Structural identifiability of compartmental models for infectious disease transmission is influenced by data type

Author names and affiliations: Emmanuelle A. Dankwa[1], Andrew F. Brouwer[2], Christl A. Donnelly[1,3,*]

[1] Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, United Kingdom
[2] Department of Epidemiology, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109-2029, USA
[3] Department of Infectious Disease Epidemiology, Faculty of Medicine, School of Public Health, Imperial College London, United Kingdom

*Corresponding author: Christl A. Donnelly, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, United Kingdom
Email address: christl.donnelly@stats.ox.ac.uk

ABSTRACT

If model identifiability is not confirmed, inferences from infectious disease transmission models may not be reliable, so they might lead to misleading recommendations. Structural identifiability analysis characterizes whether it is possible to obtain unique solutions for all unknown model parameters, given the model structure.

In this work, we studied the structural identifiability of some typical deterministic compartmental models for infectious disease transmission, focusing on the influence of the data type considered as model output on the identifiability of unknown model parameters, including initial conditions.

We defined 26 model versions, each having a unique combination of underlying compartmental structure and data type(s) considered as model output(s). Four compartmental model structures and three common data types in disease surveillance (incidence, prevalence and detected vector counts) were studied.

The structural identifiability of some parameters varied depending on the type of model output. In general, models with multiple data types as outputs had more structurally identifiable parameters, than did models with a single data type as output.

This study highlights the importance of a careful consideration of data types as an integral part of the inference process with compartmental infectious disease transmission models.

Keywords: structural identifiability, infectious disease transmission, compartmental models, data types

1.1 INTRODUCTION

Goals of infectious disease transmission modelling often include making inferences about the underlying transmission process, predicting the future course of an epidemic given a range of interventions, or estimating what would have happened in a counterfactual scenario. A defined model is fitted to a given data set (frequently an incidence time series generated by passive surveillance). This model-fitting process is parameter estimation, where one determines parameter values or distributions corresponding to model outputs that best fit (or at least, approximate) the observed data. Parameter estimation however can only produce robust results if the model is *identifiable* (Audoly et al., 2001; Castro and de Boer, 2020; Cobelli and Distefano III, 1980; Kao and Eisenberg, 2018; Ljung and Glad, 1994; Villaverde et al., 2016; Wieland et al., 2021): that is, if it is possible, in principle, to obtain unique solutions for all unknown model parameters, given the model structure and available data. We note that other properties such as predictability (Castro et al., 2020; Scarpino and Petri, 2019) and uncertainty quantification (Capaldi et al., 2012; McCabe et al., 2021) also affect the reliability of model inferences (Massonis et al., 2021a); these are not treated here, however.

Although the subject of identifiability has received considerable attention in the systems biology and control literature (see (Wieland et al., 2021) for a recent review), it is inconsistently applied in the infectious disease modelling literature. Relatively few studies exist on the identifiability analysis of infectious disease models, e.g., (Brouwer et al., 2018; Eisenberg et al., 2013; Evans et al., 2005; Kao and Eisenberg, 2018; Massonis et al., 2021a; Tuncer et al., 2016; Tuncer and Le, 2018), and the practice of routinely checking the identifiability of these models *before* parameter estimation is not widespread. Nevertheless, identifiability is required to make meaningful inferences on model parameters and, consequently, to provide reliable evidence to inform public health policymaking.

In a non-identifiable model, parameter sets with similar values may yield considerably different model predictions (Kao and Eisenberg, 2018; Roda et al., 2020). Thus, a failure to consider identifiability could result in misleading recommendations, as has been previously noted (Kao and Eisenberg, 2018; Massonis et al., 2021; Roda et al., 2020; Roosa and Chowell, 2019),

some of which could have serious consequences. For example, Kao and Eisenberg demonstrated using a dengue transmission model that two sets of parameters which fit the incidence data comparably well yielded very different predictions for incidence after an intervention is applied (see Fig. 9 in their paper) (Kao and Eisenberg, 2018). Roda and colleagues also showed that the lack of identifiability in COVID-19 transmission models could lead to extreme variability in predictions (Roda et al., 2020).

A distinction is made between the two types of identifiability: *structural identifiability* and *practical identifiability*. Structural identifiability, a concept first introduced by Bellman and Astrom in 1970, is a property of the model structure and associated measurement function (i.e., the function of model variables that is to be observed) and does not depend on the quantity or quality of the observed data. It addresses the question: Given an error-free model structure, and assuming noise-free, infinite data, do unique solutions exist for the model parameters? Structural identifiability is affected by: 1) the nature of the model parameterization (Muñoz-Tamayo et al., 2018) which influences symmetries, i.e., functional relationships between model parameters (Eisenberg and Hayashi, 2014; Hengl et al., 2007; Massonis et al., 2021a; Villaverde, 2022); and 2) the data type considered as model output (Balsa-Canto et al., 2010; Chis et al., 2011; Massonis et al., 2021a; Tuncer and Le, 2018). Practical identifiability, on the other hand, is related to the adequacy of the available observed data for the estimation problem at hand (Balsa-Canto et al., 2010; Brouwer et al., 2017; Kao and Eisenberg, 2018; Miao et al., 2011; Raue et al., 2009; Tuncer et al., 2016; Tuncer and Le, 2018). The corresponding question is: Do the data contain enough information to infer the model parameters? Structural identifiability is a necessary, but not sufficient, condition for practical identifiability; that is, a structurally non-identifiable parameter cannot be practically identifiable, and a structurally identifiable parameter could be practically non-identifiable depending on the data available (Cobelli and Distefano III, 1980; Eisenberg et al., 2013). In this work, we are considering structural identifiability of infectious disease transmission models.

Several studies have demonstrated the influence of the type of observed data on the structural identifiability of infectious disease models. For example, Tuncer and Le studied a Susceptible-Infected-Treated-Recovered epidemic model which becomes structurally identifiable only when both cumulative incidence rates and the number of treated individuals is observed (Tuncer and Le, 2018). In the same work, the authors explored the identifiability of a Susceptible-Exposed-Infected-Recovered model and showed that the type of structural identifiability for two parameters (recovery rate and length of latent period) depended on whether the observed data were cumulative incidence or prevalence. Similar works include (Evans et al., 2005), on the structural identifiability of a seasonally forced SIR model with prevalence and a proportion of the incidence as outputs; (Eisenberg et al., 2013), on the identifiability of parameters of compartmental models for cholera with prevalence as output; (Tuncer et al., 2016), on the identifiability of an immune-epidemiological model for Rift Valley fever with time-series data of viremia levels as output; (Kao and Eisenberg, 2018), on the identifiability of a dengue transmission model with various types of human and mosquito incidence data as outputs; and more recently, (Massonis et al., 2021a), on the structural identifiability of a wide range of COVID-19 transmission models with a variety of surveillance data types as outputs.

However, few of these studies (e.g., (Eisenberg et al., 2013; Evans et al., 2005)) have explicitly studied the identifiability of unknown initial conditions (ICs). Other studies have either assumed known ICs (e.g., (Tuncer and Le, 2018)) or have implicitly considered unknown ICs through assessment of the *observability* of model states (Massonis et al., 2021a); i.e., whether the state variable trajectories can be uniquely determined from observed data. (Structural identifiability has been considered as a particular case of observability (Massonis et al., 2021a; Sedoglavic, 2002; Tunali and Tarn, 1987; Villaverde, 2019).) Here, we explicitly consider ICs as unknown parameters in all models and analyse their structural identifiability given various data types. Often values are assumed for ICs, but careful analysis often reveals that parameter estimates depend on these IC assumptions. We can ask under what circumstances ICs can be uniquely determined from observed data. Although this question might technically be

considered one about observability, when the ICs are reframed as parameters, the question is one of identifiability. Thus, our work adds to the literature by examining how the structural identifiability of ICs of classic compartmental models change with data type. Additionally, we employ a publicly available web-based toolbox, SIAN (Hong et al., 2019), to analyse the structural identifiability of model parameters, allowing us to demonstrate the utility of such tools.

Specifically, we consider four compartmental structures (SIR, SLIR, SLIR with vaccination and relapse and a vector-borne disease model with SLIR for hosts and SLI for vectors) and three common data types in disease surveillance (incidence, prevalence and detected vector counts). Using SIAN, we analyse the structural identifiability of unknown parameters in 26 model versions, each a unique combination of underlying compartmental structure and data type considered as model output. We use the term “model version” to refer to a compartmental structure-output(s) combination; e.g., SIR with incidence, or SLIR with incidence and prevalence.

Although the compartmental structures and data types we consider are by no means exhaustive, our work is intended to demonstrate the importance of identifiability and to be instructive for those seeking to apply these techniques to their own models. We have therefore made available all input codes and output files to facilitate reproducibility: https://github.com/emmanuelle-dankwa/structural-identifiability-epi-models.

The paper is outlined as follows. In Section 1.2, we introduce the general modelling framework and notation and provide formal definitions of relevant structural identifiability concepts. Here, we also introduce the four compartmental structures, briefly introduce the software toolbox utilized, present the model versions examined and finally, outline the structural identifiability analysis performed. Section 1.3 presents the results and Section 1.4 presents a discussion of results. Concluding remarks are given in Section 1.5.

1.2 METHODS

1.2.1 General modelling framework and formal definitions

Consider a deterministic ordinary differential equation (ODE) infectious disease transmission model $\mathcal{M}$ of the form

$$\mathcal{M} \coloneqq \begin{cases} \dot{\boldsymbol{X}}(t) = f\big(\boldsymbol{X}(t), \boldsymbol{p}, \boldsymbol{u}(t)\big) \\ \boldsymbol{y}(t) = g(\boldsymbol{X}(t), \boldsymbol{p}) \\ \boldsymbol{X}_{t_0} = \boldsymbol{X}(t_0) \end{cases}, \tag{1}$$

with observations on the interval $t_0 \leq t \leq T$, where $\dot{\boldsymbol{X}}(\boldsymbol{t})$ is a system of non-linear ODEs, $\boldsymbol{X}(\boldsymbol{t}) \in R^{n_X}$ is a vector of time-varying disease states and the unique solution to the system $\dot{\boldsymbol{X}}(\boldsymbol{t})$, $\boldsymbol{p} \in R^{n_p}$ is a vector of constant unknown model parameters, $\boldsymbol{y}(t) \in R^{n_y}$ is a vector of time-dependent model outputs corresponding to a specific data type (for example, case incidence rates), $\boldsymbol{u}(t) \in R^{n_u}$ is a time-dependent input vector, g is the measurement equation (which defines the relationship between $\boldsymbol{X}(\boldsymbol{t})$, $\boldsymbol{p}$ and $\boldsymbol{y}$), and $\boldsymbol{X}_{t_0} \subset R^{n_X}$ is a vector of the known ICs. Note that unknown components of $\boldsymbol{X}_{t_0}$ are included in $\boldsymbol{p}$ and that f and g are vectors of analytic functions of their arguments.

The formal definition of structural identifiability for a model and its parameters is given below. The structural identifiability of a parameter may either be *local* (i.e., holding only within a limited region of the parameter space or about a given point) or *global* (i.e., holding (almost) everywhere within the parameter space) (Ljung and Glad, 1994).

Definition 1 [Parameter structural identifiability] (Cobelli and Distefano III, 1980; Ljung and Glad, 1994)

A parameter $p_i \in \boldsymbol{p}$ is structurally globally identifiable (s.g.i.) on the time interval $[t_0, T]$ for a given output $\boldsymbol{y}$ if a unique solution exists for p_i; that is, if and only if for almost any $\boldsymbol{p}^*$ and

almost any IC (i.e., excluding degenerate values), $\boldsymbol{y}(\boldsymbol{X}, \hat{\boldsymbol{p}}) = \boldsymbol{y}(\boldsymbol{X}, \boldsymbol{p}^*)$ *implies* $\hat{p}_i = p_i^*$. *Otherwise,* p_i *is structurally globally non-identifiable.*

A parameter $p_i \in \boldsymbol{p}$ *is structurally locally identifiable (s.l.i.) on the time interval* $[t_0, T]$ *for a given output* $\boldsymbol{y}$ *if there exists a neighbourhood* $V(\boldsymbol{p})$ *of the parameter space within which a unique solution exists for* p_i. *Otherwise,* p_i *is structurally non-identifiable (s.n.i.).*

Definition 2 [Model structural identifiability] (Cobelli and Distefano III, 1980; Ljung and Glad, 1994)

The model $\mathcal{M}$ *is s.g.i. for a given output* $\boldsymbol{y}$ *if every* $p_i \in \boldsymbol{p}$ *is s.g.i.* given $\boldsymbol{y}$.

The model $\mathcal{M}$ *is s.l.i. for a given output* $\boldsymbol{y}$ *if at least one* $p_i \in \boldsymbol{p}$ *is s.l.i. given* $\boldsymbol{y}$ *and if no* $p_i \in \boldsymbol{p}$ *is s.n.i.*

The model M *is s.n.i. for a given output* $\boldsymbol{y}$ *if at least one* $p_i \in \boldsymbol{p}$ *is s.n.i. given* $\boldsymbol{y}$.

1.2.2 Model structures

The most basic model structure we consider is the SIR model. For simplicity, we assume no demography, no migration, homogenous populations, and a constant, unknown population size N. In this SIR model, there are three mutually exclusive compartments, each corresponding to a distinct infection state: Susceptible S, Infectious I and Recovered (and immune) R. Susceptible individuals become infected at a rate $\beta \mathrm{I}/N$ where β is the transmission rate and is equal to the product of the contact rate and the probability that a contact will successfully result in an infection. Infectious individuals recover at a rate γ. These dynamics can be described by the following set of ODEs:

$$\textbf{SIR}: \begin{cases} \dfrac{dS}{dt} = -\dfrac{\beta \mathrm{SI}}{N} \\ \dfrac{dI}{dt} = \dfrac{\beta \mathrm{SI}}{N} - \gamma I \end{cases} . \qquad (2)$$

$$\frac{dR}{dt} = \gamma I$$

ICs for the S, I and R states will be denoted by $\mathrm{S}(0), \mathrm{I}(0)$ and $\mathrm{R}(0)$, respectively. At any time $t \geq 0$, $\mathrm{N} = \mathrm{S}(t) + \mathrm{I}(t) + \mathrm{R}(t)$. For the SIR model, we consider two outputs: incidence, $y_1 = \beta \mathrm{SI}/N$, and prevalence, $y_2 = \mathrm{I}/N$. In this context, incidence is defined as the number of new cases arising within a given time period, while prevalence is defined as the infectious proportion of the population at a given time point. In many situations, incidence data are presented as cumulative incidence; cumulative incidence contains the same information from an identifiability perspective, but, for statistical reasons, it is preferable to convert cumulative incidence to incidence before fitting (King et al., 2015). Incidence data are often generated through passive surveillance (e.g., number of new cases reported each day from a hospital system), while prevalence data may be generated through active surveillance (e.g., door-to-door data collection; testing of people at random regardless of symptoms). In reality, both incidence and prevalence are subject to bias from reporting rates and asymptomatic infection. Some studies explicitly include a reporting rate parameter κ in their measurement equations, or the effect can be implicitly accounted for in β or N. For this reason, N does not necessarily correspond to population numbers from a census of the catchment region, and thus we treat it as an unknown quantity.

For diseases with a non-negligible latent period (e.g., COVID-19 (Liu et al., 2020)), the SIR model can be modified to include a latent state L. The modified dynamics are described by the following set of equations:

SLIR:
$$\begin{cases} \frac{dS}{dt} = -\frac{\beta \mathrm{SI}}{N} \\ \frac{dL}{dt} = \frac{\beta \mathrm{SI}}{N} - \alpha L \\ \frac{dI}{dt} = \alpha L - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases}, \quad (3)$$

where $1/\alpha$ represents the length of the latent period. Let $\mathrm{L}(0)$ denote the IC for the latent state. For all $t \geq 0$, $\mathrm{N} = \mathrm{S}(t) + \mathrm{L}(t) + \mathrm{I}(t) + \mathrm{R}(t)$. We study an equivalent set of outputs as for the SIR model: incidence, $y_3 = \alpha L$, and prevalence, $y_4 = \mathrm{I}/N$.

For diseases where a relapse of symptoms is possible after a period of remission (e.g., hepatitis A), we can include a compartment Q to represent the remission state. In this model, we also allow for immunity by vaccination. The dynamics of this SLIRQ (Susceptible-Latent-Infectious-Recovered (or immune)-Remission) model as adapted from Dankwa et al. (2021) are as follows. In this model, individuals in the R compartment are immune, either as a result of vaccination or past infection. Susceptible individuals become exposed at a rate $\beta \mathrm{I}/N$ and move to the latent state, where they remain for $1/\alpha$ time units, after which they become infectious. A proportion, $1-\eta$, of infectious individuals recover temporarily, moving to the remission state for a period of $1/\sigma$ time units, after which they experience a relapse of symptoms, becoming infectious. The remaining proportion, η, of infectious individuals recover permanently and become immune. The recovery rate is γ. A number $v(t)$ of individuals are vaccinated at time t and become immune. These dynamics are captured by the following set of ODEs:

SLIRQ:

$$\begin{cases} \frac{dS}{dt} = -\frac{\beta \mathrm{SI}}{N} - v\frac{S}{N} \\ \frac{dL}{dt} = \frac{\beta \mathrm{SI}}{N} - \alpha L \\ \frac{dI}{dt} = \alpha L - \gamma I + \sigma Q \\ \frac{dR}{dt} = v\frac{S}{N} + \eta\gamma I \\ \frac{dQ}{dt} = (1-\eta)\gamma I - \sigma Q \end{cases} \quad (4)$$

The IC corresponding to the remission state will be denoted by $\mathrm{Q}(0)$. For all $t \geq 0$, $\mathrm{N} = \mathrm{S}(t) + \mathrm{L}(t) + \mathrm{I}(t) + \mathrm{R}(t) + \mathrm{Q}(t)$. We consider the same set of outputs as before: incidence, $y_5 = \alpha L + \sigma Q$, and prevalence, $y_6 = \mathrm{I}/N$.

Finally, we introduce a SLIR/SLI model structure suitable for vector-borne diseases, and adapted from the works of Ngwa and Shu (2000) and Kao and Eisenberg (2018), who apply the model to malaria and dengue, respectively. In the model, infection dynamics within the host population are explained via a SLIR model, as in equation (3), while the dynamics in the vector population are explained via a SLI model, thus a SLIR/SLI model. Transmission can only occur between individuals of different populations, i.e., host-to-vector or vector-to-host. Like in the previous models, we assume constant sizes for both populations: let N_h and N_v represent the sizes of the host and vector populations, respectively. We use subscripts "h" and "v" to represent compartments for hosts and vectors, respectively. Thus, we have $N_h = S_h(t) + L_h(t) + I_h(t) + R_h(t)$ and $N_v = S_v(t) + L_v(t) + I_v(t)$, $\forall t \geq 0$.

The pathogen transmission rate from host to vector β_{hv} is equal to the product of the contact rate between host and vector (in malaria for example, this may be the human biting rate of mosquitoes) and the probability of successful transmission from an infectious host to a susceptible vector. Similarly, the transmission rate from vector to host β_{vh} is equal to the product of the contact rate between vector and host and the probability of successful transmission from an infectious vector to a susceptible host. Infected hosts become infectious

after a latency period of $1/\alpha_h$ time units and remain infectious for a period of $1/\gamma_h$ time units before recovery. Recovered hosts become immune to the disease. Infectious hosts transmit the pathogen to susceptible vectors at a rate $\beta_{hv} I_h/N_h$. Infected vectors become infectious after a latency period of $1/\alpha_v$ time units. Infectious vectors transmit the pathogen to susceptible hosts at a rate $\beta_{vh} I_v/N_h$. Within each population, we assume equal birth and death rates: μ_h and μ_v for hosts and vectors, respectively, so no disease-related mortality is incorporated. The SLIR/SLI model is represented by the following system of ODEs:

$$\textbf{SLIR/SLI}: \begin{cases} \frac{dS_h}{dt} = \mu_h N_h - \frac{\beta_{vh} S_h I_v}{N_h} - \mu_h S_h \\ \frac{dL_h}{dt} = \frac{\beta_{vh} S_h I_v}{N_h} - \alpha_h L_h - \mu_h L_h \\ \frac{dI_h}{dt} = \alpha_h L_h - \gamma_h I_h - \mu_h I_h \\ \frac{dR_h}{dt} = \gamma_h I_h - \mu_h R_h \\ \frac{dS_v}{dt} = \mu_v N_v - \frac{\beta_{hv} S_v I_h}{N_h} - \mu_v S_v \\ \frac{dL_v}{dt} = \frac{\beta_{hv} S_v I_h}{N_h} - \alpha_v L_v - \mu_v L_v \\ \frac{dI_v}{dt} = \alpha_v L_v - \mu_v I_v \end{cases} . \quad (5)$$

The ICs for the SLIR/SLI model will be denoted by (listed in order of states): $S_h(0), E_h(0), I_h(0), R_h(0), S_v(0), L_v(0)$ and $I_v(0)$.

The following outputs are studied: 1) incidence in hosts (host incidence), $y_7 = \alpha_h L_h$; 2) prevalence in hosts (host prevalence), $y_8 = I_h/N_h$; 3) incidence in vectors (vector incidence), $y_9 = \alpha_v L_v$; and 4) detected vector counts, $y_{10} = \lambda_v(S_v + L_v + I_v)$, λ_v is an unknown vector detection rate.

1.2.3 Toolbox employed

In this study, we employ the SIAN (Structural Identifiability Analyser) (Hong et al., 2019) software tool for structural identifiability analysis. The algorithm implemented in SIAN, proposed by Hong et al. (2020), is based on a combination of differential algebra and Taylor series approaches to structural identifiability analysis. SIAN is implemented in Maple and is available as a web application: https://maple.cloud/app/6509768948056064.

Here, we are interested in assessing both local and global structural identifiability of model parameters, including ICs. Therefore, although other toolboxes exist which are capable of assessing the local and global structural identifiability of $\mathcal{M}$ (e.g., COMBOS (Meshkat et al., 2014), DAISY (Bellu et al., 2007) and GenSSI 2.0 (Ligon et al., 2018)), we employ SIAN because it uniquely possesses the following combination of characteristics as desired for this study. First, it is capable of assessing both local and global identifiability of model parameters. Second, it provides identifiability results for parameter-based ICs. Third, it is available as a web application and accepts a simple text-based input, hence more accessible than toolboxes which require program installation or knowledge of a particular programming language. This latter characteristic is a particularly desirable one for a structural identifiability analysis software, as it addresses a potential barrier to the application of structural identifiability analysis. A comparison of the performance and features of toolboxes for structural identifiability analysis of ODE models is beyond the scope of the study. Interested readers may consult Ligon et al. (2018) and Hong et al. (2019).

For a given model, SIAN typically produces one of the following results for the structural identifiability of each model parameter: s.g.i., s.l.i. or s.n.i. SIAN is also capable of computing identifiable combinations, although we do not employ that functionality here.

1.2.4 Structural identifiability assessments

Structural identifiability analysis was conducted in four stages, each stage designed to reflect a possible scenario that may be encountered when modelling infectious disease transmission. Across these stages, we studied the structural identifiability of model parameters given three common data types as model outputs – incidence, prevalence, and detected vector counts (the latter only applicable to SLIR/SLI). We analysed 26 ODE model versions, assuming in all cases constant, unknown population sizes. For each model, we assessed the structural identifiability of all unknown parameters, including ICs.

Stages are now described.

- **Stage one (single outputs, all parameters unknown)**: Structural identifiability analysis was conducted for models defined with a single data type as output and assuming all parameters were unknown. This scenario is typical in the initial stages of an outbreak of an emerging pathogen, when little is known of pathogen epidemiology and consequently, natural history parameters or transmission rates. Furthermore, in such scenarios, as data are often limited, only one type of data may be available for parameter estimation. It is therefore of interest to determine which parameters are structurally identifiable in such contexts. Therefore, for SIR, SLIR and SLIRQ, we assessed the structural identifiability of model parameters given either incidence or prevalence data. For SLIR/SLI, output was host incidence or host prevalence. We do not consider vector data at this stage, as such data are less likely to be available during the early stages of an emerging vector-borne disease outbreak. Thus, at this stage, eight model versions were analysed.

- **Stage two (single outputs, only natural history parameters known)**: In the case of an endemic disease which has been widely studied (e.g., malaria in sub-Saharan Africa), a high level of certainty may be obtained on the values of natural history parameters. In modelling transmission of such diseases, knowledge of natural history parameters may be assumed and hence these parameters may be treated as known quantities in the model. Stage two considers this scenario. For the model

versions analysed at stage one, we assumed all natural history parameters to be known and re-evaluated the structural identifiability of the other (unknown) model parameters, i.e., all ICs, transmission rate parameters, and for the SLIR/SLI models, the demography parameters, additionally. This analysis enabled us to identify how the structural identifiability properties of unknown parameters change once other parameters in the model are assumed known. As in stage one, eight model versions were analysed at this stage.

- **Stage three (multiple outputs, all parameters unknown)**: In instances where surveillance capacities are strengthened in the face of an emerging outbreak, it is possible to observe more than one type of data. For example, in the context of a vector-borne disease outbreak, there may be, in addition to host incidence data, data on the size of the vector population, as could be obtained through traps in the case of mosquitoes (for mosquito-borne diseases), or field signs, in the case of badgers (for bovine tuberculosis). In stage three, we studied the structural identifiability of model parameters in these "data-rich" scenarios by defining models to have at least two output types. All parameters were treated as unknown, as in stage one. Thus, we were able to compare results obtained at this stage to results at stage one (with single outputs) to assess the influence of additional outputs on parameters' structural identifiability.

  For the SIR, SLIR and SLIRQ structures, outputs were incidence and prevalence. For the SLIR/SLI structure, we studied two output combinations. One comprised host incidence and host prevalence, reflecting a scenario in which host infection data are available but vector data are absent, while the other comprised both host and vector data: host incidence, host prevalence, vector incidence and detected vector counts. Thus, five model versions were analysed at this stage.

- **Stage four (multiple outputs, only natural history parameters known)**: Here, we consider the five model versions analysed at stage three, but assuming knowledge of natural history parameters, as in stage two. Thus, we could compare the structural identifiability of parameters at this stage to corresponding results: 1) at stage two, to determine whether additional outputs improved parameters' structural identifiability after some parameters have been assumed known; and 2) at stage three, to determine how structural identifiability of parameters improved with knowledge of natural history parameters, given multiple outputs.

1.3 RESULTS

Structural identifiability results of model parameters assessed at stages one, two, three, and four are presented in Table 1, **Error! Reference source not found.**, **Error! Reference source not found.**, and **Error! Reference source not found.**, respectively. For some models, SIAN was unable to complete global identifiability calculations but provides results for local identifiability. For these model versions, parameters assessed as being s.l.i. by SIAN are referred to in this study as being *at least* s.l.i., given that they may potentially be s.g.i.

Results are now discussed by stage.

Stage one (single outputs, all parameters unknown): See Table 1. When all parameters were assumed unknown and single outputs considered, all models except the SLIRQ models are s.n.i. All parameters of the SLIRQ model are s.l.i., irrespective of output type. In the SIR and SLIR models with output as prevalence, the transmission rate β is s.g.i. However, with output as incidence, β becomes s.n.i. We should note, however, that β/N is an identifiable combination (meaning that its value is identifiable even if the constituent parameters are not), and the assumption of unknown N is the reason that both $R(0)$ and β are s.n.i. in these two models. The IC for the recovered compartment $R(0)$ is s.n.i. in all SIR and SLIR models studied at stage one but is at least s.l.i. in both SLIRQ models (i.e., given incidence or prevalence as output). In the SLIR/SIR model with incidence as output, the IC corresponding

to the recovered compartment for hosts $R_h(0)$ is at least s.l.i. when output is host incidence but s.n.i. when output is host prevalence. The transmission rate parameter and all ICs corresponding to the vector population are s.n.i. with host prevalence or host prevalence as output, while other parameters associated with the vector population (birth rate μ_v and parameter controlling the length of latent period α_v) are at least s.l.i.

Stage two (single outputs, only natural history parameters known): See **Error! Reference source not found.**. Assuming knowledge of the natural history parameters in the SIR, SLIR and SLIR/SLI models did not lead to an improvement of the structural identifiability of parameters which were s.n.i. at stage one (where all parameters – including natural history parameters – were unknown), irrespective of output type. However, for the SLIRQ models, the structural identifiability of unknown parameters (β and ICs) is seen to improve with the assumption of knowledge of natural history parameters: these parameters are s.g.i. at this stage but were at least s.l.i at stage one.

Stage three (multiple outputs, all parameters unknown): See **Error! Reference source not found.**. When incidence and prevalence data are considered jointly as outputs in the same model, structural identifiability of the SIR, SLIR and SLIRQ models improves considerably compared to stage one. All parameters in these models which were s.n.i. at stage one become s.g.i. For example, β is s.n.i. in the SIR model with incidence only as output; however, with the addition of prevalence data as an output in the model, β becomes s.g.i. Likewise, $R(0)$ is s.n.i. in all SIR and SLIR models with single outputs (either incidence or prevalence; Table 1) but becomes s.g.i. when these outputs are considered simultaneously.

For the SLIR/SLI model, all parameters associated with the host population are at least s.l.i. when host incidence and host prevalence data are joint model outputs. However, the ICs and transmission rate parameter associated with the vector population are s.n.i., as in stage one when these outputs were considered separately (Table 1).

Stage four (multiple outputs, natural history parameters known): See **Error! Reference source not found.**. Even when natural history parameters are assumed known, the ICs and

transmission rate parameter associated with the vector population in the SLIR/SLI model remain s.n.i. with host prevalence and host incidence as joint model outputs. It is only with the addition of vector data (vector incidence and detected vector counts) as outputs that these parameters become s.g.i.

1.4 DISCUSSION

In this work, we have studied the structural identifiability of 26 ODE model versions, each with a unique combination of underlying compartmental structure (SIR, SLIR, SLIRQ or SLIR/SLI) and data type considered as model output (incidence, prevalence or detected vector counts).

The consideration of multiple data types as outputs generally improved models' structural identifiability. Indeed, when only single outputs were considered (Table 1, **Error! Reference source not found.**), all models except the SLIRQ-structured models were s.n.i. However, when these models were defined to have at least two data types as outputs, all but one model become s.g.i. (**Error! Reference source not found.**, **Error! Reference source not found.**).

The exception – the SLIR/SLI model with outputs as host incidence and host prevalence – had its transmission rate parameter and ICs for the vector population remaining s.n.i. despite having host incidence and host prevalence as model outputs (**Error! Reference source not found.**), and even after all natural history parameters in the model were assumed known (**Error! Reference source not found.**). However, when vector-related data (vector incidence and detected vector counts) were added as outputs in the model, these parameters become s.g.i. (**Error! Reference source not found.**, **Error! Reference source not found.**), suggesting that data on host infection alone (incidence, prevalence or both) are not sufficient to identify these vector-related parameters.

We found it surprising that the other vector-related parameters studied – μ_v, the vector birth rate and α_v, the parameter controlling the length of the latent period – were at least s.l.i. given host incidence or host prevalence (Table 1), since we expected vector-related parameters to be non-identifiable in the absence of vector data. We thus checked with other structural identifiability software capable of computing global results – GenSSI 2.0 (Ligon et al., 2018), COMBOS (Meshkat et al., 2014) and DAISY (Bellu et al., 2007) – but none of this were able to complete computations. That these vector-related parameters are identifiable with host data is not yet clear to us and it is a question we continue to explore. We suspect that these parameters are likely not practically identifiable from typically available host incidence data, even if they are structurally identifiable.

Assuming knowledge of the natural history parameters did not seem to improve the structural identifiability of parameters in the majority of single-output models (Table 1, **Error! Reference source not found.**), likely because all natural history parameters were at least s.l.i. (in those models in which they were treated as unknown parameters; Table 1), indicating that they were not in identifiable combinations. Hence, fixing the values of these parameters appeared not to have influenced existing symmetries.

We note that for all SIR and SLIR models with single outputs (incidence or prevalence), the IC corresponding to the recovered compartment $R(0)$ is s.n.i., and its structural identifiability does not improve even when natural history parameters in these models are assumed known (Table 1, **Error! Reference source not found.**). Only with the simultaneous analysis of multiple data types as outputs does $R(0)$ become s.g.i. (**Error! Reference source not found.**). It is interesting to observe this "synergy-like" effect: separately, neither incidence nor prevalence is sufficient for the identification of $R(0)$, but considered jointly, these data prove adequate to identify $R(0)$. In this case, the at-risk population size N is identifiable if both incidence and prevalence are observed, allowing determination of $R(0)$. More broadly, it is

helpful to *pre-determine* which data types will lead to structural identifiability when used separately or in combination with new, external parameter information. We recommend that formal methods for pre-determination, such as the use of identifiable parameter combinations, be used in the development of study designs: these methods may result in more efficient data collection to support inference for the specific research question.

Our results on the IC of the recovered state in models with unknown N are consistent with those of Massonis et al. (2021a) who, in a structural identifiability analysis of several compartmental COVID-19 transmission models with known N, found that the recovered state is "almost never observable". That is, its value over time cannot be determined from the given data, although it could potentially be observable with a single measurement (such as the initial condition or a later serosurvey). It is not surprising that if $R(0)$ is not identifiable in models assuming known N (Massonis et al., 2021a) that it would not be identifiable in models with unknown N (our results).

An important question then arises: what sources of data are useful to inform the IC of the recovered/immune state in scenarios where this state is not directly observed? Expert knowledge or seroprevalence estimates based on representative studies may be helpful in this regard. Where these data are not readily available, the IC for the recovered or immune state has often been set to zero; however, if the true value is different from zero, other parameters need to be interpreted accordingly and the assumptions need to be stated clearly. The transmission rate and the at-risk population size N, in particular, need to be interpreted in the context of the assumptions made about the ICs, as well as any assumptions about the reporting rate and asymptomatic fraction of cases. The distinction may be particularly important when trying to mechanistically interpret the transmission rate as a product of constituent parameters (e.g., contact rate times probability of infection) or when connecting N to catchment census data. More broadly, simulation studies and sensitivity analysis may be

needed to understand the specific influences of IC values on one's parameter estimates and thus the robustness of one's inferences.

Our study is a relevant contribution to the literature as it explicitly considers ICs and population sizes as unknown in models which have been mostly studied assuming these quantities are known. Data on ICs or population size may not always be available or able to be measured directly, hence the need to study identifiability in such scenarios. Also, as we had complete control over structure-output combinations, we were able to modify model characteristics such that the cause for a change in identifiability results could be precisely identified. In addition, unlike most previous studies, we provide input code for all analyses conducted, to serve as a model to individuals who may be new to structural identifiability analysis. To further facilitate increased adoption of structural identifiability analysis, we chose to use a web-based structural identifiability analysis tool, which accepts simple text-based inputs. This eliminates potential barriers to adoption such as the need for program installation or proficiency in a programming language.

Despite these strengths, some limitations exist. First, when models were complex (i.e., having more than four states, or multiple outputs and several parameters), it was generally challenging for SIAN (and other toolboxes used) to produce complete results. More work is needed on scaling toolboxes to match the increasing complexity of modern epidemic models. Second, it would have been desirable to use multiple toolboxes for all analysis, as that would have facilitated the detection of potentially problematic results; however as stated earlier, SIAN was the only publicly available toolbox – as far as we know – which had the combination of functionalities required for this study: 1) ability to assess both local and global identifiability of model parameters; 2) ability to assess identifiability of unknown ICs, and 3) possibility to implement without requiring program installation or specialized programming language skills. Work on developing more accessible toolboxes with a range of relevant functionalities is therefore warranted. Third, the selection of compartmental models studied here is limited.

Similarly, although the set of data types examined here comprises some of the most commonly measured in disease surveillance, it is not representative of the wide variety of possible data types; for example, we did not directly consider detected incidence (i.e., incidence allowing for underreporting, although we do acknowledge that it is important to account for in real data and contributed to our decision to assume that N is unknown). Our work is intended to be primarily illustrative, providing the rationale for assessing structural identifiability and some approaches. We also note that the work here is relevant regardless of downstream decisions to take a frequentist or Bayesian approach to parameter estimation from real data, though we do note that making a choice of informative prior distributions on parameters or initial conditions is akin to changing the assumptions of what is known or unknown, which may impact the identifiability of other aspects of the model.

Our work focused on deterministic, compartmental ODE models. It would be desirable to extend our study to cover stochastic models (Browning et al., 2020); models which incorporate population structure (e.g., age-structured or spatial models); time-varying parameters, which have been shown to address structural identifiability issues due to their role in breaking symmetries in the model structure (Massonis et al., 2021a); and additional data types such as the number of recovered individuals (Massonis et al., 2021a) and environmental surveillance (Brouwer et al., 2019; Eisenberg et al., 2013). A critical caveat exists, however: the available structural identifiability toolboxes only allow for deterministic ODEs, although they could be used to establish proxy identifiability results for stochastic differential equation models (Browning et al., 2020). More research is needed towards developing identifiability analysis tools suited to stochastic models.

So far, we have focused on answering the question: Given a model $\mathcal{M}$, which data types can make model parameters more structurally identifiable? Our discussions have therefore originated from an output (or data type) perspective. Less attention has been paid to the influence of the rest of the model structure on the identifiability of model parameters. The

alternative question, therefore, and one that is necessary for data-limited settings, is: Which structural modifications on the system of ODEs $\dot{\boldsymbol{X}}(\boldsymbol{t})$ will improve the structural identifiability of $\mathcal{M}$? Some approaches have been suggested. One approach involves reparameterizing the model with the aim to reduce the number of parameters, concentrating particularly on identifiable combinations (Eisenberg and Hayashi, 2014; Massonis et al., 2021a, 2021b; Meshkat et al., 2014; Wieland et al., 2021). Another approach centers on simplifying model complexity by reducing the number of features/states (Massonis et al., 2021a) and another entails non-dimensionalizing (Kao and Eisenberg, 2018) or scaling some state variables (Brouwer et al., 2018; Eisenberg et al., 2013). These considerations are outside the scope of the current discussion but are important to the broader goal of developing infectious disease models for useful inference.

It is important to note that although a model may be s.n.i, it may be useful for drawing inferences, if these are limited to the structurally identifiable parameters of the model (Janzén et al., 2016; Massonis et al., 2021a). For example, with the SIR model with incidence, studied at stage one (Table 1), inference may be made on $\gamma,\ S(0)$ and $I(0)$ but not on β or $R(0)$, since β and $R(0)$ are s.n.i while $\gamma,\ S(0)$ and $I(0)$ are s.g.i.

In this work, we have demonstrated the influence of data types on structural identifiability of model parameters. A careful consideration of the type of data available for parameter estimation is therefore advised as a relevant initial step in performing inference with infectious disease transmission models.

1.5 CONCLUSIONS

We have studied the structural identifiability of parameters of various compartmental models for infectious disease transmission. We have demonstrated the influence of data types on structural identifiability by considering different data types as model outputs and examining

how structural identifiability of unknown parameters, including ICs, varied with varying outputs. The structural identifiability of some parameters varied depending on the type of model output, and single-output models were often not structurally identifiable. In general, the inclusion of additional data types as outputs improved structural identifiability of parameters. Attention ought therefore to be paid to the type(s) of observed data at hand, prior to estimating model parameters, given that data types influence a model's structural identifiability and consequently, the robustness of resulting inferences.

Acknowledgments

The authors thank Gleb Pogudin for helpful support on the implementation of SIAN.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. E.A.D. was supported by a studentship at the Department of Statistics, University of Oxford. A.F.B. was supported by the National Science Foundation (grant DMS1853032) and the National Institutes of Health (grant U01GM110712). C.A.D. was supported by joint Centre funding from the UK Medical Research Council and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union. C.A.D. was funded on grants from the UK National Institute for Health Research (NIHR) [Vaccine Efficacy Evaluation for Priority Emerging Diseases: PR-OD-1017-20007 and HPRU in Emerging and Zoonotic Infections: NIHR200907. The views expressed in this publication are those of the authors and not necessarily those of their funding institutions.

1.6 REFERENCES

Audoly, S., Bellu, G., D'Angio, L., Saccomani, M.P., Cobelli, C., 2001. Global identifiability of nonlinear models of biological systems. IEEE Trans. Biomed. Eng. 48, 55–65.

Balsa-Canto, E., Alonso, A.A., Banga, J.R., 2010. An iterative identification procedure for dynamic modeling of biochemical networks. BMC Syst. Biol. 4, 1–18.

Bellu, G., Saccomani, M.P., Audoly, S., D'Angiò, L., 2007. DAISY: A new software tool to test global identifiability of biological and physiological systems. Comput. Methods Programs Biomed. 88, 52–61.

Brouwer, A.F., Eisenberg, J.N., Pomeroy, C.D., Shulman, L.M., Hindiyeh, M., Manor, Y., Grotto, I., Koopman, J.S., Eisenberg, M.C., 2018. Epidemiology of the silent polio outbreak in Rahat, Israel, based on modeling of environmental surveillance data. Proc. Natl. Acad. Sci. 115, E10625–E10633.

Brouwer, A.F., Eisenberg, M.C., Love, N.G., Eisenberg, J.N., 2019. Phenotypic variations in persistence and infectivity between and within environmentally transmitted pathogen populations impact population-level epidemic dynamics. BMC Infect. Dis. 19, 1–13.

Brouwer, A.F., Meza, R., Eisenberg, M.C., 2017. Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis. PLoS Comput. Biol. 13, e1005431.

Browning, A.P., Warne, D.J., Burrage, K., Baker, R.E., Simpson, M.J., 2020. Identifiability analysis for stochastic differential equation models in systems biology. J. R. Soc. Interface 17, 20200652.

Capaldi, A., Behrend, S., Berman, B., Smith, J., Wright, J., Lloyd, A.L., 2012. Parameter estimation and uncertainty quantification for an epidemic model. Math. Biosci. Eng. 9, 553–576.

Castro, M., Ares, S., Cuesta, J.A., Manrubia, S., 2020. The turning point and end of an expanding epidemic cannot be precisely forecast. Proc. Natl. Acad. Sci. 117, 26190–26196.

Castro, M., de Boer, R.J., 2020. Testing structural identifiability by a simple scaling method. PLOS Comput. Biol. 16, e1008248.

Chis, O.-T., Banga, J.R., Balsa-Canto, E., 2011. Structural identifiability of systems biology models: a critical comparison of methods. PloS One 6, e27755.

Cobelli, C., Distefano III, J.J., 1980. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. Am. J. Physiol.-Regul. Integr. Comp. Physiol. 239, R7–R24.

Dankwa, E.A., Donnelly, C.A., Brouwer, A.F., Zhao, R., Montgomery, M.P., Weng, M.K., Martin, N.K., 2021. Estimating vaccination threshold and impact in the 2017–2019 hepatitis A virus outbreak among persons experiencing homelessness or who use drugs in Louisville, Kentucky, United States. Vaccine 39, 7182–7190. https://doi.org/10.1016/j.vaccine.2021.10.001

Eisenberg, M.C., Hayashi, M.A., 2014. Determining identifiable parameter combinations using subset profiling. Math. Biosci. 256, 116–126.

Eisenberg, M.C., Robertson, S.L., Tien, J.H., 2013. Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. J. Theor. Biol. 324, 84–102.

Evans, N.D., White, L.J., Chapman, M.J., Godfrey, K.R., Chappell, M.J., 2005. The structural identifiability of the susceptible infected recovered model with seasonal forcing. Math. Biosci. 194, 175–197.
Hengl, S., Kreutz, C., Timmer, J., Maiwald, T., 2007. Data-based identifiability analysis of non-linear dynamical models. Bioinformatics 23, 2612–2618.
Hong, H., Ovchinnikov, A., Pogudin, G., Yap, C., 2020. Global identifiability of differential models. Commun. Pure Appl. Math. 73, 1831–1879.
Hong, H., Ovchinnikov, A., Pogudin, G., Yap, C., 2019. SIAN: software for structural identifiability analysis of ODE models. Bioinformatics 35, 2873–2874.
Janzén, D.L., Bergenholm, L., Jirstrand, M., Parkinson, J., Yates, J., Evans, N.D., Chappell, M.J., 2016. Parameter identifiability of fundamental pharmacodynamic models. Front. Physiol. 7, 590.
Kao, Y.-H., Eisenberg, M.C., 2018. Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment. Epidemics 25, 89–100.
King, A.A., Domenech de Cellès, M., Magpantay, F.M.G., Rohani, P., 2015. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. Proc. Biol. Sci. 282, 20150347. https://doi.org/10.1098/rspb.2015.0347
Ligon, T.S., Fröhlich, F., Chiş, O.T., Banga, J.R., Balsa-Canto, E., Hasenauer, J., 2018. GenSSI 2.0: multi-experiment structural identifiability analysis of SBML models. Bioinformatics 34, 1421–1423.
Liu, Z., Magal, P., Seydi, O., Webb, G., 2020. A COVID-19 epidemic model with latency period. Infect. Dis. Model. 5, 323–337.
Ljung, L., Glad, T., 1994. On global identifiability for arbitrary model parametrizations. Automatica 30, 265–276.
Massonis, G., Banga, J.R., Villaverde, A.F., 2021a. Structural identifiability and observability of compartmental models of the COVID-19 pandemic. Annu. Rev. Control 51, 441–459.
Massonis, G., Banga, J.R., Villaverde, A.F., 2021b. AutoRepar: A method to obtain identifiable and observable reparameterizations of dynamic models with mechanistic insights. Int. J. Robust Nonlinear Control 1–19.
McCabe, R., Kont, M.D., Schmit, N., Whittaker, C., Løchen, A., Walker, P.G., Ghani, A.C., Ferguson, N.M., White, P.J., Donnelly, C.A., Watson, O.J., 2021. Communicating uncertainty in epidemic models. Epidemics 37, 100520.
Meshkat, N., Kuo, C.E., DiStefano III, J., 2014. On finding and using identifiable parameter combinations in nonlinear dynamic systems biology models and COMBOS: a novel web implementation. PLoS One 9, e110261.
Miao, H., Xia, X., Perelson, A.S., Wu, H., 2011. On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM Rev. 53, 3–39.
Muñoz-Tamayo, R., Puillet, L., Daniel, J.-B., Sauvant, D., Martin, O., Taghipoor, M., Blavy, P., 2018. To be or not to be an identifiable model. Is this a relevant question in animal science modelling? Animal 12, 701–712.
Ngwa, G.A., Shu, W.S., 2000. A mathematical model for endemic malaria with variable human and mosquito populations. Math. Comput. Model. 32, 747–763.
Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., Timmer, J., 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics 25, 1923–1929.
Roda, W.C., Varughese, M.B., Han, D., Li, M.Y., 2020. Why is it difficult to accurately predict the COVID-19 epidemic? Infect. Dis. Model. 5, 271–281.
Roosa, K., Chowell, G., 2019. Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models. Theor. Biol. Med. Model. 16, 1–15.
Scarpino, S.V., Petri, G., 2019. On the predictability of infectious disease outbreaks. Nat. Commun. 10, 1–8.

Sedoglavic, A., 2002. A probabilistic algorithm to test local algebraic observability in polynomial time. J. Symb. Comput. 33, 735–755.
Tunali, E., Tarn, T.-J., 1987. New results for identifiability of nonlinear systems. IEEE Trans. Autom. Control 32, 146–154.
Tuncer, N., Gulbudak, H., Cannataro, V.L., Martcheva, M., 2016. Structural and practical identifiability issues of immuno-epidemiological vector–host models with application to rift valley fever. Bull. Math. Biol. 78, 1796–1827.
Tuncer, N., Le, T.T., 2018. Structural and practical identifiability analysis of outbreak models. Math. Biosci. 299, 1–18.
Villaverde, A.F., 2022. Symmetries in Dynamic Models of Biological Systems: Mathematical Foundations and Implications. Symmetry 14, 467.
Villaverde, A.F., 2019. Observability and structural identifiability of nonlinear biological systems. Complexity 2019. https://doi.org/10.1155/2019/8497093
Villaverde, A.F., Barreiro, A., Papachristodoulou, A., 2016. Structural identifiability of dynamic systems biology models. PLoS Comput. Biol. 12, e1005153.
Wieland, F.-G., Hauber, A.L., Rosenblatt, M., Tönsing, C., Timmer, J., 2021. On structural and practical identifiability. Curr. Opin. Syst. Biol. 25, 60–69.

Table 1 **(stage one)**: Structural identifiability of parameters and models assuming **all parameters are unknown** and given **single model outputs**: incidence (I) or prevalence (P). For the SLIR/SLI models, outputs corresponding to the host population are annotated with "(h)". Output cells are shaded according to the structural identifiability of the model given that output: a green shade indicates the model is structurally globally identifiable (s.g.i.), a yellow shade indicates the model is structurally locally identifiable (s.l.i.) and a brown shade indicates the model is structurally non-identifiable (s.n.i.).

Model structure	**Output**	**Structural identifiability of parameters**		
		s.g.i.	**s.l.i**	**s.n.i.**
SIR	I	$\gamma, S(0), I(0)$		$\beta, N, R(0)$
	P	β, γ		$N, S(0), I(0), R(0)$
SLIR	I	$\alpha, \gamma, S(0), L(0), I(0)$		$\beta, N, R(0)$
	P	β	α, γ	$N, S(0), L(0), I(0), R(0)$
SLIRQ	I		$\alpha, \beta, \eta, \gamma, \sigma, N, S(0), L(0), I(0), R(0), Q(0)$[a]	
	P		$\alpha, \beta, \eta, \gamma, \sigma, N, S(0), L(0), I(0), R(0), Q(0)$[a]	
SLIR/SLI	I (h)		$\alpha_h, \alpha_v, \beta_h, \gamma_h, \mu_h, \mu_v, N_h, S_h(0), L_h(0), I_h(0), R_h(0)$[a]	$N_v, S_v(0), L_v(0), I_v(0), \beta_v$
	P (h)		$\alpha_h, \alpha_v, \beta_h, \gamma_h, \mu_h, \mu_v$	$\beta_v, N_h, S_h(0), L_h(0), I_h(0), R_h(0), S_v(0), L_v(0), I_v(0), N_v$

[a] Parameters are <u>at least</u> s.l.i. No results were produced for global identifiability: SIAN timed out before global identifiability calculations could be completed.

Table 2 (stage two): Structural identifiability of parameters and models assuming **all natural history parameters are known (transmission and demography parameters unknown)** and given **single model outputs**: incidence (I) or prevalence (P). For the SLIR/SLI models, outputs corresponding to the host population are annotated with "(h)". Output cells are shaded according to the structural identifiability of the model given that output: a green shade indicates the model is structurally globally identifiable (s.g.i.), a yellow shade indicates the model is structurally locally identifiable (s.l.i.) and a brown shade indicates the model is structurally non-identifiable (s.n.i.).

Model structure	Output	Structural identifiability of parameters		
		s.g.i.	**s.l.i**	**s.n.i.**
SIR	I	$S(0), I(0)$		$\beta, N, R(0)$
	P	β		$N, S(0), I(0), R(0)$
SLIR	I	$S(0), L(0), I(0)$		$\beta, N, R(0)$
	P	β		$N, S(0), L(0), I(0), R(0)$
SLIRQ	I	$\beta, N, S(0), L(0), I(0), R(0), Q(0)$		
	P	$\beta, N, S(0), L(0), I(0), R(0), Q(0)$		
SLIR/SLI	I (h)		$\beta_h, \mu_h, \mu_v, N_h, S_h(0), L_h(0), I_h(0), R_h(0)$[a]	$\beta_v, N_v, S_v(0), L_v(0), I_v(0)$
	P (h)		β_h, μ_h, μ_v[a]	$\beta_v, N_h, S_h(0), L_h(0), I_h(0), R_h(0), N_v, S_v(0), L_v(0), I_v(0)$

[a] Parameters are at least s.l.i. No results were produced for global identifiability: SIAN timed out before global identifiability calculations could be completed.

Table 3 (stage three): Structural identifiability of parameters and models assuming **all parameters are unknown** and given **multiple model outputs**: outputs are incidence (I), prevalence (P) or detected vector counts (DC). For the SLIR/SLI models, outputs corresponding to the host and vector populations are annotated with "(h)" and "(v)", respectively. Output cells are shaded according to the structural identifiability of the model given that output: a green shade indicates the model is structurally globally identifiable (s.g.i.), a yellow shade indicates the model is structurally locally identifiable (s.l.i.) and a brown shade indicates the model is structurally non-identifiable (s.n.i.).

Model structure	**Output**	**Structural identifiability of parameters**		
		s.g.i.	**s.l.i**	**s.n.i.**
SIR	I, P	$\beta, \gamma, N, S(0), I(0), R(0)$		
SLIR	I, P	$\alpha, \beta, \gamma, N, S(0), L(0), I(0), R(0)$		
SLIRQ	I, P	$\alpha, \beta, \eta, \gamma, \sigma, N, S(0), L(0), I(0), R(0), Q(0)$		
SLIR/SLI	I (h), P (h)		$\alpha_h, \alpha_v, \beta_h, \gamma_h, \mu_h, \mu_v, N_h, S_h(0), L_h(0), I_h(0), R_h(0)$[a]	$\beta_v, N_v, S_v(0), L_v(0), I_v(0)$
	I (h), P (h), I (v), DC (v)	$\alpha_h, \alpha_v, \beta_h, \beta_v, \gamma_h, \lambda_v, \mu_h, \mu_v, N_h, S_h(0), L_h(0), I_h(0), R_h(0), N_v, S_v(0), L_v(0), I_v(0)$		

[a] Parameters are <u>at least</u> s.l.i. No results were produced for global identifiability: SIAN timed out before global identifiability calculations could be completed.

Table 4 (stage four): Structural identifiability of parameters and models assuming **all natural history parameters are known (transmission and demography parameters unknown)** and given **multiple model outputs**: outputs are incidence (I), prevalence (P) or detected vector counts (DC). For the SLIR/SLI models, outputs corresponding to the host and vector populations are annotated with "(h)" and "(v)", respectively. Output cells are shaded according to the structural identifiability of the model given that output: a green shade indicates the model is structurally globally identifiable (s.g.i.), a yellow shade indicates the model is structurally locally identifiable (s.l.i.) and a brown shade indicates the model is structurally non-identifiable (s.n.i.).

Model structure	**Output**	**Structural identifiability of parameters**		
		s.g.i.	**s.l.i**	**s.n.i.**
SIR	I, P	$\beta, N, S(0), I(0), R(0)$		
SLIR	I, P	$\beta, N, S(0), L(0), I(0), R(0)$		
SLIRQ	I, P	$\beta, N, S(0), L(0), I(0), R(0), Q(0)$		
SLIR/SLI	I (h), P (h)	$\beta_h, \mu_h, \mu_v, N_h, S_h(0), L_h(0), I_h(0), R_h(0)$		$\beta_v, N_v, S_v(0), L_v(0), I_v(0)$
	I (h), P (h), I (v), DC (v)	$\beta_h, \beta_v, \lambda_v, \mu_h, \mu_v, N_h, S_h(0), L_h(0), I_h(0), R_h(0), N_v, S_v(0), L_v(0), I_v(0)$		