel. Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature Communications*, 13(1):3896, 2022.

Thomas Ward, Martyn Fyles, Alex Glaser, Robert S Paton, William Ferguson, and Christopher E Overton. The real-time infection hospitalisation and fatality risk across the COVID-19 pandemic in England. *Nature Communications*, 15(1):4633, 2024.

Leighton M Watson, Michael J Plank, Bridget A Armstrong, Joanne R Chapman, Joanne Hewitt, Helen Morris, Alvaro Orsi, Michael Bunce, Christl A Donnelly, and Nicholas Steyn. Jointly estimating epidemiological dynamics of Covid-19 from case and wastewater data in Aotearoa New Zealand. *Communications Medicine*, 4(1):143, 2024.

Mike West and Jeff Harrison. *Bayesian forecasting and dynamic models*. Springer, 1997.

Erin R Whitehouse, Nancy Gerloff, Randall English, Stacie K Reckling, Mohammed A Alazawi, Meghan Fuschino, Kirsten St George, Daniel Lang, Eli S Rosenberg, Enoma Omoregie, et al. Wastewater surveillance for poliovirus in selected jurisdictions, United States, 2022–2023. *Emerging Infectious Diseases*, 30 (11):2279, 2024.

Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310): 1102–1104, 2010.

World Health Organization. "Crafting the mosaic"': a framework for resilient surveillance for respiratory viruses of epidemic and pandemic potential. Technical report, World Health Organization, 2023. URL `https://www.who.int/publications/i/item/9789240070288`. Accessed: 2025-10-16.

World Health Organization. *Implementing the integrated sentinel surveillance of influenza and other respiratory viruses of epidemic and pandemic potential by the Global Influenza Surveillance and Response System: standards and operational guidance*. World Health Organization, 2024.

A. W. C. Yan. *Influenza viral dynamics models to explore the roles of innate and adaptive immunity*. PhD thesis, The University of Melbourne, 2017. URL `https://hdl.handle.net/11343/192580`.

Fuming Yang, David Nott, and Anne M Presanis. Detecting Conflicts in Evidence Synthesis Models Using Score Discrepancies. 2025.

Yuling Yao, Aki Vehtari, Daniel Simpson, and Andrew Gelman. Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1003, 2018.

Xuejun Yu, David J Nott, and Michael Stanley Smith. Variational inference for cutting feedback in misspecified models. *Statistical Science*, 38(3):490–509, 2023.

Xiawan Zheng, Keyue Zhao, Xiaoqing Xu, Yu Deng, Kathy Leung, Joseph T Wu, Gabriel M Leung, Malik Peiris, Leo LM Poon, and Tong Zhang. Development and application of influenza virus wastewater surveillance in Hong Kong. *Water Research*, 245:120594, 2023.

Xiawan Zheng, Keyue Zhao, Bingjie Xue, Yu Deng, Xiaoqing Xu, Weifu Yan, Chao Rong, Kathy Leung, Joseph T Wu, Gabriel M Leung, et al. Tracking diarrhea viruses and mpox virus using the wastewater surveillance network in Hong Kong. *Water Research*, 255:121513, 2024.