Identifiability of linear compartmental models of infectious disease transmission

Jeremy P D'Silva, Marisa C Eisenberg

August 27, 2018

Abstract

Mathematical models are commonly used in epidemiology for generating predictions and comparing outbreak intervention strategies. These applications often require parameter estimation, in which models are fitted to data to draw mechanistic inferences about the system. Successful parameter estimation generally requires that the estimated parameters are unique: the estimation approach should result in one set of parameter estimates for a given model and data. A necessary condition for fitting a model to real data is that the fitting approach works for perfect data. This property of unique parameter estimates for a given output (data set) is called structural identifiability: a system is globally structurally identifiable (resp. locally structurally identifiable) if the map from parameters to outputs is injective (resp. has finite fibers). In this paper, we prove that a general class of stage-structured disease models used in modeling Ebola are unidentifiable from incidence data, we completely characterize the structural identifiability for a simplified form of the model, and we conjecture the identifiability of the general model. Finally, we discuss what additional parameter measurements are needed to ensure structural identifiability.

1 Introduction

Since the introduction of the Susceptible, Infectious, Recovered (SIR) model by Kermack and McKendrick in 1927 [1], mathematical modelling has been widely applied in epidemiology. Mathematical models can provide insights into outbreak dynamics and predict the future progression of an outbreak, which makes them relevant for public health officials and policymakers [2–4].

Many compartmental models of disease transmission are extensions of the SIR model, which describes disease spread in a population with a mass-action approach [2]. In an SIR model, the population is compartmentalized into "Susceptible", "Infected", and "Recovered" individuals. The $SEIR$ model includes an "Exposed" compartment, which captures the latent period (time between exposure and infection, during which the person is asymptomatic or not transmitting the disease), a feature of many diseases.

1.1 Model structure

We consider a modification of the $SEIR$ model with n infectious compartments ("stages") and a funeral transmission compartment, F. This is inspired by stage-structured Ebola transmission models [5–7], although it is applicable to a range of stage-structured infections. Stage-structured

models can be used to model diseases where the disease progression naturally fits into stages. Examples include norovirus [8, 9] and Ebola Virus Disease infections, which can be described in two stages, where some patients recover from the milder first-stage symptoms, and others progress to the severe second-stage symptoms, in which they have a high mortality rate and potentially higher transmissibility [6, 7]. The model we consider for this paper consists of Susceptible (S), Exposed (E), n Infected stages ($I_1, \cdots, I_n$), and Funeral (F). The model is given by the following equations, where we indicate the derivative with parenthetical superscript:

$$\begin{aligned}
S^{(1)} &= -(\sum_{i=1}^{n} \beta_i I_i + \beta_F F)S \\
E^{(1)} &= (\sum_{i=1}^{n} \beta_i I_i + \beta_F F)S - \gamma_0 E \\
I_1^{(1)} &= \gamma_0 E - \gamma_1 I_1 \\
I_2^{(1)} &= \delta_1\gamma_1 I_1 - \gamma_2 I_2 \\
&\vdots \\
I_j^{(1)} &= \delta_{j-1}\gamma_{j-1} I_{j-1} - \gamma_j I_j \\
&\vdots \\
I_n^{(1)} &= \delta_{n-1}\gamma_{n-1} I_{n-1} - \gamma_n I_n \\
F^{(1)} &= \delta_n\gamma_n I_n - \gamma_F F
\end{aligned} \tag{1}$$

We consider two options for the observation structure:

$$\begin{aligned}
y_1^{(1)} &= k\gamma_0 E \\
y_2^{(1)} &= k\delta_n\gamma_n I_n
\end{aligned} \tag{2}$$

$$\begin{aligned}
y_1^{(1)} &= k(\sum_{i=1}^{n} \beta_i I_i + \beta_F F)S \\
y_2^{(1)} &= k\delta_n\gamma_n I_n
\end{aligned} \tag{3}$$

With observation Eqs. (3), y_1 essentially observes exposed persons, whereas in Eqs. (2) we observe the entrance to the first infectious stage. Either option might be realistic, depending on the disease—for example, for diseases where symptoms begin before transmissibility, one might use Eq. (3), while if symptoms begin with transmissibility, cases are more likely to be detected when they enter the infectious stage (Eq. (2)).

For Ebola, the onset of symptoms and transmission coincide, and so typically we would use Eq. (2). However, using (3) simplifies the computations and is equivalent to (2) in the case where γ_0 is known. Why is this valid? Intuitively, if γ_0 is known, then $\gamma_0 E$ is uniquely determined by $k(\sum_{i=1}^{n} \beta_i I_i + \beta_F F)S$: observing $\gamma_0 E$ is like observing $k(\sum_{i=1}^{n} \beta_i I_i + \beta_F F)S$ with some delay.
Important Note For this paper, we always assume that γ_0 is known, in order to use the simpler observation structure. The biological interpretation of this statement is that we assume the latent period for the disease being modelled is known (which is the case for Ebola, and indeed for most infectious diseases).

1.2 Definition/practical relevance of identifiability

One can think of a mathematical model as a map from its parameter space to the space of output trajectories. When fitting a model to data, we think of the data as an output trajectory, that is, a point in the space of output trajectories. The goal is to determine the point or points in parameter space that the model map sends to the given output trajectory. Thus, a natural question to ask is whether the model map is injective. This question is critical for understanding the applicability of a mathematical model: successful parameter estimation is necessary in order to develop models that reflect the underlying mechanism of a disease, which in turn allows us to make useful predictions from the model. Identifiability is a way of making this notion precise.

Definitions: A model is *structurally identifiable* provided that the map from parameter space to observations is injective: i.e. each observation trajectory is produced by a unique parameter vector. A model is *locally structurally identifiable* provided that each trajectory is produced by a finite fiber in parameter space.

Structural identifiability assumes perfect knowledge of the data/observation trajectory. While this is rarely realistic in a biological setting, it is critical to study structural identifiability because being able to fit parameters uniquely to perfect data is a necessary condition for being able to fit parameters uniquely to real data.

1.3 Techniques to study identifiability

A wide range of techniques have been developed to study structural identifiability, using both numerical and analytical tools (see reviews in [10] and [11]). Of them, methods from differential algebra have emerged as an effective and useful way to prove identifiability for broad classes of models [12, 13]. As discussed above, a model is in some sense a map from parameter space to the space of output trajectories. If an ordinary differential equation (ODE) is solvable (e.g. a linear ODE), this map can be written explicitly; however, even when the ODE is not solvable, the differential algebra approach provides tools to find an implicit form of this map, called the input-output equations (IPOPs).

For ODE models consisting of rational functions, the input-output equations of the model can be written as differential polynomials, polynomials where we allow derivatives of the variables, in the observed variables. In this case, the coefficients of input-output equations in monic form contain the necessary data to determine identifiability [14]: the coefficients themselves are the identifiable parameters or combinations of parameters. Input-output equations are obtained by eliminating hidden variables to get equations that directly encode the relationship between inputs, parameters, and outputs, without hidden variables. In this paper, we obtain input-output equations using direct substitution to eliminate hidden variables. Input-output equations can also be obtained from certain generating sets of the differential ideal generated by an ODE model (namely, characteristic sets or Grobner bases) [12, 15], often using Ritt-Kolchin pseudodivision or similar reduction methods.

1.4 Summary of paper

In Section 2, we rescale the model in Eq. (1) to demonstrate that it is unidentifiable. In Section 3, we obtain input-output equations for the rescaled model; these IPOPs are also IPOPs for the original model after un-scaling. In Section 4, we use the input-output equations to characterize the identifiable combinations of the model assuming that $\gamma_j = \gamma$ for all $j > 0$ (i.e. for

the stage-structured model with stages of equal duration). Finally, we conjecture the identifiable combinations for the unequal-γ model.

2 Unidentifiability of original model through rescaling

In this section, we rescale the model given in Eq. 1; the rescaling reduces the number of parameters. The existence of a lower-dimensional rescaling implies that the original model is unidentifiable, irrespective of the identifiability of the rescaling. Intuitively, this is because a lower-dimensional rescaling implies that the original model had "too many dimensions" in its parametrization, meaning that it had parameters that did not show up uniquely in the IPOPs; those "extra" parameters can be combined to form a model with equivalent input-output structure. The appendix contains some other arguments and intuition for this well-known result.

Intuitively, we "collapse" the δ_i into the hidden variables and divide through by N, the overall population size (assumed to be constant and nonzero), so that the hidden variables represent fractions of the total population. This is akin to the process of nondimensionalization in other physical and biological models, except here we choose the rescaling to reduce the number of parameters

rather than to result in dimensionless variables. We define:

$$
\begin{aligned}
\tilde{S} &= \left(\frac{1}{N}\right)\left(\prod_{m=1}^{n}\delta_m\right)S \\
\tilde{E} &= \left(\frac{1}{N}\right)\left(\prod_{m=1}^{n}\delta_m\right)E \\
\tilde{I}_1 &= \left(\frac{1}{N}\right)\left(\prod_{m=1}^{n}\delta_m\right)I_1 \\
\tilde{I}_2 &= \left(\frac{1}{N}\right)\left(\prod_{m=2}^{n}\delta_m\right)I_2 \\
&\vdots \\
\tilde{I}_j &= \left(\frac{1}{N}\right)\left(\prod_{m=j}^{n}\delta_m\right)I_j \\
&\vdots \\
\tilde{I}_{n-1} &= \left(\frac{1}{N}\right)\delta_{n-1}\delta_n I_{n-1} \\
\tilde{I}_n &= \left(\frac{1}{N}\right)\delta_n I_n \\
\tilde{F} &= \left(\frac{1}{N}\right)F \\
\tilde{k} &= kN \\
\delta &= \prod_{i=1}^{n}\delta_i \\
\tilde{\beta}_i &= \left(\prod_{j=1}^{i-1}\delta_j\right)N\beta_i \\
\tilde{\beta}_F &= \left(\prod_{j=1}^{n}\delta_j\right)N\beta_F
\end{aligned}
\tag{4}
$$

With this rescaling, the model equations become:

$$\begin{aligned}
\tilde{S}^{(1)} &= -\frac{\delta y_1^{(1)}}{\tilde{k}} \\
\tilde{E}^{(1)} &= \frac{\delta y_1^{(1)}}{\tilde{k}} - \gamma_0 \tilde{E} \\
\tilde{I}_1^{(1)} &= \gamma_0 \tilde{E} - \gamma_1 \tilde{I}_1 \\
\tilde{I}_2^{(1)} &= \gamma_1 \tilde{I}_1 - \gamma_2 \tilde{I}_2 \\
&\vdots \\
\tilde{I}_j^{(1)} &= \gamma_{j-1} \tilde{I}_{j-1} - \gamma_j \tilde{I}_j \\
&\vdots \\
\tilde{I}_n^{(1)} &= \gamma_{n-1} \tilde{I}_{n-1} - \gamma_n \tilde{I}_n \\
\tilde{F}^{(1)} &= \gamma_n \tilde{I}_n - \gamma_F \tilde{F} \\
y_1^{(1)} &= \frac{\tilde{k}}{\delta^2} (\sum_{i=1}^{n} \tilde{\beta}_i \tilde{I}_i + \tilde{\beta}_F \tilde{F}) \tilde{S} \\
y_2^{(1)} &= \tilde{k} \gamma_n \tilde{I}_n
\end{aligned} \tag{5}$$

3 Obtaining input-output equations

Next, we will determine the input-output equations for Eq. (5). Since the model has two observed variables, there will be two input-output equations, denoted IPOP 1 and and IPOP 2.

3.1 Computation to demonstrate the process for IPOP 1

We first obtain a closed form for $\tilde{I}_j$ in terms of $\tilde{I}_n$ by recursion from $n-1$ to j. Below is the computation; for simplicity, we let $p_m = \frac{1}{\gamma_m}$. In brief, we rearrange each equation containing $\tilde{I}_j^{(1)}$ to solve for $\tilde{I}_{j-1}$, which occurs once. We then substitute that into the equation containing $\tilde{I}_{j-1}^{(1)}$, and solve so so that we have an expression for $\tilde{I}_{j-2}$ in terms of $\tilde{I}_j$. Beginning at $\tilde{I}_n$, we clearly can proceed down to $\tilde{I}_1$ and get expressions for all $\tilde{I}_j$ in terms of $\tilde{I}_n$.

$$\begin{aligned}
\tilde{I}_{n-1} &= p_{n-1}(\gamma_n \tilde{I}_n + \tilde{I}_n^{(1)}) \\
\tilde{I}_{n-2} &= p_{n-2}(\gamma_{n-1} \tilde{I}_{n-1} + \tilde{I}_{n-1}^{(1)}) \\
&= p_{n-2}(\gamma_{n-1} p_{n-1}(\gamma_n \tilde{I}_n + \tilde{I}_n^{(1)}) + (p_{n-1}(\gamma_n \tilde{I}_n + \tilde{I}_n^{(1)}))^{(1)}) \\
&= (p_{n-2} p_{n-1})(\gamma_{n-1} \gamma_n \tilde{I}_n + (\gamma_{n-1} + \gamma_n) \tilde{I}_n^{(1)} + \tilde{I}_n^{(2)}) \\
\tilde{I}_{n-3} &= (p_{n-3} p_{n-2} p_{n-1})(\gamma_{n-2}(\gamma_{n-1} \gamma_n \tilde{I}_n + (\gamma_{n-1} + \gamma_n) \tilde{I}_n^{(1)} + \tilde{I}_n^{(2)}) + (\gamma_{n-1} \gamma_n \tilde{I}_n \\
&\quad + (\gamma_{n-1} + \gamma_n) \tilde{I}_n^{(1)} + \tilde{I}_n^{(2)})^{(1)}) \\
&= \text{(continued substitution with this pattern)}
\end{aligned}$$

The important thing to observe is that each time we move down one index in $\tilde{I}_j$, we accumulate another p and a higher-order derivative.

3.2 Proof for general form of IPOP 1

Lemma 3.1. *For a model as in Eq. 5, we can write*

$$\tilde{I}_{n-m} = \left(\prod_{i=n-m}^{n-1} p_i\right) \sum_{j=0}^{m} \tilde{I}_n^{(j)}\, C_{n-j,m} \tag{6}$$

where

$$C_{n-j,m} = \sum_{\substack{\mathcal{S}\subset\{n-m+1,\cdots,n\} \\ |\mathcal{S}|=m-j}} \left(\prod_{l\in\mathcal{S}} \gamma_l\right) \tag{7}$$

$C_{n-m,m} := 1$, *and* $p_m = \frac{1}{\gamma_m}$.

Proof. We proceed recursively over m for $m \in \mathbb{N}_{n-1}$.
$\underline{m=1}$ By rearranging the equation containing $\tilde{I}_n^{(1)}$

$$\tilde{I}_{n-1} = p_{n-1}(\gamma_n \tilde{I}_n + \tilde{I}_n^{(1)})$$

Suppose the claim holds for $m = k$ (this k is different than the parameter k in the model).
$\underline{\text{m = k+1}}$

$$\begin{aligned}
\tilde{I}_{n-k-1} &= p_{n-k-1}\left(\gamma_{n-k}\tilde{I}_{n-k} + \tilde{I}_{n-k}^{(1)}\right) \\
&= p_{n-k-1}\left(\gamma_{n-k}\left(\prod_{i=n-k}^{n-1} p_i\right)\left(\sum_{j=0}^{k} \tilde{I}_n^{(j)}\, C_{n-j,k}\right) + \left(\prod_{i=n-k}^{n-1} p_i\right)\left(\sum_{j=0}^{k} \tilde{I}_n^{(j)}\, C_{n-j,k}\right)^{(1)}\right) \\
&= p_{n-k-1}\left(\gamma_{n-k}\left(\prod_{i=n-k}^{n-1} p_i\right)\left(\sum_{j=0}^{k} \tilde{I}_n^{(j)}\, C_{n-j,k}\right) + \left(\prod_{i=n-k}^{n-1} p_i\right)\left(\sum_{j=0}^{k} \tilde{I}_n^{(j+1)}\, C_{n-j,k}\right)\right) \\
&= \left(\prod_{i=n-k-1}^{n-1} p_i\right)\left(\gamma_{n-k}C_{n,k}\tilde{I}_n + \left(\sum_{j=1}^{k} \tilde{I}_n^{(j)}\gamma_{n-k}C_{n-j,k} + C_{n-(j-1),k}\right) + \tilde{I}_n^{k+1}\right)
\end{aligned}$$

Note that $\gamma_{n-k}C_{n,k} = C_{n,k+1}$ and $1 = C_{n-(k+1),k+1}$. So we've reduced to showing

$$C_{n-j,k+1} = \gamma_{n-k}C_{n-j,k} + C_{n-(j-1),k}$$

for $j \in \mathbb{N}_k$. This holds by a combinatorial argument

$$\gamma_{n-k}C_{n-j,k} + C_{n-(j-1),k} = \sum_{\substack{\mathcal{S}\subset\{n-k+1,\cdots,n\} \\ |\mathcal{S}|=k-j}} \left(\prod_{l\in\mathcal{S}} \gamma_l\right)\gamma_{n-k} + \sum_{\substack{\mathcal{S}\subset\{n-k+1,\cdots,n\} \\ |\mathcal{S}|=k-j+1}} \left(\prod_{l\in\mathcal{S}} \gamma_l\right) \tag{8}$$

The first summation expression on the RHS counts all sets $\mathcal{S}$ of the form $\mathcal{S} \subset \{\gamma_{n-k},\cdots,\gamma_n\}$, $|\mathcal{S}| = k-(j-1) = k+1-j$, and $\gamma_{n-k} \in \mathcal{S}$. The second summation expression on the RHS counts all sets $\mathcal{S}$ of the form $\mathcal{S} \subset \{\gamma_{n-k},\cdots,\gamma_n\}$, $|\mathcal{S}| = k-(j-1) = k+1-j$, $\gamma_{n-k} \notin \mathcal{S}$. The sets $\mathcal{S} \subset \{\gamma_{n-k},\cdots,\gamma_n\}$, $|\mathcal{S}| = k+1-j$ either contain γ_{n-k} or don't contain γ_{n-k}.

Together, the summations count all sets of cardinality $\mathcal{S} = k+1-j$ that are subsets of

$\{\gamma_{n-k}, \cdots, \gamma_n\} = \{\gamma_{n-(k+1)+1}, \cdots, \gamma_n\}$. Thus, equation 3 reduces to

$$\sum_{\substack{\mathcal{S}\subset\{n-k+1,\cdots,n\}\\|\mathcal{S}|=k-j}} \left(\prod_{l\in\mathcal{S}} \gamma_l\right) \gamma_{n-k} + \sum_{\substack{S\subset\{n-k+1,\cdots,n\}\\|\mathcal{S}|=k-j+1}} \left(\prod_{l\in\mathcal{S}} \gamma_l\right) \tag{9}$$

$$= \sum_{\substack{\mathcal{S}\subset\{n-(k+1)+1,\cdots,n\}\\|\mathcal{S}|=k+1-j}} \left(\prod_{l\in\mathcal{S}} \gamma_l\right) \tag{10}$$

$$= C_{n-j,k+1} \tag{11}$$

□

Corollary 3.1. *IPOP 1 in monic form is given by:*

$$0 = -\delta \left(\prod_{i=0}^{n} \gamma_i\right) y_1^{(1)} + \left(\sum_{j=0}^{n} y_2^{(1+j)} C_{n-j,n+1}\right) + y_2^{(n+2)} \tag{12}$$

Proof. We obtain a formula for E in terms of I_n (using Lemma 3.1), substitute this into the equation containing $E^{(1)}$, and substitute $\tilde{I}_n = \dfrac{y_2^{(1)}}{\gamma_n \tilde{k}}$. Finally, we multiply through by $\tilde{k}$ $(\prod_{i=0}^{n} \gamma_i)$ to clear the coefficient of the leader; this gives us monic form. □

3.3 Obtaining IPOP 2

IPOP 2 is obtained by the following process:

1. Solve the equation containing $\tilde{S}^{(1)}$ and integrate it to get expressions for $\tilde{S}$ and $\tilde{S}^{(1)}$ in terms of $y_1^{(1)}$ and y_1, respectively.
2. Substitute those expressions into the first observation equation to solve for $\tilde{F}$ in terms of observed variables
3. Substitute the expression for $\tilde{F}$ in terms of y_1, y_2, and derivatives and the expression for $\tilde{I}_n$ in terms of y_2 into the equation containing $\tilde{F}^{(1)}$.
4. After simplification, this is an IPOP.

Note In the integration in step 1, the constant of integration is in fact an initial condition for the model. Here, we are concerned with the output's dependency on the parameters (and not the initial conditions), so as long as the initial conditions are chosen to be a "reasonable" point (in this case, positive), the choice of initial condition is arbitrary. Thus, we used 1. Note also that we are assuming we know the value of this initial condition, which could affect the identifiability of the parameters—more parameters might be identifiable with the known initial condition than without it. Nonetheless, it's a reasonable assumption to say we know the initial condition, since we often assume a small number of infected individuals at the start of the epidemic and set $S(0) = 1 - I(0)$. If we wanted to check the identifiablility of the initial condition, we could have treated it as another parameter in the model. The standard differential algebra approach would be to differentiate one polynomial with S to get a polynomial containing S', whereupon we could substitute our expression for S'.

Lemma 3.2. *IPOP 2 is given by*

$$0 = \frac{\prod_{i=1}^{n}(\gamma_i)\,\tilde{k}^2}{\tilde{\beta}_1\delta^2}\left(\delta y_1^{(1)}\left(-\frac{\delta}{\tilde{k}}y_1^{(1)}\right) - \left(y_1^{(1)}+y_1^{(2)}\right)\delta\left(1-\frac{\delta}{\tilde{k}}y_1\right) + \tilde{k}\left(1-\frac{\delta}{\tilde{k}}y_1\right)^2\left(\heartsuit+\heartsuit^{(1)}\right)+\beta_F y_2^{(1)}\left(1-\frac{\delta}{\tilde{k}}y_1\right)^2\right) \tag{13}$$

where

$$\heartsuit+\heartsuit^{(1)} = \frac{y_2^{(1)}}{\tilde{k}\gamma_n}\left(\sum_{i=1}^{n}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_j}\right)\tilde{\beta}_i C_{n,n-i}\right) + \sum_{l=1}^{n-1}\frac{y_2^{(l+1)}}{\tilde{k}\gamma_n}\left(\sum_{i=1}^{n-l}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_i}\right)\tilde{\beta}_i C_{n-l,n-i}\right) \tag{14}$$

$$+\sum_{l=1}^{n-1}\frac{y_2^{(l+1)}}{\tilde{k}\gamma_n}\left(\sum_{i=1}^{n-l+1}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_i}\right)\tilde{\beta}_i C_{n-l+1,n-i}\right) \tag{15}$$

$$+\frac{y_2^{(n+1)}}{\tilde{k}\gamma_n}\left(\tilde{\beta}_1\left(\prod_{i=1}^{n-1}\frac{1}{\gamma_i}\right)\right) \tag{16}$$

Proof. This proof follows the process given above. We first introduce some notation:

$$\heartsuit = \sum_{i=1}^{n}\tilde{\beta}_i\tilde{I}_i \tag{17}$$

so the first observation equation is written

$$0 = -y_1^{(1)} + \frac{\tilde{k}}{\delta}(\heartsuit + \tilde{\beta}_F\tilde{F})\tilde{S} \tag{18}$$

Following the process above, we obtain some identities; this integration step is where the initial condition choice occurs (discussed in the note above).

$$\begin{aligned}\tilde{S}^{(1)} &= -\frac{\delta y_1^{(1)}}{\tilde{k}}\\ \tilde{S} &= 1-\frac{\delta y_1}{\tilde{k}}\end{aligned} \tag{19}$$

We rearrange the first output equation to solve for $\tilde{F}$:

$$\tilde{F} = \frac{\delta\tilde{y_1}^{(1)} - \tilde{k}\heartsuit\tilde{S}}{\tilde{k}\tilde{\beta}_F\tilde{S}} \tag{20}$$

Naming the numerator and denominator f and g, respectively, and applying the quotient rule, we get

$$\begin{aligned}f &= \delta y_1^{(1)} - \tilde{k}\heartsuit\tilde{S}\\ g &= \tilde{k}\tilde{\beta}_F\tilde{S}\\ \tilde{F}^{(1)} &= \frac{gf'-fg'}{g^2}\end{aligned} \tag{21}$$

Substituting into the equation for $\tilde{F}^{(1)}$, we obtain

$$0 = -\frac{gf'-fg'}{g^2} + \gamma_n\tilde{I}_n - \frac{f}{g} \tag{22}$$

After clearing the denominator, we obtain

$$0 = fg' - gf' + g^2\gamma_n\tilde{I}_n - gf \tag{23}$$

Substituting in f and g, we get

$$\begin{aligned} 0 = &\left(\delta y_1^{(1)} - \tilde{k}\heartsuit\tilde{S}\right)\left(\tilde{k}\tilde{\beta}_F\tilde{S}^{(1)}\right) - \left(\tilde{k}\tilde{\beta}_F\tilde{S}\right)\left(\delta y_1^{(2)} - \tilde{k}\heartsuit^{(1)}\tilde{S} - \tilde{k}\heartsuit\tilde{S}^{(1)}\right) \\ &+ \left(\tilde{k}\tilde{\beta}_F\tilde{S}\right)^2\frac{y_2^{(1)}}{\tilde{k}} - \left(\tilde{k}\tilde{\beta}_F\tilde{S}\right)\left(\delta y_1^{(1)} - k\heartsuit\tilde{S}\right) \end{aligned} \tag{24}$$

We first note that the $\tilde{k}^2\tilde{\beta}_F\heartsuit\tilde{S}\tilde{S}^{(1)}$ parts of the first and second terms cancel out. We eliminate them, and dividing through by $\tilde{k}\tilde{\beta}_F$, we get

$$0 = \left(\delta y_1^{(1)}\right)\tilde{S}^{(1)} - \tilde{S}\left(\delta y_1^{(2)} - \tilde{k}\heartsuit^{(1)}\tilde{S}\right) + \tilde{k}\tilde{\beta}_F\tilde{S}^2\frac{y_2^{(1)}}{\tilde{k}} - \tilde{S}\left(\delta y_1^{(1)} - \tilde{k}\heartsuit\tilde{S}\right) \tag{25}$$

Factoring, we obtain

$$0 = \delta y_1^{(1)}\tilde{S}^{(1)} - \left(y_1^{(1)} + y_1^{(2)}\right)\delta\tilde{S} + \tilde{k}\tilde{S}^2\left(\heartsuit + \heartsuit^{(1)}\right) + \tilde{\beta}_F y_2^{(1)}\tilde{S}^2 \tag{26}$$

Substituting in $\tilde{S}$ and $\tilde{S}^{(1)}$ from above, and rearranging terms a bit, this looks like

$$\begin{aligned} 0 = &\delta y_1^{(1)}\left(-\frac{\delta}{\tilde{k}}y_1^{(1)}\right) - \left(y_1^{(1)} + y_1^{(2)}\right)\delta\left(1 - \frac{\delta}{\tilde{k}}y_1\right) + \\ &\tilde{k}\left(1 - \frac{\delta}{\tilde{k}}y_1\right)^2\left(\heartsuit + \heartsuit^{(1)}\right) + \beta_F y_2^{(1)}\left(1 - \frac{\delta}{\tilde{k}}y_1\right)^2 \end{aligned} \tag{27}$$

In order for this to be an input-output equation, we must write out the explicit form of $\heartsuit + \heartsuit^{(1)}$ in terms of output equations. We note that $\heartsuit$ consists of monomial terms of I_j for $j \in \mathbb{N}_n$; by Lemma 1, I_j can be written as a differential polynomial of I_n; substituting the expression from Lemma 1, we obtain

$$\begin{aligned} \heartsuit_n &= \sum_{i=1}^{n}\tilde{\beta}_i\tilde{I}_i = \sum_{i=1}^{n}\tilde{\beta}_i\tilde{I}_{n-(n-i)} \\ &= \sum_{i=1}^{n}\tilde{\beta}_i\left(\prod_{j=1}^{n-1}\frac{1}{\gamma_i}\right)\left(\sum_{l=0}^{n-i}\tilde{I}_n^{(l)}C_{n-l,n-i}\right) \\ &= \sum_{l=0}^{n-1}\tilde{I}_n^{(l)}\left(\sum_{i=1}^{n-l}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_i}\right)\tilde{\beta}_i C_{n-l,n-i}\right) \end{aligned} \tag{28}$$

With this, we can write

$$\begin{aligned} \heartsuit + \heartsuit^{(1)} = &\tilde{I}_n\left(\sum_{i=1}^{n}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_j}\right)\tilde{\beta}_i C_{n,n-i}\right) + \sum_{l=1}^{n-1}\tilde{I}_n^{(l)}\left(\sum_{i=1}^{n-l}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_i}\right)\tilde{\beta}_i C_{n-l,n-i}\right) \\ &+ \sum_{l=0}^{n-2}\tilde{I}_n^{(l+1)}\left(\sum_{i=1}^{n-l}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_i}\right)\tilde{\beta}_i C_{n-l,n-i}\right) \\ &+ \tilde{I}_n^{(n)}\left(\tilde{\beta}_1\left(\prod_{i=1}^{n-1}\frac{1}{\gamma_i}\right)\right) \end{aligned} \tag{29}$$

The second line of this equation becomes, after shifting indices of the summation,

$$\sum_{l=1}^{n-1}\tilde{I}_n^{(l)}\left(\sum_{i=1}^{n-l+1}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_i}\right)\tilde{\beta}_i C_{n-l+1,n-i}\right)$$

After substituting in an expression for $\tilde{I}_n$ in terms of observed variables, we obtain

$$\begin{aligned}\heartsuit + \heartsuit^{(1)} = &\frac{y_2^{(1)}}{\tilde{k}\gamma_n}\left(\sum_{i=1}^{n}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_j}\right)\tilde{\beta}_i C_{n,n-i}\right) + \sum_{l=1}^{n-1}\frac{y_2^{(l+1)}}{\tilde{k}\gamma_n}\left(\sum_{i=1}^{n-l}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_i}\right)\tilde{\beta}_i C_{n-l,n-i}\right)\\ &+\sum_{l=1}^{n-1}\frac{y_2^{(l+1)}}{\tilde{k}\gamma_n}\left(\sum_{i=1}^{n-l+1}\left(\prod_{j=i}^{n-1}\frac{1}{\gamma_i}\right)\tilde{\beta}_i C_{n-l+1,n-i}\right)\\ &+\frac{y_2^{(n+1)}}{\tilde{k}\gamma_n}\left(\tilde{\beta}_1\left(\prod_{i=1}^{n-1}\frac{1}{\gamma_i}\right)\right)\end{aligned} \tag{30}$$

The leader of 27 is the leader of $\heartsuit^{(1)}$ times ${\tilde{y_1}}^2$ with some coefficients. The coefficient of this term is $\tilde{\beta}_1\left(\frac{\delta}{\tilde{k}}\right)^2\frac{1}{\left(\prod_{i=1}^{n}\gamma_i\right)}$. After dividing through 27 by these coefficients, it is an IPOP in monic form. □

4 Complete identifiability information of simplified model

We make the simplification that all infectious stages have the same length, that the time spent in each I_j compartment is equal. In the model, this means that we set $\gamma_i = \gamma$ for all $i \in \mathbb{N}_n$, where γ is the inverse stage length. This is an assumption used for various diseases; it is applicable if the stages are known to be of equal length, or if the discretized stage structure is being used to approximate a continuously varying infection. This assumption often arises when assuming a multistage structure without a knowledge of the lengths of individual stages, or when approximating a gamma distributino for length of the infectious period [16–18].

4.1 Proof of identifiability for simplified model

Theorem 4.1. *If we assume that $\gamma_i = \gamma$ for all $i \in \mathbb{N}_n$, then the rescaled model given in Eq. (5) is identifiable.*

Proof. Recall that the coefficients of the IPOPs in monic form are the identifiable parameters or combinations of the model. Thus, we will rewrite the IPOPs (obtained in Corollary 3.1 and Lemma 3.2, respectively) using our simplifying assumption, and read off the identifiable combinations or parameters.

First, the $C_{a,b}$ will become binomial coefficients: instead of each term of the sum acquiring distinct γ_i, the sum will now accumulate a certain number of γ to an exponent (specifically, the coefficient of γ is the number of ways of choosing the sets S in the definition of $C_{a,b}$ in Lemma

3.1, and the exponent is the cardinality of those sets).

$$\begin{aligned}
C_{n,n-i} &= \sum_{\substack{S\subset\{i+1,\cdots,n\}\\|S|=n-i}} \left(\prod_{l\in S}\gamma\right) = \gamma^{n-i}\\
C_{n-l,n-i} &= \sum_{\substack{S\subset\{i+1,\cdots,n\}\\|S|=n-i-l}} \left(\prod_{m\in S}\gamma_m\right) = \binom{n-i}{n-i-l}\gamma^{n-i-l}\\
C_{n-l+1,n-i} &= \sum_{\substack{S\subset\{i+1,\cdots,n\}\\|S|=n-i-l+1}} \left(\prod_{m\in S}\gamma_m\right) = \binom{n-i}{n-i-l+1}\gamma^{n-i-l+1}\\
C_{n-j,n+1} &= \sum_{\substack{S\subset\{0,\cdots,n\}\\|S|=n+1-j}} \left(\prod_{m\in S}\gamma_m\right) = \binom{n}{n-j}\gamma^{n-j} + \binom{n}{n-j-1}\gamma_0\gamma^{n-j-1}
\end{aligned} \tag{31}$$

With the above coefficient expression, IPOP 1, obtained in Corollary 3.1, becomes

$$0 = y_1^{(1)}\delta\gamma^{n+1} - \sum_{j=0}^{n}\left(\binom{n}{n-j}\gamma^{n-j} + \binom{n}{n-j-1}\gamma_0\gamma^{n-j-1}\right)y_2^{(1+j)} - y_2^{(n+2)} \tag{32}$$

From IPOP 1, we can see that the coefficient of $y_2^{(n+1)}$ is $n\gamma+\gamma_0$. Since γ_0 is assumed to be known, γ is identifiable.

With the above coefficient expressions, $\heartsuit + \heartsuit^{(1)}$, defined in Lemma 3.2, becomes

$$\begin{aligned}
\heartsuit + \heartsuit^{(1)} = {} & \frac{y_2^{(1)}}{\tilde{k}\gamma_n}\left(\sum_{i=1}^{n}\gamma\tilde{\beta}_i\right) + \sum_{l=1}^{n-1}\frac{y_2^{(l+1)}}{\tilde{k}\gamma_n}\left(\sum_{i=1}^{n-l}\frac{1}{\gamma^{l-1}}\tilde{\beta}_i\binom{n-i}{n-i-l} + \sum_{i=1}^{n-l+1}\frac{1}{\gamma^{l}}\tilde{\beta}_i\binom{n-i}{n-i-l+1}\right)\\
& + \frac{y_2^{(n+1)}}{\tilde{k}\gamma_n}\left(\tilde{\beta}_1\frac{1}{\gamma^{n-1}}\right)
\end{aligned} \tag{33}$$

In IPOP 2 (Lemma 3.2), the coefficients independent of $\heartsuit + \heartsuit^{(1)}$ are:

$$\begin{pmatrix}
\frac{\gamma^n\tilde{k}}{\tilde{\beta}_1}\\
\frac{\gamma^n\tilde{k}^2}{\tilde{\beta}_1\delta}\\
\frac{\gamma^n\tilde{k}}{\tilde{\beta}_1\delta}
\end{pmatrix} \tag{34}$$

Index these terms by a through c, and note that they are identifiable combinations (as a property of the differential algebra method). Performing an injectivity test (i.e. supposing that evaluation of $(a,b,c,)$ is equivalent to evaluation of (a',b',c') under the model map): substituting $c=c'$ into $b=b'$ implies that $\tilde{k}$ is identifiable. Rewriting the equality $a=a'$ as $\frac{a}{c}=\frac{a'}{c'}$ means that $\frac{\delta}{\tilde{k}}$ is identifiable, but $\tilde{k}$ is known, hence δ is identifiable. Since γ and $\tilde{k}$ are identifiable, we can divide by them in $a=a'$ to get $\tilde{\beta}_1$.

Note that $\left(1-\frac{\delta}{\tilde{k}}y_1\right)$ is multiplied by $\heartsuit+\heartsuit^{(1)}$, but since δ and $\tilde{k}$ are known, we can divide by them and look at the coefficients of $\heartsuit+\heartsuit^{(1)}$. Since γ is identifiable from IPOP 1, from the coefficients of $\heartsuit+\heartsuit^{(1)}$, we can obtain all the β in a recursive process as follows: decrement l from $n-1$ to 1, and each l is a linear combination of $\tilde{\beta}_1, \tilde{\beta}_2, \cdots \tilde{\beta}_{n-l+1}$. From $l-1$, β_{n-l} is known, so the linear combination can be solved for $\beta n - l + 1$. For example, when $l = n-1$, the term is a linear combination of $\tilde{\beta}_1$ and $\tilde{\beta}_2$, but $\tilde{\beta}_1$ is known, so we can solve for $\tilde{\beta}_2$ (i.e. by substitution of $\tilde{\beta}_1, \gamma$ into the identifiable combination, we find that $\tilde{\beta}_2$ is injective). This gives us $\tilde{\beta}_2, \cdots, \tilde{\beta}_n$.

Finally, the very first term of $\left(\heartsuit+\heartsuit^{(1)}\right)$ has the power $y_2^{(1)}$; note that the $\tilde{\beta}_F$ term in IPOP 2 (13) also has a power of $y_2^{(1)}$. Thus, in addition to being a linear combination of β_1 through β_n, the monomials of $y_2^{(1)}\left(1-\frac{\delta}{\tilde{k}}y_1\right)^2$ also accumulate a β_F in their coefficients. But $\tilde{\beta}_1, \cdots, \tilde{\beta}_n$ are identifiable, so βF is also identifiable by the same argument.

So we have shown that this rescaled model, with $\gamma_i = \gamma$ for $1 \leq i \leq n$, is identifiable. □

We now examine the identifiable combinations of the non-rescaled model (Eq. 5).

Corollary 4.1. *In the non-rescaled model given in Eqs.* (1) *and* (3)*, the identifable combinations are given by* kN, $\prod_{i=1}^{n}\delta_i$, $\left(\prod_{j=1}^{i-1}\delta_j\right)N\beta_i$, *and* $N\beta_F$.

Proof. Because the rescaled model is identifiable (Theorem 4.1), the rescaling function's components are identifiable combinations, given by:

$$\begin{aligned}\tilde{k} &= kN\\ \delta &= \prod_{i=1}^{n}\delta_i\\ \tilde{\beta}_i &= \left(\prod_{j=1}^{i-1}\delta_j\right)N\beta_i\\ \tilde{\beta}_F &= \left(\prod_{j=1}^{n}\delta_j\right)N\beta_F\end{aligned} \tag{35}$$

Since $\tilde{\beta}_F = \delta N\beta_F$ and δ is known, we can obtain $N\beta_F$ separately from the $\tilde{\beta}_F$ combination, yielding the desired identifiable combinations.

□

4.2 Interpretation of identifiability result (Corollary 4.1)

The identifiable combination kN is quite intuitive. The parameter k represents the "reporting rate", i.e., the fraction of cases and deaths that is measured; N is the effective population size. Thus, kN being an identifiable combination means that the product of the reporting rate and the effective population is an identifiable combination.

As an example: given 100 cases in a region with $1,000,000$ people, either an effective outbreak population is $N = 1,000,000$ with a reporting rate of $k = 0.1$ (10 percent), or an effective outbreak population of $N = 100,000$ with perfect reporting $k = 1$ would result in the same model trajectory, because 100 cases in $100,000$ is equivalent to 1000 cases in $1,000,000$ in a mass-action model. In particular, the first parameter set corresponds to an outbreak in which the entire population is at

risk, but reporting is a fraction, so the actual number of cases is larger than reported. The second parameter combination corresponds to an outbreak where only a fraction of the population is at risk, but there is perfect reporting, so the actual number of cases is exactly the reported number.

The intuition for δ (the product of the "worsening" or "disease progression" rates from each stage) being identifiable from incidence data, but not the individual δ being identifiable, is that incidence data has observation of cases, at the beginning of $\tilde{I}_1$ and observation of deaths, at the end of $\tilde{I}_n$. Thus, we can roughly determine the total number of people that recover, because they count towards the observation of cases but not deaths. However, fraction $1-\delta_j$ of patients in stage I_j recover for all j, and the incidence observation, i.e. the difference between cases and deaths, cannot distinguish between patients who recover at different stages of infection. Thus, we can get the total recovery rate (equivalent to 1 minus the death rate), but we cannot get the individual recovery rates, so we can identify δ, but not the δ_j.

Similarly, the β_i are not identifiable, but what is identifiable is $N\beta_i$ times the previous $i-1$ δ_j. Because incidence data does not allow us to determine how many infected persons enter each I_j stage, we cannot separate the transmission from the jth stage from the fraction of infected persons who enter that stage. However, if the δ_j are known, then we can estimate $\beta_i N$, and determine the relative balance of transmission in each stage. This is practically applicable because it gives an understanding of the transmission rates for each stage of the infection; this can allow epidemiologists to identify the stages that are contributing the most to transmission.

Perhaps the most surprising result is that $N\beta_F$ is identifiable. This means that the product of funeral transmission and effective population can be determined from perfect observation. This is very practical because it offers a way to quantify the amount of funeral transmission (up to population size), even if individual δ_i are not known. If individual δ_i are known, then $N\beta_i$ can be found, so that the relative contribution of I_i transmission and F transmission could be determined. As discussed above, understanding the infectiousness of various stages can help inform quarantine policies by identifying the stages at which most transmission occurs.

5 Conjecture (Identifiability of general model)

Conjecture 5.1. *For the $SEI\cdots IF$ model with no relationship assumed between the γ_i, and with γ_0 known, we conjecture that the model is locally structurally identifiable: the cardinality of the fiber is $n!$, and each element in the fiber represents a different permutation of $(\gamma_1,\gamma_2,\cdots,\gamma_n)$ (an element of S_n), along with the necessary adjustments of $\tilde{\beta}_i$ to compensate for the permutation of the γ_j. Furthermore, we conjecture that $\tilde{k}$, δ, and $\tilde{\beta}_F$ are identifiable, and that the γ_i are identifiable up to ordering; that is, the values of the elements of the set $\{\gamma_0,\gamma_1,\gamma_2,\cdots,\gamma_n\}$ are "identifiable up to ordering."*

5.1 Computations for small n

For $n=2$, there are six possible parameter vectors that give equivalent trajectories, obtained by performing an injectivity test on the input-output equations in the computer algebra environment

Mathematica. They are:

$$\begin{aligned}
&\left(\tilde{k}, \gamma_0, \gamma_1, \gamma_2, \gamma_F, \delta, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_F\right)\\
&\left(\tilde{k}, \gamma_0, \gamma_2, \gamma_1, \gamma_F, \delta, \tilde{\beta}_1, \frac{-\tilde{\beta}_1\gamma_1 + \tilde{\beta}_2\gamma_1 + \tilde{\beta}_1\gamma 2}{\gamma_2}, \tilde{\beta}_F\right)\\
&\left(\tilde{k}, \gamma_1, \gamma_0, \gamma_2, \gamma_F, \delta, \frac{\gamma_0\tilde{\beta}_1}{\gamma_1}, \tilde{\beta}_2, \tilde{\beta}_F\right)\\
&\left(\tilde{k}, \gamma_1, \gamma_2, \gamma_0, \gamma_F, \delta, \frac{\gamma_0\tilde{\beta}_1}{\gamma_1}, \frac{-\gamma_0^2\tilde{\beta}_1 + \gamma_0\gamma_1\tilde{\beta}_2 + \gamma_0\gamma_2\tilde{\beta}_1}{\gamma_1\gamma_2}, \tilde{\beta}_F\right)\\
&\left(\tilde{k}, \gamma_2, \gamma_0, \gamma_1, \gamma_F, \delta, \frac{\gamma_0\tilde{\beta}_1}{\gamma_2}, \frac{-\gamma_1\tilde{\beta}_1 + \gamma_1\tilde{\beta}_2 + \gamma_2\tilde{\beta}_1}{\gamma_2}, \tilde{\beta}_F\right)\\
&\left(\tilde{k}, \gamma_2, \gamma_1, \gamma_0, \gamma_F, \delta, \frac{\gamma_0\tilde{\beta}_1}{\gamma_2}, \frac{-\gamma_0^2\tilde{\beta}_1 + \gamma_0\gamma_1\tilde{\beta}_2 + \gamma_0\gamma_2\tilde{\beta}_1}{\gamma_1\gamma_2}, \tilde{\beta}_F\right)
\end{aligned} \tag{36}$$

Since γ_0 is known, we can eliminate four of the parameter vectors, so there are 2 possible parameter vectors, one that corresponds to preserving the order of γ_1, γ_2 and the other switching the order with a compensation of $\tilde{\beta}_i$.

We performed the same computation for $n = 3$, and found 24 parameter vector, which again consist of permutations of $\gamma_0, \cdots, \gamma_3$ along with adjustments to the $\tilde{\beta}_i$. Since γ_0 is known, we eliminate all the permutations that switch γ_0, and find that there are 6 possible parameter vectors, each of which corresponds to a permutation in S_3 of $\{\gamma_1, \gamma_2, \gamma_3\}$.

5.2 Partial results for general model

We prove a few partial results for the identifiability of the general model.

Lemma 5.1. *1. IPOP 1 is insensitive to the order of γ.*

2. The product $\prod_{i=1}^n \gamma_i$ and the parameter (combination) δ is identifiable.

3. The parameters $\tilde{k}$, δ, and β_1 are identifiable.

Proof. 1. Each of the coefficients in IPOP 1 is a sum over the products within all (fixed-size) subsets of $\{\gamma_0, \cdots, \gamma_n\}$ (Lemma 3.1, Corollary 3.1). Thus, switching any γ_i and γ_j does not change the total number of times γ_i or γ_j appears, nor does it change the coefficients by which γ_i or γ_j are multiplied.

2. We note that from the summation term in IPOP 1, when $j = 0$, the coefficient is the product $\gamma_0\gamma_1\cdots\gamma_n$. That coefficient is identifiable, i.e. if we are performing an injectivity test, $\gamma_0\gamma_1\cdots\gamma_n = g_0g_1\cdots g_n$. Further, we assumed γ_0 is known, so the identifiable combination is $\gamma_1\cdots\gamma_n$. Then in the test for the coefficient of $y_1^{(1)}$, $\delta\gamma_0\gamma_1\cdots\gamma_n = dg_0g_1\cdots g_n$, but we can substitute $\gamma_0\gamma_1\cdots\gamma_n = g_0g_1\cdots g_n$ into the expression and cancel to see that $\delta = d$, i.e., δ is identifiable.

3. By an almost identical argument to the one in Theorem 4.1 (Eq. 34 and subsequent argument), replacing γ^n by $\prod_{i=1}^n \gamma_i$, $\tilde{k}$, δ, and $\tilde{\beta}_1$ are identifiable.

□

With Lemma 5.1, we have proven part of the conjecture; what remains is to prove that all permutations of $\gamma_1, \cdots, \gamma_n$ are present in a fiber of an observation trajectory, and that no other points are present in the fiber.

6 Conclusion

In this paper, we have demonstrated the unidentifiability of n-infectious-stage compartmental disease models with a funeral compartment (Section 2). We have proven the complete identifiability information for a reduced case of the model, with infectious stages of equal duration (Theorem 4.1). Finally, we have obtained partial results for the identifiability of the general model (Lemma 5.1), and conjectured that the most general model is locally identifiable, with a particular structure (Conjecture 5.1).

References

[1] and, "A contribution to the mathematical theory of epidemics," Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 115, no. 772, pp. 700–721, 1927.

[2] L. J. Allen, F. Brauer, P. Van den Driessche, and J. Wu, Mathematical epidemiology, vol. 1945. Springer, 2008.

[3] W. E. R. Team, "Ebola virus disease in west africathe first 9 months of the epidemic and forward projections," New England Journal of Medicine, vol. 371, no. 16, pp. 1481–1495, 2014.

[4] E. T. Lofgren, M. E. Halloran, C. M. Rivers, J. M. Drake, T. C. Porco, B. Lewis, W. Yang, A. Vespignani, J. Shaman, J. N. Eisenberg, et al., "Opinion: Mathematical models: A key tool for outbreak response," Proceedings of the National Academy of Sciences, vol. 111, no. 51, pp. 18095–18096, 2014.

[5] M. A. Hayashi and M. C. Eisenberg, "Changing burial practices explain temporal trends in the 2014 ebola outbreak," arXiv preprint arXiv:1709.07319, 2017.

[6] J. P. DSilva and M. C. Eisenberg, "Modeling spatial invasion of ebola in west africa," Journal of theoretical biology, vol. 428, pp. 65–75, 2017.

[7] M. C. Eisenberg, J. N. Eisenberg, J. P. D'Silva, E. V. Wells, S. Cherng, Y.-H. Kao, and R. Meza, "Forecasting and uncertainty in modeling the 2014-2015 ebola epidemic in west africa," arXiv preprint arXiv:1501.05555, 2015.

[8] B. Lopman, K. Simmons, M. Gambhir, J. Vinjé, and U. Parashar, "Epidemiologic implications of asymptomatic reinfection: a mathematical modeling study of norovirus," American journal of epidemiology, vol. 179, no. 4, pp. 507–512, 2013.

[9] K. Simmons, M. Gambhir, J. Leon, and B. Lopman, "Duration of immunity to norovirus gastroenteritis," Emerging infectious diseases, vol. 19, no. 8, p. 1260, 2013.

[10] O.-T. Chis, J. R. Banga, and E. Balsa-Canto, "Structural identifiability of systems biology models: a critical comparison of methods," PloS one, vol. 6, no. 11, p. e27755, 2011.

[11] H. Miao, X. Xia, A. S. Perelson, and H. Wu, "On identifiability of nonlinear ode models and applications in viral dynamics," SIAM review, vol. 53, no. 1, pp. 3–39, 2011.

[12] S. Audoly, G. Bellu, L. D'Angio, M. P. Saccomani, and C. Cobelli, "Global identifiability of nonlinear models of biological systems," IEEE Transactions on biomedical engineering, vol. 48, no. 1, pp. 55–65, 2001.

[13] G. Bellu, M. P. Saccomani, S. Audoly, and L. DAngiò, "Daisy: A new software tool to test global identifiability of biological and physiological systems," Computer methods and programs in biomedicine, vol. 88, no. 1, pp. 52–61, 2007.

[14] M. Eisenberg, "Generalizing the differential algebra approach to input-output equations in structural identifiability," arXiv preprint arXiv:1302.5484, 2013.

[15] N. Meshkat, M. Eisenberg, and J. J. DiStefano III, "An algorithm for finding globally identifiable parameter combinations of nonlinear ode models using gröbner bases," Mathematical biosciences, vol. 222, no. 2, pp. 61–72, 2009.

[16] Z. Feng, Y. Zheng, N. Hernandez-Ceron, H. Zhao, J. W. Glasser, and A. N. Hill, "Mathematical models of ebola—consequences of underlying assumptions," Mathematical biosciences, vol. 277, pp. 89–107, 2016.

[17] X. Wang, Y. Shi, Z. Feng, and J. Cui, "Evaluations of interventions using mathematical models with exponential and non-exponential distributions for disease stages: The case of ebola," Bulletin of mathematical biology, vol. 79, no. 9, pp. 2149–2173, 2017.

[18] N. Hernandez-Ceron, Z. Feng, and P. Van den Driessche, "Reproduction numbers for discrete-time epidemic models with arbitrary stage distributions," Journal of Difference Equations and Applications, vol. 19, no. 10, pp. 1671–1693, 2013.

7 Appendix 1

7.1 Rescaling yields an equivalent identifiability structure

This is a well-known result in the literature, but we decided to include a few justifications here for readers' convenience.

Little Lemma Consider a dynamical system given by

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), \theta) \\ y(t) &= h(x(t), \theta) \end{aligned} \tag{37}$$

Define $\Phi(\theta)$ to be evaluation of the system $(\dot{x}, y)$ with the parameters θ for all initial conditions (i.e. the solution space of $(\dot{x}, y)$ given parameters θ). Suppose we can write $\theta^* = q(\theta)$ for $q : \mathbb{R}^n \to \mathbb{R}^m$, $m < n$, and the system $(\dot{x}^*, y^*)$, defined by

$$\begin{aligned} \dot{x}(t) &= f^*(t, x^*(t), u(t), \theta^*) \\ y(t) &= h^*(x^*(t), u(t), \theta^*) \end{aligned} \tag{38}$$

is equivalent to $(\dot{x}, y)$. By equivalence, we mean that for all θ, the evaluations are equal: $\Phi(\theta) = \Phi^*(q(\theta^*))$ for all equivalent pairs of initial conditions.

In this case, $(\dot{x}^*, y^*)$ and (x, y) have equivalent identifiability. In particular, if $(\dot{x}^*, y^*)$ is identifiable, then (x, y) is unidentifiable with identifiable combinations given by $q(\theta)$.

Proof: We first require that the system $(\dot{x}, y)$ is reduced; that is, that for each $p_i \in \theta$, the system $(\dot{x}, y)$ is sensitive to p_i.

If $(\dot{x}^*, y^*)$ is identifiable, then by the definition of equivalence, $q(\theta)$ is identifiable from (x, y). Namely, estimating θ^* from perfect observation of y^* is equivalent to estimating $q(\theta)$ from perfect observation of y. We say that (x, y) cannot have "worse" identifiability than $(\dot{x}^*, y^*)$.

Furthermore, in this case, (x, y) cannot have "better" identifiability than $(\dot{x}^*, y^*)$. Because the two systems are equivalent, if we can obtain $s(\theta)$ from perfect observation of y, and $q = t \circ s$ for some $t : \mathbb{R}^l \to \mathbb{R}^m$, then we can observe $t^{-1}(\theta^*)$ from perfect observation of y^*. But if $l > m$, then we have successfully parametrized the solution manifold with maps of two different dimensions (see note subsequent to this lemma). This is clearly a contradiction; so up to differentiable, almost-everywhere locally invertible functions from $\mathbb{R}^m$ to $\mathbb{R}^m$, q is unique (i.e. there is no better identifiable parametrization, up to the restriction above).

Now suppose that $(\dot{x}^*, y^*)$ is unidentifiable. Then $\exists\ \theta_1^*, \theta_2^* \in \mathbb{R}^m$ such that $\theta_1^* \neq \theta_2^*$ and $\Phi^*(\theta_1^*) = \Phi^*(\theta_2^*)$. Take $q^{-1}(\theta_i^*)$, $i \in \mathbb{N}_2$. Because $(\dot{x}^*, y^*)$ is equivalent to (x, y), $\Phi(q^{-1}(\theta_i^*)) = \Phi^*(q^{-1}(\theta_i^*))$ for $i \in \{1, 2\}$. So $(\dot{x}, y)$ is unidentifiable (at least as bad as $(\dot{x}^*, y^*)$).

To show that $(\dot{x}, y)$ cannot be worse (have higher dimensionality with respect to parameters) than $(\dot{x}^*, y^*)$, we add the restriction that the model must have some identifiable rescaling. Then, by the proof for an identifiable rescaling, any unidentifiable "un"-rescaling of the identifiable model must have equivalent identifiability to the rescaled version, so by considering $(\dot{x}^*, y^*)$ and $(\dot{x}^*, y^*)$ to be "un"-rescalings of a common, equivalent identifiable rescaling, the two systems must be equivalent; in particular, $(\dot{x}, y)$ must have the same dimensionality with respect to parameters as $(\dot{x}^*, y^*)$. □

Note We can consider an identifiable parameterization as a patch map almost everywhere for the manifold consisting of the solutions of the system of differential equations y, where the map is defined by sending a set of parameters to solutions of the system of differential equations satisfying $(\dot{x}, y)$.

8 Appendix 2

8.1 Utilities

There are many different forms of the C coefficients needed, depending on the exact summations and subscripts. Here are some of the forms I found helpful.

$$\begin{aligned} C_{n-j,m} &= \sum_{\substack{S\subset\{n-m+1,\cdots,n\}\\|S|=m-j}} \left(\prod_{l\in S}\gamma_l\right) \\ C_{n-j,n+1} &= \sum_{\substack{S\subset\{0,\cdots,n\}\\|S|=n+1-j}} \left(\prod_{l\in S}\gamma_l\right) \\ C_{n-l,n-j} &= \sum_{\substack{S\subset\{j+1,\cdots,n\}\\|S|=n-j-l}} \left(\prod_{l\in S}\gamma_l\right) \end{aligned} \tag{39}$$

8.2 Computations for $n = 4$

I found it helpful when writing lemmas and obtaining IPOPs to do example computations ("sanity checks"). I found that $n = 4$ is the optimal n for manual computation: large enough to allow you to see the binomial patterns appearing, but small enough that the computations are not prohibitively hard. Thus, I include some of those computations here for the benefit of the reader.

For IPOP 1, we care about Lemma 1 and writing I_j in terms of I_n.

$$\begin{aligned}
\tilde{I}_3 = \tilde{I}_{4-1} &= \prod_{i=4-1}^{4-1} \frac{1}{\gamma_i} \left(\sum_{j=0}^{1} \tilde{I}_4^{(j)} C_{4-j,1} \right) \\
&= \frac{1}{\gamma_3} \left(\gamma_4 \tilde{I}_4 + \tilde{I}_4^{(1)} \right) \\
\tilde{I}_2 = \tilde{I}_{4-2} &= \prod_{i=4-2}^{4-1} \frac{1}{\gamma_i} \left(\sum_{j=0}^{2} \tilde{I}_4^{(j)} C_{4-j,2} \right) \\
&= \frac{1}{\gamma_2\gamma_3} \left(\sum_{j=0}^{2} \tilde{I}_4^{(j)} \left(\sum_{\substack{S \subset \{\gamma_3,\gamma_4\} \\ |S|=2-j}} \left(\prod_{l \in S} \gamma_l \right) \right) \right) \\
&= \frac{1}{\gamma_2\gamma_3} \left(\tilde{I}_4 \left(\gamma_3\gamma_4 \right) + \tilde{I}_4^{(1)} \left(\gamma_3 + \gamma_4 \right) + \tilde{I}_4^{(2)} \right) \\
\tilde{I}_1 = \tilde{I}_{4-3} &= \prod_{i=4-3}^{4-1} \frac{1}{\gamma_i} \left(\sum_{j=0}^{3} \tilde{I}_4^{(j)} C_{4-j,3} \right) \\
&= \frac{1}{\gamma_1\gamma_2\gamma_3} \left(\sum_{j=0}^{3} \tilde{I}_4^{(j)} \left(\sum_{\substack{S \subset \{\gamma_2\gamma_3,\gamma_4\} \\ |S|=3-j}} \left(\prod_{l \in S} \gamma_l \right) \right) \right) \\
&= \frac{1}{\gamma_1\gamma_2\gamma_3} \left(\tilde{I}_4 \left(\gamma_2\gamma_3\gamma_4 \right) + \tilde{I}_4^{(1)} \left(\gamma_2\gamma_3 + \gamma_2\gamma_4 + \gamma_3\gamma_4 \right) + \tilde{I}_4^{(2)} \left(\gamma_2 + \gamma_3 + \gamma_4 \right) + \tilde{I}_4^{(3)} \right)
\end{aligned} \tag{40}$$

This checks Lemma 1 (you get the same pattern as the "continued substitution" list of computations in Section 2). For IPOP 2, we need an expression for $\heartsuit$.

$$\begin{aligned}
\heartsuit &= \sum_{i=1}^{4} \tilde{\beta}_i \tilde{I}_i \\
&= \tilde{\beta}_4 \tilde{I}_4 + \frac{\tilde{\beta}_3}{\gamma_3} \left(\gamma_4 \tilde{I}_4 + \tilde{I}_4^{(1)} \right) + \frac{\tilde{\beta}_2}{\gamma_2\gamma_3} \left(\tilde{I}_4 \left(\gamma_3\gamma_4 \right) + \tilde{I}_4^{(1)} \left(\gamma_3 + \gamma_4 \right) + \tilde{I}_4^{(2)} \right) \\
&\quad + \frac{\tilde{\beta}_1}{\gamma_1\gamma_2\gamma_3} \left(\tilde{I}_4 \left(\gamma_2\gamma_3\gamma_4 \right) + \tilde{I}_4^{(1)} \left(\gamma_2\gamma_3 + \gamma_2\gamma_4 + \gamma_3\gamma_4 \right) + \tilde{I}_4^{(2)} \left(\gamma_2 + \gamma_3 + \gamma_4 \right) + \tilde{I}_4^{(3)} \right) \\
&= \tilde{I}_4 \left(\tilde{\beta}_4 + \frac{\tilde{\beta}_3}{\gamma_3}(\gamma_4) + \frac{\tilde{\beta}_2}{\gamma_2\gamma_3}(\gamma_3\gamma_4) + \frac{\tilde{\beta}_1}{\gamma_1\gamma_2\gamma_3}(\gamma_2\gamma_3\gamma_4) \right) \\
&\quad + \tilde{I}_4^{(1)} \left(\frac{\tilde{\beta}_3}{\gamma_3} + \frac{\tilde{\beta}_2}{\gamma_2\gamma_3}(\gamma_3 + \gamma_4) + \frac{\tilde{\beta}_1}{\gamma_1\gamma_2\gamma_3}(\gamma_2\gamma_3 + \gamma_2\gamma_4 + \gamma_3\gamma_4) \right) \\
&\quad + \tilde{I}_4^{(2)} \left(\frac{\tilde{\beta}_2}{\gamma_2\gamma_3} + \frac{\tilde{\beta}_1}{\gamma_1\gamma_2\gamma_3}(\gamma_2 + \gamma_3 + \gamma_4) \right) \\
&\quad + \tilde{I}_4^{(3)} \left(\frac{\tilde{\beta}_1}{\gamma_1\gamma_2\gamma_3} \right)
\end{aligned} \tag{41}$$