1 **Recognizing Structural Non-Identifiability: When Experiments Don't Provide Information about**

2 **Important Parameters and Misleading Models Can Still Have Great Fit**

3 *Philip J. Schmidt[1,*], Monica B. Emelko[1], Mary E. Thompson[2]*

4 *[1] Department of Civil & Environmental Engineering, University of Waterloo, 200 University Ave. W,*
5 *Waterloo, Ontario N2L 3G1, Canada*

6 *[2] Department of Statistics & Actuarial Science, University of Waterloo, 200 University Ave. W, Waterloo,*
7 *Ontario N2L 3G1, Canada*

8 *[*] Address correspondence to Philip Schmidt, University of Waterloo, 200 University Ave. W, Waterloo,*
9 *Ontario N2L 3G1, Canada; tel. +1(519)8884567 x36745; pj2schmidt@uwaterloo.ca*

10   **ABSTRACT**

11   In the quest to model various phenomena, the foundational importance of parameter identifiability to

12   sound statistical modelling may be less well appreciated than goodness of fit. Identifiability concerns the

13   quality of objective information in data to facilitate estimation of a parameter, while non-identifiability

14   means there are parameters in a model about which the data provide little or no information. In purely

15   empirical models where parsimonious good fit is the chief concern, non-identifiability (or parameter

16   redundancy) implies over-parameterization of the model. In contrast, non-identifiability implies under-

17   informativeness of available data in mechanistically derived models where parameters are interpreted as

18   having strong practical meaning. This study explores illustrative examples of structural non-identifiability

19   and its implications using mechanistically derived models (for repeated presence/absence analyses and

20   dose-response of *E. coli* O157:H7 and norovirus) drawn from quantitative microbial risk assessment.

21   Following algebraic proof of non-identifiability in these examples, profile likelihood analysis and Bayesian

22   Markov Chain Monte Carlo with uniform priors are illustrated as tools to help detect model parameters

23   that are not strongly identifiable. It is shown that identifiability should be considered during experimental

24   design and ethics approval to ensure generated data can yield strong objective information about all

25   mechanistic parameters of interest. When Bayesian methods are applied to a non-identifiable model, the

26   subjective prior effectively fabricates information about any parameters about which the data carry no

27   objective information. Finally, structural non-identifiability can lead to spurious models that fit data well

28   but can yield severely flawed inferences and predictions when they are interpreted or used

29   inappropriately.

30   Key Words: parameter redundancy; quantitative microbial risk assessment (QMRA); dose-response;

31   research ethics; Bayesian analysis

32   **SUMMARY**

33   When structurally non-identifiable, practically relevant model parameters are inherently inestimable.

34   Overlooking this can favour spurious inferences informed too strongly by questionable assumptions.

1. INTRODUCTION

There is a well-known aphorism attributed to George Box that "all models are wrong but some are useful". Quantitative microbial risk assessment (QMRA) (Haas, Rose, & Gerba, 2014) requires a degree of faith in the models upon which it depends because it concerns pathogens that are difficult to quantify accurately, infection processes that are difficult to explore in detail, and epidemiological consequences that are difficult to measure and attribute to specific exposure pathways. Moreover, QMRA can be riddled with many types of variability (e.g. temporal, spatial, person-to-person) and uncertainty (in parameter estimation and model form), concerns about representativeness, and assumptions that are difficult to validate. Nonetheless, there is continued growth in its use as a tool to motivate exploration and understanding of risks and to facilitate decision-making.

QMRA integrally depends upon statistical modelling. A statistical model is a mathematical representation of a set of assumptions linking random variables (e.g. data not yet observed) and other variables that are non-random (e.g. model parameters, known variables, observed data). Identifiability and non-identifiability (also called parameter redundancy) concern the capacity to obtain a unique estimate of a model parameter from data of a particular type. Formally, structural non-identifiability occurs when different sets of values of model parameters yield identical distributions of data (Silvey, 1975; Prakasa Rao, 1992), resulting in models that do not have unique maximum likelihood estimates. Prakasa Rao (1992) notes that "estimation of a parameter is not meaningful unless it is identifiable" while Kreutz, Raue, Kaschek, and Timmer (2013) note that non-identifiability means "the data provides no information about the respective parameter component".

While statistical models are widely used in scientific research, discussion of parameter identifiability is largely limited to formally-trained statisticians. Indeed, introductory probability and statistics text books (especially for engineers and scientists) generally do not feature identifiability, non-identifiability,

58 parameter redundancy, or Fisher Information while some include model fitting material. Analysis of

59 parameter identifiability has been extensively discussed in biostatistics (Catchpole & Morgan, 1997; Raue

60 et al., 2009; Cole, Morgan, & Titterington, 2010; Little, Heidenreich, & Li, 2010; Kreutz et al., 2013,

61 Maiwald et al., 2016). In contrast, it has rarely been addressed in relation to QMRA, with a few exceptions

62 in which it has been highlighted but not extensively discussed (Schmidt, 2015; Brouwer, Weir, Eisenberg,

63 Meza, & Eisenberg, 2017). The biostatistics literature on structural non-identifiability largely focuses upon

64 parameters that are an unnecessary (redundant) addition to a model because they provide categorically

65 no improvement in model fit. In contrast, this study focuses on the other side of the same proverbial coin:

66 that non-identifiability of parameters can indicate under-informativeness of the type of data collected

67 rather than over-parameterization of the model. When a model is mechanistically derived with

68 parameters that have practically meaningful interpretations, arbitrary omission of any redundant

69 parameter has no effect on model fit but can have grave consequences upon statistical inferences.

70 Fig. 1 graphically illustrates the consequences of ignoring structural non-identifiability in model

71 development. Here, parts of the model are *exclusively* supported by subjective beliefs (e.g. potentially

72 inaccurate assumptions, informative Bayesian priors) rather than being principally founded in objective

73 information (e.g. measured data, carefully controlled experimental conditions, rigorously justified

74 assumptions). Data of the same type are fundamentally incapable of disproving flawed subjective beliefs

75 about structurally non-identifiable parameters, and the resulting over-dependence on these beliefs can

76 lead to incorrect scientific inferences. If a model form is appropriate, then future collection of more or

77 better information would continue to support it, refine it, or build upon it with diminishing parameter

78 uncertainty. If, on the other hand, the model form contains flawed subjective beliefs that cannot be

79 challenged by objective information in the data, then future collection of better information could

80 disprove part or all of the model leading to dramatic shifts in statistical inferences. The term 'spurious

81   model' is used herein to describe scenarios in which weak parameter identifiability affords a model

82   sufficient flexibility to contort itself to fit the data despite flawed subjective beliefs.

83   This work provides a brief introduction to the topic of structural non-identifiability and the potential perils

84   of ignoring it. This analysis is presented with examples from and application to QMRA but should be

85   broadly relevant to statistical modelling in general. Comprehensive review of literature on identifiability

86   and associated mathematical principles, or the important topic of practical non-identifiability (e.g. Raue

87   et al., 2009), is beyond the scope of this work. Section 2 provides some practical examples of structural

88   non-identifiability along with relatively simple algebraic proofs. Some applied approaches to evaluate

89   parameter identifiability are addressed in Section 3. Section 4 demonstrates and discusses implications of

90   structural non-identifiability in experimental design and the development and use of models. Finally,

91   Section 5 synthesizes key concepts from this work into recommendations for developing objectively

92   supported models.

93   **2. EXAMPLES OF STRUCTURAL NON-IDENTIFIABILITY IN QUANTITATIVE MICROBIAL RISK ASSESSSMENT**

94   Three examples of structural non-identifiability are drawn from exposure assessment and dose-response

95   models that may be used in QMRA to aid protection of public health from waterborne or foodborne

96   pathogens. They are further considered in Sections 3 and 4 to illustrate diagnosis and implications of

97   structural non-identifiability. For each scenario, 1) the statistical model mechanistically linking observed

98   data to parameters of interest is described, 2) the likelihood function that encapsulates all objective

99   information in the data is provided, and 3) structural non-identifiability is algebraically proved.

100   In each example considered herein, structural non-identifiability of a set of parameters (represented by

101   vector $\boldsymbol{\theta}$) from a particular statistical model can be proved by finding a lower-dimensional function $\boldsymbol{\psi}(\boldsymbol{\theta})$

102   such that the likelihood may be written as a function of $\boldsymbol{\psi}$ without the parameters which $\boldsymbol{\theta}$ comprises. For

103   example, one parameter (ψ) contains all the available information about two parameters (ρ, λ) in the

104 structurally non-identifiable model considered in Section 2.1. Such structural non-identifiability results in

105 a ridge in the likelihood function (Cole et al., 2010) along which there are many model fits that are equally

106 and optimally supported by the available data. This is analogous to a system of equations from which the

107 solution is indeterminate because (1) there are fewer equations than variables or (2) some of the

108 equations do not carry independent information. A brief introduction to such structural non-identifiability

109 is provided in the Supplementary content. Other comparatively complex mathematical approaches that

110 involve determining the rank of the Hessian matrix or Fisher information matrix (Cole et al., 2010; Little

111 et al., 2010) are not considered herein.

112 **2.1. Fitting a Concentration Distribution to a Set of Non-Repeated Presence/Absence Analyses**

113 Among the traditional culture-based methods for quantifying microorganisms, there is a long history of

114 methods involving presence/absence analyses (McCrady, 1915; Cochran, 1950; Pouillot, Hoelzer, Chen, &

115 Dennis, 2013). Assuming a homogeneous concentration of target microorganisms, random dispersion

116 (e.g. no aggregation) of these microorganisms, and independence of aliquots drawn from the source, the

117 number of these microorganisms in each aliquot should be Poisson-distributed (Student, 1907; Emelko,

118 Schmidt, & Reilly, 2010). The mean of this Poisson distribution is the product of concentration (c) and

119 aliquot volume (V), and an individual presence/absence analysis will be positive ($X = 1$) with probability

120 $1 - e^{-cV}$ if one or more culturable target microorganisms are present in the aliquot, and negative ($X = $

121 0) with probability e^{-cV} otherwise. This assumes perfect analytical recovery (i.e. no losses of culturable

122 target microorganisms (Petterson, Dumoutier, Loret, & Ashbolt, 2009)) and specificity (i.e. no false-

123 positive detections) of the sample processing procedure. The resulting likelihood function for

124 concentration given a measured volume and presence/absence result (represented by presence indicator

125 variable X) may be expressed as Equation 1.

126 $$L(c; V, X) = (1 - e^{-cV})^X (e^{-cV})^{1-X}, c > 0 \tag{1}$$

127   A single presence/absence analysis is commonly regarded as yielding qualitative data because it does not

128   allow estimation of the microbial concentration. A positive result may be interpreted as censored count

129   data because one or more discrete microorganisms is needed for detection to occur while a negative

130   result is essentially a count of zero (Chik, Schmidt, & Emelko, 2018). A negative result cannot prove

131   absence in the source because low concentration, small aliquot volume, or imperfect analytical recovery

132   can cause non-detects when target microorganisms are present. The likelihood function associated with

133   a positive result is monotonic increasing towards a horizontal asymptote at 1 (because non-detects

134   become practically impossible at high concentrations). It is explained in Section 3.1 that such a positive

135   result (or group of repeated positive results) is an example of practical non-identifiability. A negative result

136   is associated with a monotonic decreasing likelihood function (because a positive result becomes

137   practically impossible at very low concentrations) with maximum likelihood at a concentration of zero. A

138   suite of repeated presence/absence analyses from the same source, however, allows quantitative

139   estimation (if there is at least one negative result) by the most probable number (MPN) method (Cochran,

140   1950). It determines the value of concentration that maximizes the likelihood function shown in Equation

141   2, in which subscript i on the volume and presence/absence indicator denotes the ith of n aliquots.

$$L(c; \{V_i\}, \{X_i\}) = \prod_{i=1}^{n}(1 - e^{-cV_i})^{X_i}(e^{-cV_i})^{1-X_i}, c > 0 \tag{2}$$

143   If all model assumptions are valid, the maximum likelihood estimator of concentration will converge upon

144   the true underlying concentration as more aliquots are analyzed due to the property of consistency

145   (Silvey, 1975). Cochran (1950) noted that the precision of the most probable number method with a

146   constant aliquot volume becomes poorer for aliquot volumes at which either positive or negative results

147   become rare. The optimal aliquot volume—$V_{opt} \cong 1.5936/c$, corresponding to a 79.68% probability of

148   detection—can be evaluated using the Fisher information for the model in Equation 2 with constant

149   aliquot volume as shown in the Supplementary Content. It is common knowledge that some experimental

150   designs are inherently more informative about a model's parameters than others (further addressed in

151   Section 4.1), and this can be explored in a mathematically rigorous way using Fisher information.

152   The exposure assessment module of QMRA often requires description of how concentrations vary over

153   time (in water) or among portions (in food). Equation 2 can be modified to include a distribution for

154   variation in concentration among samples as well as repeated presence/absence analyses for each sample

155   (e.g. Pouillot et al., 2013; Schmidt, Pintar, Fazil, Flemming, et al., 2013). Suppose, hypothetically, that some

156   researchers want to determine the average pathogen concentration for use in QMRA and assume that

157   concentration varies according to a gamma distribution with shape parameter ρ and scale parameter λ.

158   To reduce cost and the amount of sample processing, they are considering carrying out only one

159   presence/absence analysis per sample (each with the same volume) rather than carrying out a full MPN

160   analysis for each sample. Given $c \sim gamma(\rho, \lambda)$, the marginal probability of detection[1] is $1 -$

161    $(V\lambda + 1)^{-\rho}$. The likelihood function for the proposed experimental design and statistical model is shown

162   in Equation 3 with subscript i on the presence/absence indicator denoting the ith of n samples.

163    $$L(\rho, \lambda; V, \{X_i\}) = \prod_{i=1}^{n} [1 - (V\lambda + 1)^{-\rho}]^{X_i} [(V\lambda + 1)^{-\rho}]^{1-X_i}, \rho > 0, \lambda > 0 \qquad (3)$$

164   This model is just a series of independent Bernoulli trials with equal probability of success (detection)

165    $\psi(\rho, \lambda) = 1 - (V\lambda + 1)^{-\rho}$. Accordingly, the number of detections obtained in n samples, $S = \sum_{i=1}^{n} X_i$, is

166   binomially distributed $(S \sim binomial(n, \psi))$. S is a sufficient statistic for such a model (Silvey, 1975),

167   meaning that it carries all the information available in the data $\{X_i\}$ for estimation of ψ. Critically, such an

168   experimental design essentially yields only one datum (S) from which estimation of two model parameters

169   (ρ, λ) is impossible—this is due to structural non-identifiability. Specifically, it is possible to estimate ψ,

170   but collecting further data of this type can never enable estimation of both ρ and λ (or mean $\mu = \rho\lambda$ and

---

[1] If $Y \sim Poisson(cV)$ and $c \sim gamma(\rho, \lambda)$, then $p_Y(y) = \int_0^\infty \left[\frac{e^{-cV}(cV)^y}{y!}\right] \left[\frac{1}{\Gamma(\rho)\lambda^\rho} c^{\rho-1} e^{-c/\lambda}\right] dc = \frac{\Gamma(y+\rho)}{y!\Gamma(\rho)} \left(\frac{V\lambda}{V\lambda+1}\right)^y \left(\frac{1}{V\lambda+1}\right)^\rho$ and $p_Y(y > 0) = 1 - p_Y(0) = 1 - (V\lambda + 1)^{-\rho}$.

171   standard deviation $\sigma = \rho^{0.5}\lambda$) at the same time. Any of a spectrum of gamma distributions conforming to

172   the maximum likelihood estimate $\hat{\psi}_{MLE} = 1 - (\lambda V + 1)^{-\rho} = S/n$ is equally and optimally supported by

173   the available data; thus, maximum likelihood can be obtained for any arbitrary value of ρ by finding the

174   corresponding value of λ (or vice versa) for which $\hat{\psi}_{MLE} = 1 - (\lambda V + 1)^{-\rho}$. Analysis of parameter

175   identifiability during the experimental design stage would reveal that such a sampling plan is inherently

176   incapable of providing the needed information. Accordingly, an unsuitable study design can be averted

177   before time and resources are wasted collecting inadequately informative data.

178   It becomes possible to fit such a two-parameter model for variability in concentration among samples to

179   this type of data if there are presence/absence analyses using two or more sample volumes. This happens

180   because the structural non-identifiability represented by $\psi(\rho, \lambda) = 1 - (V\lambda + 1)^{-\rho}$ requires a constant

181   sample volume. When Equation 3 is modified slightly by varying the sample volumes, the full set of

182   experimental   data $\{X_i, V_i | i = 1,2, \dots n\}$ can   only   be   reduced   to   a   set   of   sufficient   statistics

183   $\{S_j | j = 1,2, \dots m\}$ corresponding to the m unique sample volume values. The model's parameters are

184   identifiable, though perhaps not strongly so; estimates of ρ, λ may be much less precise than those

185   obtained using a better experimental design, as illustrated in Section 4.1.

186   **2.2. Fitting an Exact Beta-Poisson Dose-Response Model to Data from a Single Dose Group**

187   The dose-response relationship for *E. coli* O157:H7 has been explored using data from a foodborne

188   outbreak in a school in Japan (Teunis, Takumi, & Shinagawa, 2004). Stored samples of the contaminated

189   food and the known amount of food allotted to each exposed pupil or teacher enabled quantification of

190   exposure. Pupils and teachers are believed to have been exposed to mean doses of 31 and 35 colony

191   forming units, respectively. In all, 208 of 828 pupils and 7 of 43 teachers became infected. Separate exact

192   beta-Poisson models were fit to data from each group based upon a hypothesis that dose-response may

193   differ between pupils and teachers.

194   This model assumes that the number of pathogens consumed is Poisson-distributed with mean N (mean

195   dose), each pathogen consumed by a particular host has a probability r of successfully replicating to

196   initiate infection, and r varies among subjects (Schmidt, Pintar, Fazil, & Topp, 2013; Nilsen & Wyller, 2016)

197   according to the beta distribution $r \sim beta(\alpha, \beta)$ with shape parameters α and β. The resulting

198   probability of infection is $P = 1 - {_1F_1}(\alpha, \alpha + \beta; -N)$, where ${_1F_1}()$ is Kummer's confluent

199   hypergeometric function. The likelihood function for a set of subjects (e.g. pupils) exposed to equal mean

200   doses is shown in Equation 4 with subscript i on the infection status indicator (X) denoting the ith of n

201   exposed subjects.

202
$$L(\alpha, \beta; N, \{X_i\}) = \prod_{i=1}^{n} \left[1 - {_1F_1}(\alpha, \alpha + \beta; -N) \right]^{X_i} \left[{_1F_1}(\alpha, \alpha + \beta; -N) \right]^{1-X_i}, \alpha > 0, \beta > 0 \quad (4)$$

203   When all subjects are exposed to the same mean dose, this model is just a series of independent Bernoulli

204   trials with equal probability of success (infection) $\psi(\alpha, \beta) = 1 - {_1F_1}(\alpha, \alpha + \beta; -N)$. Accordingly, the

205   number of infections obtained in n subjects, $S = \sum_{i=1}^{n} X_i$, is binomially distributed $\left(S \sim binomial(n, \psi) \right)$

206   and is a sufficient statistic for this model. Considering dose-response of pupils separately from teachers,

207   this outbreak yields only one datum (S) from which estimation of two model parameters (α,β) is

208   impossible due to structural non-identifiability. Specifically, it is possible to estimate ψ, but collecting

209   further data of this type can never enable estimation of both α and β at the same time. For the pupils,

210   any of a spectrum of exact beta-Poisson models conforming to the maximum likelihood estimate $\hat{\psi}_{MLE} =$

211    $1 - {_1F_1}(\alpha, \alpha + \beta; -31) = 208/828$ has equal and optimal objective support from the available data.

212   The resulting implications upon Bayesian parameter uncertainty analysis are addressed in Section 4.2.

213   **2.3. Fitting an Exponential Dose-Response Model with Uncontrolled Aggregation**

214   Structural non-identifiability was relatively easy to prove in the preceding examples because each

215   experiment was just a series of independent Bernoulli trials with equal probability of success, which is

216   insufficient information to fit a two-parameter model. Using variable volumes in the first example or

217   testing several mean dose values in dose-response experiments resolves such non-identifiability by

218   varying the probability of success among the Bernoulli trials. In a reanalysis of published norovirus dose-

219   response models, Schmidt (2015) explored an example of structural non-identifiability that occurs despite

220   testing many values of the administered mean dose.

221   In the aggregated exponential dose-response model, it is assumed that pathogens are aggregated to an

222   unknown degree, the number of aggregates in a dose is Poisson-distributed, and the number of pathogens

223   per aggregate is log-series distributed with parameter a. Additionally, it is assumed that each pathogen

224   consumed by a particular host has probability r of successfully replicating to initiate infection and that r

225   is constant among hosts. The probability of infection (P) is described by Equation 5. The resulting

226   likelihood function is shown in Equation 6 with subscript i on the infection status indicator (X) and

227   administered mean dose (N) denoting the ith of n exposed subjects.

228
$$P = 1 - \left(\frac{1-a}{1-a+ar}\right)^{N(1-a)/a} \tag{5}$$

229   $L(a, r; \{N_i\}, \{X_i\}) = \prod_{i=1}^{n} \left[1 - \left(\frac{1-a}{1-a+ar}\right)^{N(1-a)/a}\right]^{X_i} \left[\left(\frac{1-a}{1-a+ar}\right)^{N(1-a)/a}\right]^{1-X_i}, 0 < a < 1, 0 < r \leq 1$ (6)

230   With the substitution $e^{-\psi} = \left(\frac{1-a}{1-a+ar}\right)^{(1-a)/a}$, Equation 5 simplifies to $P = 1 - e^{-\psi N}$. Accordingly,

231   Equation 6 depends only upon ψ and not upon the parameters of interest (a, r), proving structural non-

232   identifiability. Specifically, it is possible to estimate ψ, but collecting further data of this type can never

233   enable estimation of both a and r at the same time. Any of a spectrum of aggregated exponential dose-

234   response models conforming to the maximum likelihood estimate[2] $\hat{\psi}_{MLE} = \frac{1-a}{a} ln\left(1 + \frac{ar}{1-a}\right)$ has equal

235   and optimal support from the available data. This model is used in Section 4.3 along with simulated data

236   to illustrate how structural non-identifiability can lead to spurious models that have good fit to the data

---

[2] This includes the exponential model $P = 1 - e^{-\hat{\varphi}_{MLE}N}$ obtained in the limit as $a \to 0$.

237 but yield incorrect inferences and predictions. Considering norovirus dose-response in particular, Schmidt

238 (2015) described how the aggregation parameters in the aggregated exact beta-Poisson model (Teunis et

239 al., 2008) and aggregated fractional Poisson model (Messner, Berger, & Nappier, 2014) behave as tuning

240 parameters and may lead to spurious models.

241 Uncontrolled aggregation is a confounding factor in this experimental design that precludes estimation of

242 parameter r. For such reasons, ensuring disaggregation has always been foundational in many aspects of

243 quantitative microbiology (Eisenhart & Wilson, 1943). To avert wastefully administering pathogens to

244 humans in order to collect inadequately informative data, dose-response experiments such as this should

245 not be conducted if the pathogens are aggregated to an unknown degree. This invokes a need for review

246 of experimental design by a qualified statistician as a condition for ethics approval of such experiments.

247 **3. APPROACHES TO SEARCH FOR STRUCTURAL NON-IDENTIFIABILITY IN A STATISTICAL MODEL**

248 Recognizing and algebraically proving structural non-identifiability in a statistical model given a particular

249 set of data can be much more complicated than the simple examples shown in Section 2. This section

250 provides two exploratory techniques to seek evidence of structural non-identifiability: assessment of

251 profile likelihood and Bayesian analysis with uniform priors. These methods can also reveal practical non-

252 identifiability, which occurs when a model's parameters are strictly identifiable from data of a particular

253 type but the available data happen to lack strong information about a parameter. These approaches may

254 be applied in two ways: 1) using simulated data during the experimental design stage or 2) exploring

255 parameter identifiability while fitting a model to experimental data. The premise for using simulated data

256 is that a large sample of data consistent with the planned experimental design should facilitate precise

257 estimation of all parameters of the statistical model used to generate the data unless some parameters

258 are structurally non-identifiable.

259   Simulated data from an aggregated exponential with immunity dose-response model (Schmidt, 2015) are

260   used to illustrate these approaches. This model (Equation 7) adds an immunity parameter (ϕ) to the

261   aggregated exponential model (Equation 5) to represent a portion of the population of subjects who

262   cannot be infected regardless of the dose to which they are exposed. Data (Table I) were simulated using

263   arbitrary parameter values $\phi = 0.25$ and $r = 0.1$ as well as disaggregation of the administered pathogens

264   ($a \to 0$) as is typically assumed in most dose-response experiments. Parameters a,r are structurally non-

265   identifiable (as shown in Section 2.3) unless one or the other has a known or measured value, while ϕ is

266   identifiable. To provide clear evidence of this structural non-identifiability, a large number of subjects was

267   simulated for each of several mean dose values spanning the region where the probability of infection

268   rises rapidly from zero to the maximum. In practice, however, the most informative administered doses

269   are generally not known a priori and it is impractical to administer pathogens to large numbers of human

270   volunteers. Dose-response experiments need to include some high mean doses yielding a maximal

271   probability of infection to facilitate precise estimation of ϕ (Schmidt, 2015).

272
$$P = (1 - \phi)\left[1 - \left(\frac{1-a}{1-a+ar}\right)^{N(1-a)/a}\right] \tag{7}$$

273   **3.1. Profile Likelihood Analysis**

274   Structural non-identifiability is often manifested as a ridge in the likelihood surface along which

275   alternative sets of parameter values have equal and optimal support from the data. Profile likelihood

276   analysis (Raue et al., 2009; Kreutz et al., 2013) facilitates detection of such features in models with many

277   parameters and an easily evaluated and optimized likelihood function. The profile likelihood for

278   parameter θ_1 is obtained by maximizing the likelihood function at each of a suite of specified values of θ_1.

279   For a two-parameter model, this approach resembles looking at the topography of an island from far

280   offshore with views from the south and west corresponding to the profiles for the two parameters. A

281   profile likelihood with a unique maximum is indicative of an identifiable parameter, while maximum

282   likelihood along a plateau may be indicative of structural non-identifiability.

283   Raue et al. (2009) also discuss 'practical non-identifiability', where weak information about a parameter

284   that is technically identifiable leads to a profile likelihood that has a unique maximum but diminishes to a

285   high plateau in one direction or the other. For example, the shape parameters of a beta distribution can

286   be practically non-identifiable if the distribution's variance could plausibly be zero: the profile likelihood

287   of each shape parameter would diminish to a high plateau for large parameter values (corresponding to

288   vanishing variance). A key distinction is that practical non-identifiability can be progressively resolved by

289   collecting more data. In contrast, structural non-identifiability can never be resolved by collecting more

290   data unless the experimental design is changed to control or facilitate estimation of the otherwise non-

291   identifiable parameters. Repeated presence/absence analyses that are all positive (present) may be an

292   unusual instance of practical non-identifiability. In this case, the concentration would otherwise be

293   identifiable but the likelihood rises asymptotically towards unity at high concentrations. A unique

294   maximum does not emerge until one negative (non-detect) result is obtained.

295   Fig. 2 shows each profile likelihood of the aggregated exponential with immunity dose-response model fit

296   to the Table I data. The *mle* function in the *stats4* package of R (R Core Team, 2017) was used for

297   optimization (see code in Supplementary Content). The immunity parameter (ϕ) has an identifiable

298   maximum likelihood estimate (Fig. 2a). Due to the structural non-identifiability shown in Section 2.3, the

299   degree of aggregation (transformed to the more practically meaningful mean aggregate size $\mu =$

300    $-a/[(1-a)ln(1-a)]$) and host susceptibility among the non-immune (r) each have a flat profile

301   likelihood (Figs. 2b and 2c). Though structurally non-identifiable, these parameters are called 'set

302   identifiable' because the maximum likelihood estimate of ψ corresponds to ranges of possible values of

303   the non-identifiable parameters that do not fully span the parameter space. In this example, the maximum

304   likelihood estimate of ψ corresponds to values of the mean aggregate size between 1 and 9.3 and values

305   of the host susceptibility among the non-immune between 0.1074 and 1. Parameter values outside of

306   these ranges are still plausible, but can only correspond to values of ψ with smaller likelihood than its

307   maximum likelihood estimate. The numerous data are strongly informative about ψ (Fig. 2d), so the

308   profile likelihood curves fall off sharply outside of the ranges corresponding to this set identifiability.

309   **3.2. Bayesian Markov Chain Monte Carlo Analysis with Uniform Priors**

310   Profile likelihood analysis may become i

311   mpractical if the likelihood function cannot be explicitly evaluated or easily optimized, as often occurs in

312   hierarchical statistical models. If independent uniform priors are used in Bayesian parameter uncertainty

313   analysis, the posterior and likelihood will share the same shape for a particular parameterization and

314   scatter plots of pairs of untransformed parameter values drawn from the posterior by Markov Chain

315   Monte Carlo (MCMC) may reveal evidence of a ridge in the likelihood surface. Proper uniform priors (i.e.

316   with fixed bounds) may be necessary to ensure a proper posterior, and MCMC convergence and mixing

317   must be adequate to provide a representative sample from the posterior.

318   An aggregated exponential with immunity model was fit to the Table I simulated dose-response data using

319   OpenBUGS (version 3.2.3, rev 1012) to implement MCMC (see code in Supplementary Content). A uniform

320   prior for each parameter ($U(0,1)$), default updater algorithms, and generated initial values for the Markov

321   Chain were used. The posterior distribution is represented with 10,000 iterations (specifically, every

322   hundredth of one million iterations following a burn-in of 1,000 iterations). Convergence and mixing were

323   visually assessed using three chains and history plots in OpenBUGS.

324   The resulting ϕ, r (Fig. 3a) and ϕ, a (Fig. 3b) scatter plots indicate that the immunity parameter (ϕ) is

325   strongly identifiable, with a 95% credible interval (0.1736 to 0.2845) encompassing the true value of 0.25.

326   These plots suggest non-identifiability of the other parameters (r, a) because there is a ridge in the

327   posterior that spans much of the parameter space and the posterior density does not diminish in one

328   direction. The a, r scatter plot (Fig. 3c) provides compelling evidence of non-identifiability because a ridge

329   (following the known structural non-identifiability characterized by $\psi(a,r) = \frac{1-a}{a} ln \left(1 + \frac{ar}{1-a} \right)$) spans

330   much of the parameter space without diminishing posterior density in either direction. Moreover, a mode

331   would be expected near the actual values of the parameters used to generate the simulated data if these

332   parameters were identifiable.

333   **4. IMPLICATIONS OF NON-IDENTIFIABILITY**

334   Structural non-identifiability is not just an obscure mathematical concept—it can have grave

335   consequences when it is ignored in experimental design and the development and use of models. Here,

336   the examples of structural non-identifiability shown in Section 2 are used to explore practical implications

337   for experimental design, model fitting and parametric uncertainty analysis, and model-based inference.

338   These implications may also apply in instances of weak parameter identifiability.

339   **4.1. Uninformative or Weakly Informative Data Should Be Avoided in Experimental Design**

340   It is generally understood that sensible experimental design is necessary to ensure experiments can yield

341   scientifically useful results. If a particular theoretically derived model form is anticipated before

342   conducting the experiment, variance decomposition (Schmidt, Emelko, & Thompson, 2014) can provide

343   insight into strategies to improve the quality of information in experimental data. Alternatively, simulation

344   studies may be a helpful precursor to carrying out an experiment. Here, simulation is used to illustrate

345   how structural non-identifiability is a particular concern in experimental design because corresponding

346   experiments provide no information to differentiate among a suite of model fits that would be equally

347   supported by the data. It is essential for scientists to recognize experimental designs that can be

348   foreknown to lead to structurally non-identifiable models before resources are wasted collecting

349   inadequately informative data. This is particularly important if experiments require ethics approval: an

350   experimental design that can be foreknown to provide inadequate information about important model

351   parameters should not be approved.

352   The models discussed in Section 2.1 are used to provide an illustrative comparison of experimental designs

353   corresponding to identifiable, weakly identifiable, and structurally non-identifiable model parameters.

354   Table II summarizes six considered scenarios. In each, variability of the pathogen concentrations was

355   simulated using a gamma distribution with $\rho = 1.5625$ and $\lambda = 6.4$ (corresponding to mean $\mu =$

356    $10\ \mathrm{MPN/mL}$ and standard deviation $\sigma = 8\ \mathrm{MPN/mL}$). The number of sampling events and the number

357   of presence/absence analyses per sampling event (and their respective volumes) vary among scenarios.

358   The second trio of scenarios adds data to the first trio to illustrate the effects of collecting further data

359   with differing degrees of parameter identifiability. The data from each scenario were analyzed using

360   OpenBUGS (as described in Section 3.2) to implement MCMC (see code in Supplementary Content).

361   Uniform priors (between -1 and 2) were used for the base-10 logarithm of the mean and standard

362   deviation of the gamma distribution, so scatter plots (Fig. 4) show the mean and standard deviation in

363   logarithmic scale.

364   In Fig. 4a, the MCMC sample from the posterior clusters to some extent around the true parameter values

365   because the model is identifiable. Notably, the density of points from the MCMC sample does not tail off

366   at low values of standard deviation. This pattern is consistent with practical non-identifiability and occurs

367   because the small number of data available can still be plausibly explained by a constant concentration

368   (or any trivially small value of the standard deviation). Fig. 4b is indicative of weaker parameter

369   identifiability relative to Fig. 4a because there is a lesser degree of clustering of the MCMC sample from

370   the posterior around the true parameter values. Additionally, even the larger number of data is unable to

371   provide compelling evidence against a trivially small variance of the gamma distribution. Scenarios A2 and

372   B2 have equal numbers of presence/absence analyses, but Scenario B2 involves more work (sampling

373   events) to provide less useful information. It is known (from Section 2.1) that the model corresponding to

374   Fig. 4c is structurally non-identifiable and that collecting further data of the same type improves the

375   estimation of $\psi(\rho, \lambda) = 1 - (V\lambda + 1)^{-\rho}$ without ever allowing estimation of $\mu = \rho\lambda$ and $\sigma = \rho^{0.5}\lambda$. A

376   simulation study such as this can help to choose a preferred experimental design before wasting resources

377   on inadequately informative data.

378   **4.2. Bayesian Analyses of Structurally Non-Identifiable Models are Unduly Influenced by the Prior**

379   Given a particular model form, Bayesian analysis provides a framework to merge objective information

380   from data (the likelihood) with subjective beliefs of the analyst (the prior) to provide a characterization of

381   uncertainty in the model's parameters given both sources of information (the posterior). The prior's effect

382   usually becomes progressively muted as more data strengthen the informativeness of the likelihood.

383   Relatively uninformative priors also reduce the effect of subjective beliefs upon the posterior. This is

384   generally desirable in science so that inferences are founded in defensible objective information rather

385   than being too strongly influenced by subjective beliefs. When a model is structurally non-identifiable

386   given the type of data available, it is imperative for the analyst to recognize and clearly concede that the

387   data carry no objective information about some parameters and that posterior information about these

388   parameters is determined by the prior alone. In this way, greater scientific insight can be facilitated by

389   motivating more informative experimental work in the future or more rigorous exploration of key

390   assumptions. Without due recognition of structural non-identifiability and its implications, it could be

391   claimed that Bayesian methods were used injudiciously to model one's way out of a confounded

392   experiment or otherwise uninformative data. Bayesian analysis of non-identifiable models without a well-

393   justified informative prior can be biased because the prior effectively fabricates information missing from

394   the data and creates an illusion of strong data-centric science. This is illustrated using the example from

395   Section 2.2.

396   Fitting the two-parameter exact beta-Poisson dose-response model to one datum (208 of 828 pupils at a

397   mean dose of 31 *E. coli* O157:H7) leads to structural non-identifiability characterized by $\hat{\psi}_{MLE} = 1 -$

398    $_1F_1(\alpha, \alpha + \beta; -31) = 208/828$ (Fig. 5a). Thus, a spectrum of exact beta-Poisson models passing through

399   the point $(N, \psi) = (31, 208/828)$ are equally supported by the data (Fig. 5b). The set of such models is

400   bounded by two extremes of the variance of $r \sim beta(\alpha, \beta)$: the variance is minimized by the exponential

401   model $P = 1 - exp(-rN)$ where $r = \alpha/(\alpha + \beta)$ and is maximized by the fractional Poisson model $P =$

402    $(1 - \phi)(1 - exp(-N))$ with immunity parameter $\phi = \beta/(\alpha + \beta)$. Notably, the latter model makes the

403   questionable assertion that 75% of pupils would be immune to any dose of *E. coli* O157:H7. There is a 27-

404   fold (1.43-log)[3] difference in low-dose risks between these extremes, so it is important to consider

405   whether or not Bayesian analysis can reliably aid model fitting and inferences in this scenario.

406   Teunis et al. (2004) recognized that "the data still only represent a single point in a dose-response

407   relation", but nonetheless undertook a Bayesian analysis to draw inferences about the exact beta-Poisson

408   model's two parameters without noting the inevitable structural non-identifiability this causes. They used

409   "prior specifications that allow an extremely wide range of parameter values, whereby they may be

410   assumed noninformative": a uniform prior (uniform(0,1)) on $u = \alpha/(\alpha + \beta)$ and a broad normal prior

411   (normal(0,10)[4]) on $v = log_{10}(\alpha + \beta)$. MCMC was then used to characterize uncertainty in the dose-

412   response relationship and to assert that the 90% posterior probability of infection of a pupil exposed to

413   just one *E. coli* O157:H7 is between 0.072 and 0.274. Additionally, a posterior mode at $(\alpha, \beta) =$

---

[3] One extreme is the fractional Poisson model with $\phi = 1 - \psi \cong 0.2512$, while the other extreme is the exponential model with $r = -ln(1 - \psi)/31 \cong 0.009332$. The low-dose ($N \ll 1$) linear approximations for dose-response are $\frac{d}{dN}\Big|_{N \ll 1} (1 - 620/828) \times (1 - exp(-N)) \cong 0.2512$ and $\frac{d}{dN}\Big|_{N \ll 1} 1 - exp(-0.009332 \times N) \cong 0.009332$. $log_{10}(0.2512/0.009332) \cong 1.43$

[4] It is assumed herein that this normal prior has a variance of 10 because the notation "normal(0,10)" is ambiguous.

414     $(0.0844, 1.442)$[5] was presented as the exact beta-Poisson dose-response relation for pathogenic *E. coli* in

415   children.

416   Critically, there is no such thing as a "noninformative" prior in Bayesian analysis; there are only relatively

417   uninformative priors, and even these become informative when applied to a structurally non-identifiable

418   model. In cases of structural non-identifiability, a sufficiently broad prior will provide good coverage of

419   the spectrum of model fits equally and optimally supported by the data. However, any differences in

420   posterior density along the non-identifiability relationship (e.g. $\psi(\alpha, \beta) = 1 - {}_1F_1(\alpha, \alpha + \beta; -31)$), as

421   illustrated in Fig. 6, would be determined exclusively by the prior (i.e. with categorically no support from

422   the data). An MCMC sample from the posterior for pupils (with adequate convergence/mixing) should

423   consist only of points close to the curve $\hat{\psi}_{MLE} = 1 - {}_1F_1(\alpha, \alpha + \beta; -31) = 208/828$ because $\hat{\psi}_{MLE}$ is

424   estimated with good precision from results for 828 pupils, but any preference to a specific part of the

425   curve is determined by the informative prior alone.

426   Given information that 208 of 828 pupils became infected at a mean dose of 31 *E. coli* O157:H7, the data

427   provide categorically no objective information to draw inferences about the probability of infection at

428   other mean doses. A one-parameter model can be fit, but goodness of fit cannot be tested because the

429   data carry zero degrees of freedom[6]. A two-parameter model cannot be fit without assuming the value of

430   one parameter or using a strongly informative prior, and goodness of fit cannot be tested because the

431   data carry no degrees of freedom. Many QMRA researchers (e.g. Messner et al., 2014) have raised

432   concerns about low-dose extrapolation (drawing inferences outside the range of tested mean doses) even

433   for models with tested goodness of fit. In this case, using the model determined from the posterior mode

---

[5] Actually, the posterior mode is $(\alpha, \beta) = (0.071, 0.929)$, as determined by the intersection of maximal ridges in the prior and likelihood along $v = 0$ (or $\alpha + \beta = 1$) and $1 - {}_1F_1(\alpha, \alpha + \beta; -31) = 208/828$, respectively.

[6] In dose-response experiments, the degrees of freedom are the number of dose values tested minus the number of free parameters.

434   is even more dubious because it not only extrapolates beyond the mean dose encountered in the outbreak

435   but also uses a subjectively derived model for which it is impossible to test goodness of fit. Therefore,

436   these *E. coli* dose-response models are not useful for QMRA relative to published models objectively

437   supported by data from which goodness of fit can be tested.

438   **4.3. Non-Identifiability Facilitates Spurious Model Fit**

439   Ideally, abundant data allow both precise estimation of a model's parameters and testing the model's

440   goodness of fit. Theoretically derived models include many assumptions, and informative data can provide

441   evidence of an invalid assumption through poor model fit. The less informative the data are, the easier it

442   is for a model incorporating an inappropriate assumption to have adequate fit. Conversely, adding more

443   parameters to a model than the data can support can improve the fit of a model incorporating a flawed

444   assumption. In the extreme of structural non-identifiability, a value of one of the non-identifiable

445   parameters can be chosen almost at random[7] and the model will obligingly contort itself to fit this

446   assumption with no loss of fit. This effectively deprives the data of their ability to speak through poor fit

447   if an inappropriate value of a non-identifiable (or weakly identifiable) parameter is assumed. This may

448   have little consequence if the model is regarded as only an empirical fit applicable in essentially the same

449   conditions, but is immensely problematic if the model is used to make mechanistic inferences or applied

450   under substantially different conditions. The result is a spurious model that fits the data well while

451   misrepresenting the relationship between specific parameters and the data.

452   The dose-response data in Table I (simulated with $\phi = 0.25$, $a \rightarrow 0$ and $r = 0.1$) can be used to illustrate

453   how structural non-identifiability can result in misleading models that have great fit to the data. Maximum

---

[7] This can happen unintentionally if algorithms for maximum likelihood estimation yield incorrect or misleading results. Relatively flat likelihood functions are prone to computational error in optimization, and many algorithms will not indicate that the likelihood has a non-unique maximum (structural non-identifiability). The Solver function in Microsoft Excel™, for example, yields different maximum likelihood estimates for different starting points (using either the GRG Nonlinear or Evolutionary method) if parameters are structurally non-identifiable.

454   likelihood fits of several special cases of the aggregated exponential with immunity model to the data are

455   compared (Table III), particularly with and without assuming that the pathogens were disaggregated in

456   the simulations. When the degree of aggregation is known, the data are too informative for the model to

457   contort itself to provide good fit to a flawed assumption. Accordingly, the exponential with immunity

458   model (the right one with respect to how the simulated data were generated) fits the data 10^{17} times

459   better than the fractional Poisson model, soundly rejecting the flawed assumption that $r = 1$. When the

460   degree of aggregation is not known, the data cannot enable concurrent estimation of parameters a and

461   r due to structural non-identifiability. Thus, for almost any assumed value of one of these parameters

462   (subject to the constraints of set identifiability discussed in Section 3.1), the model is readily able to

463   contort itself to achieve great fit.

464   The aggregated fractional Poisson model has a spurious fit in this example that could mislead the analyst

465   into believing that these pathogens were aggregated with a mean cluster size of 9.312 when they were

466   actually disaggregated. Additionally, the analyst is left with no evidence against the false assumption that

467    $r = 1$ because the aggregation parameter behaved wholly as a tuning parameter to compensate for the

468   incorrect assumption. Empirically, this model shares the form $P(N) = 0.7782 \times \left(1 - exp(-0.1074N)\right)$

469   with any instance of the structurally non-identifiable aggregated exponential with immunity model

470   featuring $\hat{\psi}_{MLE} = \frac{1-a}{a} ln \left(1 + \frac{ar}{1-a}\right) = 0.1074$. There is no harm in applying it as a partially mechanistic

471   model under essentially the same conditions, which effectively treats parameter ψ as the combined effect

472   of unknown parameters a and r. Substantial bias can arise, however, if $\hat{\psi}_{MLE} = 0.1074$ is misinterpreted

473   (e.g. as a mean aggregate size of 9.312 in this example) or misapplied (e.g. to a scenario in which the

474   parameters it comprises have changed such that the numerical value of parameter ψ has changed but the

475   modeller cannot know it). Dose-response models in QMRA are often described as semi-mechanistic, yet

476   it is common to remove the aggregation parameter from norovirus dose-response models to represent

477   disaggregated viruses in environmental waters (United States Environmental Protection Agency, 2014;

478   World Health Organization, 2016). This is contrary to empirical or semi-mechanistic treatment of the fitted

479   models and requires all mechanistic assumptions in the model to be valid.

480   For example, the aggregated fractional Poisson model $P(N) = 0.7782 \times \left(1 - exp(-N/9.312)\right)$ has a

481   good empirical fit to the Table I data (a log-likelihood of -18.71). Regarding it as mechanistically valid by

482   removing the fitted aggregation parameter accepts the assumption $r = 1$ as true and justifies the poor

483   empirical fit (a log-likelihood of -75.50) of the resulting fractional Poisson model $P(N) = 0.7782 \times \left(1 - \right.$

484      $\left. exp(-N)\right)$ as a consequence of pathogen aggregation with a mean aggregate size of 9.312. This model is

485   spurious because the data were actually generated with no aggregation and $r = 0.1$, and it over-states

486   low-dose risks by 0.97 orders of magnitude (a factor of 9.312). Additional examples of spurious or

487   misapplied dose-response models compromising risk inferences are provided in the Supplementary

488   Content. These types of misrepresentations of dose-response can have grave consequences upon

489   decision-making in the water industry and beyond, and demonstrate that critical thinking regarding dose-

490   response models needs to go well beyond the basic goal of good model fit. Additionally, future

491   experimental work could disprove a spurious model, leading to dramatic shifts in statistical inferences.

492   This example raises questions about injudicious use of the Akaike information criterion (AIC) and the

493   likelihood ratio test for model selection. The AIC is calculated as $AIC = -2 \times ln(L_{max}) + 2k$, where L_{max}

494   is the maximum likelihood and k is the number of free parameters in the model. Due to structural non-

495   identifiability, the aggregated fractional Poisson ($k = 2$) and aggregated exponential with immunity ($k =$

496   3) models share a maximum log-likelihood of -18.71. These models have AICs of 41.42 and 43.42,

497   respectively, so the spurious aggregated fractional Poisson model is preferred. Likewise, a likelihood ratio

498   test will favour the spurious aggregated fractional Poisson model because the more generalized

499   aggregated exponential with immunity model provides no improvement in fit. In each case, any model

500   with an arbitrarily fixed value of one of the non-identifiable parameters would be chosen. Once again, this

501   is acceptable if the model is used only as an empirical fit to the data in essentially the same conditions,

502   but is scientifically unsound in the case of a theoretically derived model with practically meaningful

503   parameters. In these cases, it is misguided to base model selection on fit alone. Modellers need to

504   recognize that sometimes a generalized model with more plausible assumptions is more appropriate,

505   even if structural non-identifiability precludes determination of fitted parameter values.

506   Bayesian analysis of a spurious model is also problematic, as Fig. 7 shows for the spurious aggregated

507   fractional Poisson model. Although the immunity parameter is estimated with good accuracy, the mean

508   aggregate size is not. The MCMC sample from the posterior (generated using OpenBUGS as described

509   above) asserts that the 95% credible interval for the mean aggregate size is 6.5–14.2. This misleading

510   conclusion arises from the aggregation parameter behaving wholly as a tuning parameter to compensate

511   for an incorrect assumed value of the host susceptibility among the non-immune ($r = 1$). The more non-

512   identifiable or weakly identifiable parameters a model has, the harder it is to root out incorrect

513   assumptions because there are more parameters behaving as tuning parameters to facilitate good fit.

514   **5. RECOMMENDATIONS FOR DEVELOPING OBJECTIVELY SUPPORTED MODELS**

515   Although the topic of parameter identifiability is familiar to most formally trained statisticians, this work

516   exemplifies the need for better awareness of non-identifiability among applied modellers, particularly in

517   the context of quantitative microbial risk assessment. Moreover, this work highlights the potentially

518   serious implications of ignoring non-identifiability in the design and approval of experiments, model-

519   fitting, and model-based inference, particularly in relatively mechanistic models where parameters have

520   important practical meaning rather than just being a means to achieve goodness-of-fit.

521   Evaluating identifiability of parameters in a mechanistic model (either algebraically or by simulation)

522   before an experiment is carried out can help to preempt wasteful experimental work that is inherently

523   incapable of yielding adequately informative data. This may be particularly necessary for experiments

524   requiring ethics approval. With simulated or experimental data, structural non-identifiability can be

525   explored using profile likelihood analysis or Bayesian analysis with uniform priors. Bayesian analyses of

526   models featuring structurally non-identifiable parameters (e.g. for *E. coli* O157:H7) should be viewed with

527   skepticism because they mask data that are otherwise too uninformative for statistical inference with a

528   subjective prior, and this can lead to bias. Although eliminating a redundant parameter may be sensible

529   in strictly empirical modelling, doing so injudiciously in a theoretically derived model (by assuming a value

530   of a practically meaningful parameter without rigorous justification) leads to spurious models that have

531   great fit but provide misleading mechanistic inferences. Sometimes a model deemed over-parameterized

532   by the Akaike information criterion or a likelihood ratio test has more realistic assumptions. Conversely,

533   almost any flawed assumption can be fit by adding more parameters to a model than the data can inform.

534   Thus, modelling must not be done haphazardly by focusing upon model fit alone, but must be done

535   prudently with careful consideration of all assumptions and recognition of structural non-identifiability

536   and its important implications. Given these major implications of structural non-identifiability upon

537   applied statistical modelling in mechanistic scenarios, the topic and its implications should at least be

538   noted in any probability and statistics text book featuring basic model fitting material such as goodness

539   of fit tests and the Akaike Information Criterion.

540   **ACKNOWLEDGEMENTS**

REFERENCES

Brouwer, A. F., Weir, M. H., Eisenberg, M. C., Meza, R., & Eisenberg, J. N. (2017). Dose-response relationships for environmentally mediated infectious disease transmission models. *PLoS Computational Biology, 13*(4), e1005481.

Catchpole, E. A., & Morgan, B. J. (1997). Detecting parameter redundancy. *Biometrika, 84*(1), 187-196.

Chik, A.H., Schmidt, P.J., & Emelko M.B. (2018). Learning something from nothing: The critical importance of rethinking microbial non-detects. *Frontiers in Microbiology, 9*2304.

Cochran, W.G. (1950). Estimation of bacterial densities by means of the" most probable number". *Biometrics, 6*(2), 105-116.

Cole, D.J., Morgan, B.J., & Titterington, D.M. (2010). Determining the parametric structure of models. *Mathematical Biosciences, 228*(1), 16-30.

Eisenhart, C., & Wilson, P.W. (1943). Statistical methods and control in bacteriology. *Bacteriological Reviews, 7*(2), 57-137.

Emelko, M.B., Schmidt, P.J., & Reilly, P.M. (2010). Particle and microorganism enumeration data: Enabling quantitative rigor and judicious interpretation. *Environmental Science and Technology, 44*(5), 1720-1727.

Haas, C.N., Rose, J.B., & Gerba, C.P. (2014). *Quantitative Microbial Risk Assessment* (2nd ed.). New York, NY: John Wiley and Sons.

Kreutz, C., Raue, A., Kaschek, D., & Timmer, J., 2013. Profile likelihood in systems biology. *The FEBS Journal, 280*(11), 2564-2571.

Little, M.P., Heidenreich, W.F., Li, G. (2010). Parameter identifiability and redundancy: Theoretical considerations. *PLoS ONE, 5*(1), e8915.

565   Maiwald, T., Hass, H., Steiert, B., Vanlier, J., Engesser, R., Raue, A., Kipkeew, F., Bock, H.H., Kaschek, D.,

566   Kreutz, C., & Timmer, J. (2016). Driving the model to its limit: Profile likelihood based model reduction.

567   *PLoS ONE, 11*(9), e0162366.

568   McCrady, M.H. (1915). The numerical interpretation of fermentation-tube results. *Journal of Infectious*

569   *Diseases, 17*(1), 183-212.

570   Messner, M.J., Berger, P., & Nappier, S.P. (2014). Fractional Poisson—A simple dose-response model for

571   human norovirus. *Risk Analysis, 34*(10), 1820-1829.

572   Nilsen, V., & Wyller, J. (2016). QMRA for drinking water: 1. Revisiting the mathematical structure of single-

573   hit dose-response models. *Risk Analysis, 36*(1), 145-162.

574   Petterson, S.R., Dumoutier, N., Loret, J.F., & Ashbolt, N.J. (2009). Quantitative Bayesian predictions of

575   source water concentration for QMRA from presence/absence data for *E. coli* O157: H7. *Water Science*

576   *and Technology, 59*(11), 2245-2252.

577   Pouillot, R., Hoelzer, K., Chen, Y., & Dennis, S. (2013). Estimating probability distributions of bacterial

578   concentrations in food based on data generated using the most probable number (MPN) method for use

579   in risk assessment. *Food Control, 29*(2), 350-357.

580   Prakasa Rao, B.L.S. (1992). *Identifiability in Stochastic Models: Characterization of Probability*

581   *Distributions*. San Diego, CA: Academic Press Inc.

582   R Core Team, (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R

583   Foundation for Statistical Computing. https://www.R-project.org/.

584   Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., & Timmer, J. (2009).

585   Structural and practical identifiability analysis of partially observed dynamical models by exploiting the

586   profile likelihood. *Bioinformatics, 25*(15), 1923-1929.

587   Schmidt, P.J., Pintar, K.D., Fazil, A.M., Flemming, C.A., Lanthier, M., Laprade, N., Sunohara, M.D., Simhon,

588   A., Thomas, J.L., Topp, E., Wilkes, G. (2013). Using *Campylobacter* spp. and *Escherichia coli* data and

589   Bayesian microbial risk assessment to examine public health risks in agricultural watersheds under tile

590   drainage management. *Water Research, 47*(10), 3255-3272.

591   Schmidt, P.J., Pintar, K.D., Fazil, A.M., Topp, E. (2013). Harnessing the theoretical foundations of the

592   exponential and beta-Poisson dose-response models to quantify parameter uncertainty using Markov

593   chain Monte Carlo. *Risk Analysis, 33*(9), 1677-1693.

594   Schmidt, P.J., Emelko, M.B., & Thompson, M.E. (2014). Variance decomposition: A tool enabling strategic

595   improvement of the precision of analytical recovery and concentration estimates associated with

596   microorganism enumeration methods. *Water Research, 55*, 203-214.

597   Schmidt, P.J. (2015). Norovirus dose–response: Are currently available data informative enough to

598   determine how susceptible humans are to infection from a single virus? *Risk Analysis, 35*(7), 1364-1383.

599   "Student" (Gosset WS). (1907). On the error of counting with a haemacytometer. *Biometrika, 5*(3), 351-

600   360.

601   Silvey, S.D. (1975). *Statistical Inference*. London, United Kingdom: Chapman and Hall Ltd..

602   Teunis, P., Takumi, K., & Shinagawa, K. (2004). Dose response for infection by *Escherichia coli* O157:H7

603   from outbreak data. Risk Analysis, 24(2), 401-407.

604   Teunis, P.F., Moe, C.L., Liu, P., Miller S.E., Lindesmith, L., Baric, R.S., Le Pendu, J., & Calderon, R.L. (2008).

605   Norwalk virus: How infectious is it? *Journal of Medical Virology, 80*(8), 1468-1476.

606   United States Environmental Protection Agency. (2014). *Microbiological Risk Assessment (MRA) Tools,*

607   *Methods, and Approaches for Water Media*; EPA-820-R-14-009. Washington, DC: Office of Water.

608   World Health Organization. (2016). *Quantitative Microbial Risk Assessment: Application for Water Safety*

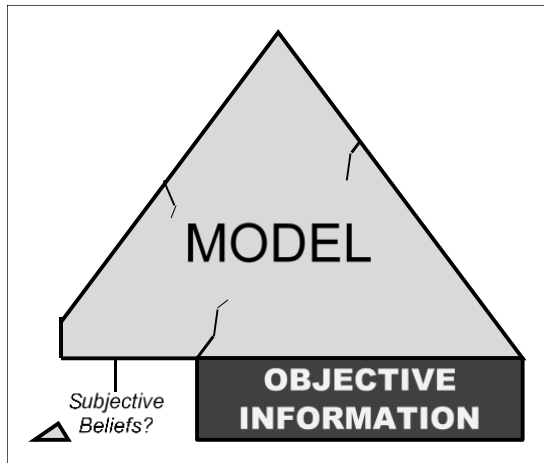609   *Management*. Geneva, Switzerland: WHO Document Production Services.

Fig. 1. A model that is not completely supported by a firm foundation of reliable objective information may depend too heavily upon subjective beliefs and can lead to incorrect scientific inferences. In such cases, collection of better data in the future could disprove flawed subjective assumptions underpinning the model.
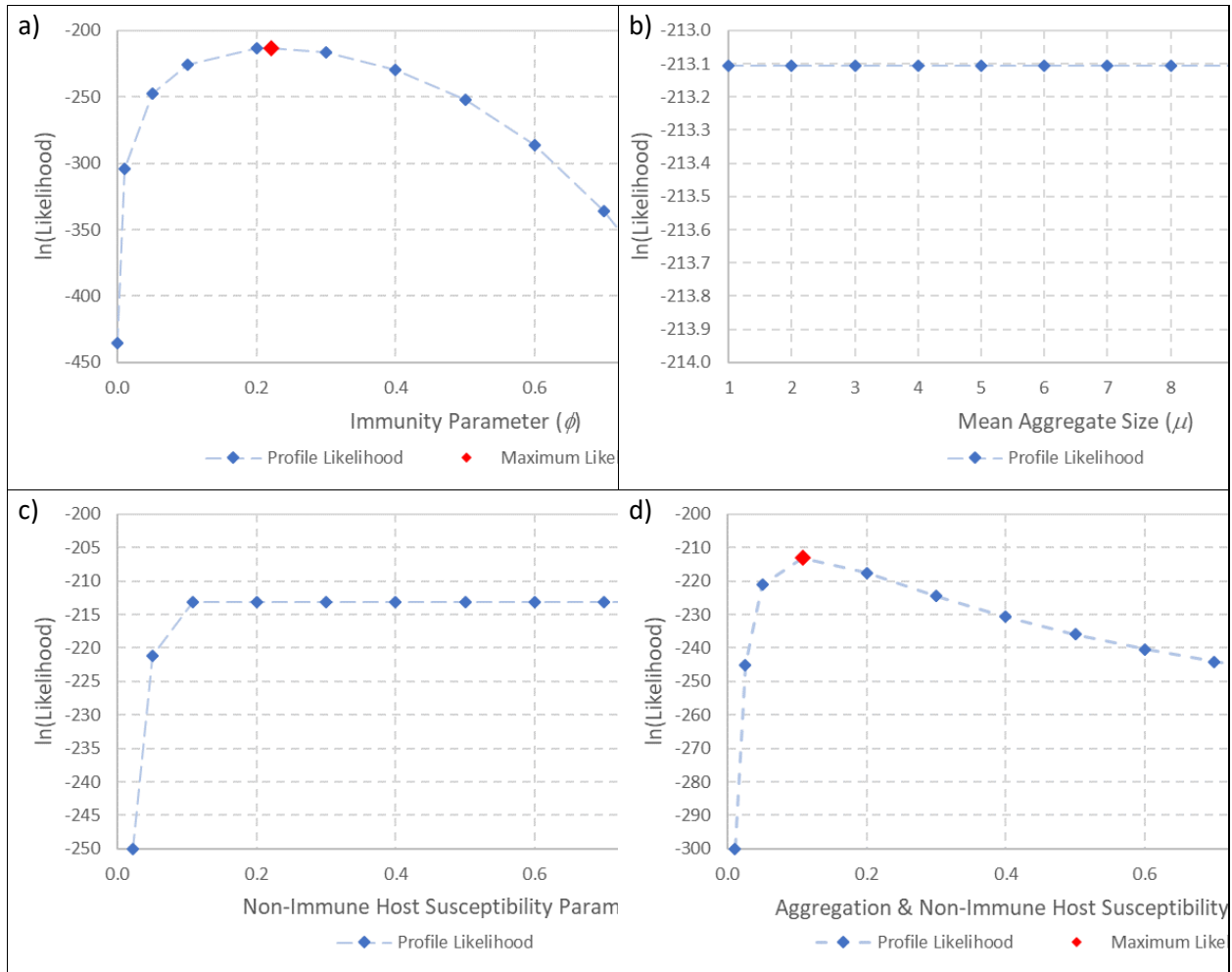
Fig. 2. The profile likelihood determined from fitting an aggregated exponential with immunity dose-response model to the Table 1 data is shown for a) the immunity parameter (ϕ), b) the mean aggregate size $\mu = -a/[(1-a)ln(1-a)]$, c) the non-immune host susceptibility parameter (r), and d) the parameter $\psi(a,r) = \frac{1-a}{a}ln\left(1+\frac{ar}{1-a}\right)$. Parameters ϕ, ψ are identifiable with maximum likelihood estimates as shown (0.22176 and 0.10738, respectively). Parameters μ, r are structurally non-identifiable, with plateaus across only a portion of the parameter space (i.e. $1 \le \mu \le 9.3$ and $0.1074 \le r \le 1$) illustrating set identifiability.
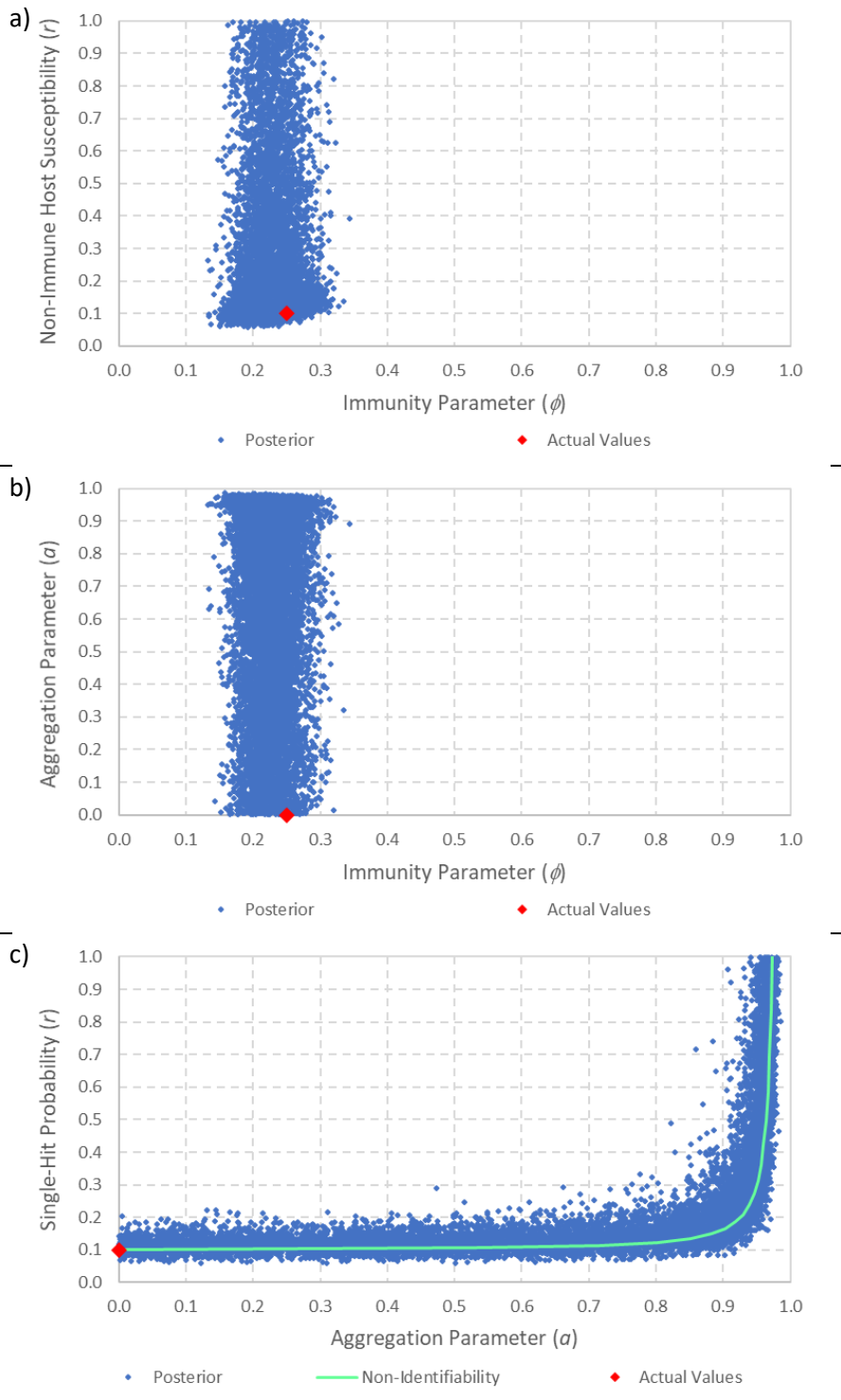
Fig. 3. Scatter plots obtained by MCMC are shown for a) ϕ, r, b) ϕ, a, and c) a, r based upon Bayesian analysis of the aggregated exponential with immunity dose-response model using uniform priors and the simulated data in Table 1. These plots are illustrative of the appearance of structural non-identifiability in MCMC results. The actual values of the parameters used to generate the simulated data are shown, as well as the structural non-identifiability characterized by $\psi(a, r) = \frac{1-a}{a} \ln\left(1 + \frac{ar}{1-a}\right)$. The identifiability of ϕ (i.e. clustering about a unique value of ϕ that is close to its actual value) and set identifiability of a, r (i.e. $0 \leq a \leq 0.9731$ and $0.1074 \leq r \leq 1$) are evident in these scatter plots.
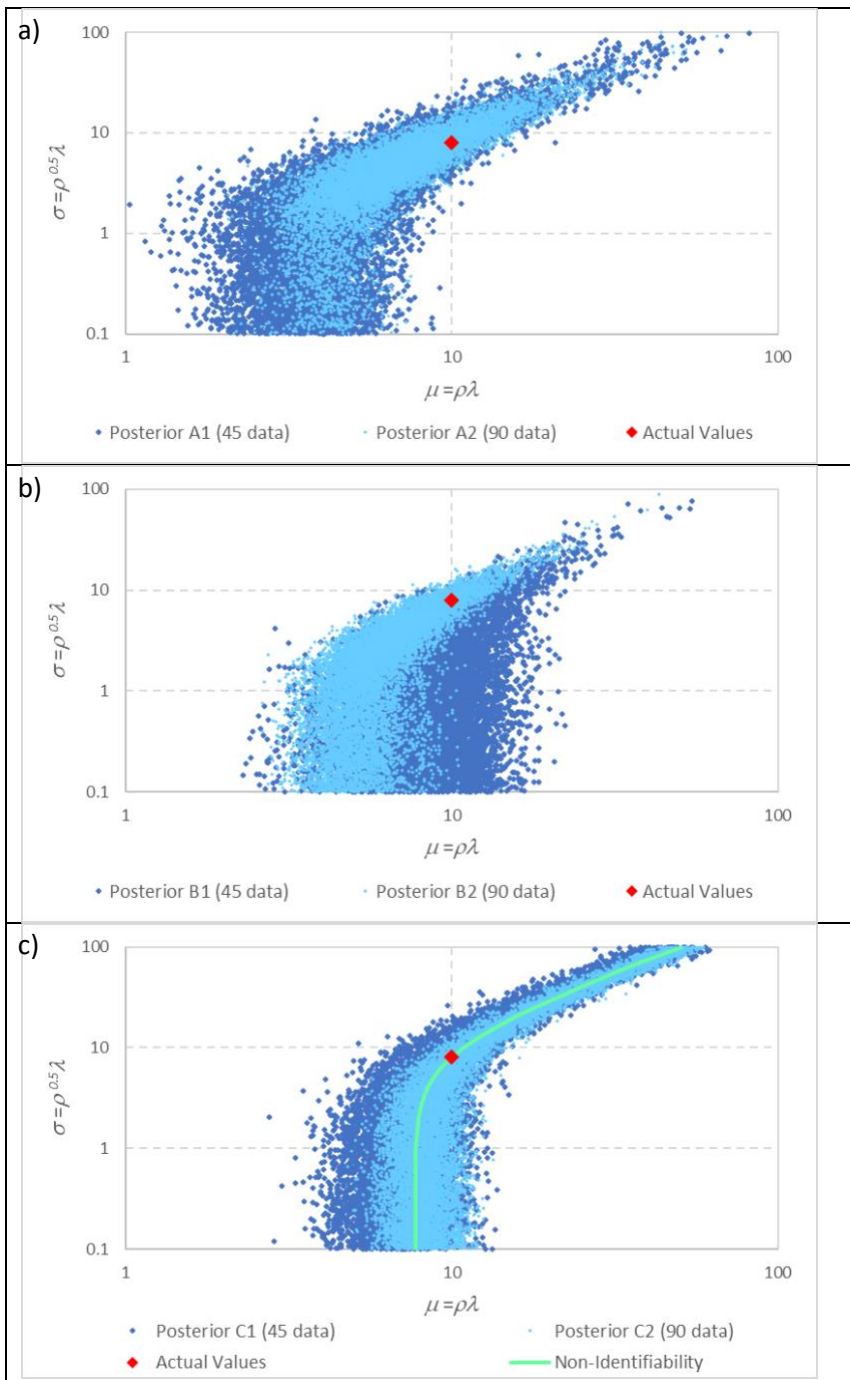
a)

| Posterior A1 (45 data) | Posterior A2 (90 data) | ◆ Actual Values |

b)

| Posterior B1 (45 data) | Posterior B2 (90 data) | ◆ Actual Values |

c)

| Posterior C1 (45 data) | Posterior C2 (90 data) |
| ◆ Actual Values | Non-Identifiability |

Fig. 4. Scatter plots show the mean (μ) and standard deviation (σ) of a gamma distribution for temporal concentration variability estimated from various types of presence/absence data. Three alternative experimental designs are considered as outlined in Table 2: a) five or ten samples each having three 1 mL aliquots, three 0.1 mL aliquots, and three 0.01 mL aliquots; b) 45 or 90 samples each with just one aliquot of varying volume (1 mL, 0.1 mL, or 0.01 mL); and c) 45 or 90 samples each with just one aliquot of 0.1 mL. The first experimental design is relatively informative (converging upon the actual values of the model's parameters with relatively few data), the second is weakly informative (an inefficient experimental design requiring more data to converge upon the actual values of the models parameters), and the third is structurally non-identifiable (an inappropriate experimental design from which it is fundamentally impossible to estimate the model's parameters regardless of how many data of this type are collected). The structural non-identifiability characterized by $\psi(\rho, \lambda) = 1 - (V\lambda + 1)^{-\rho}$ is shown.
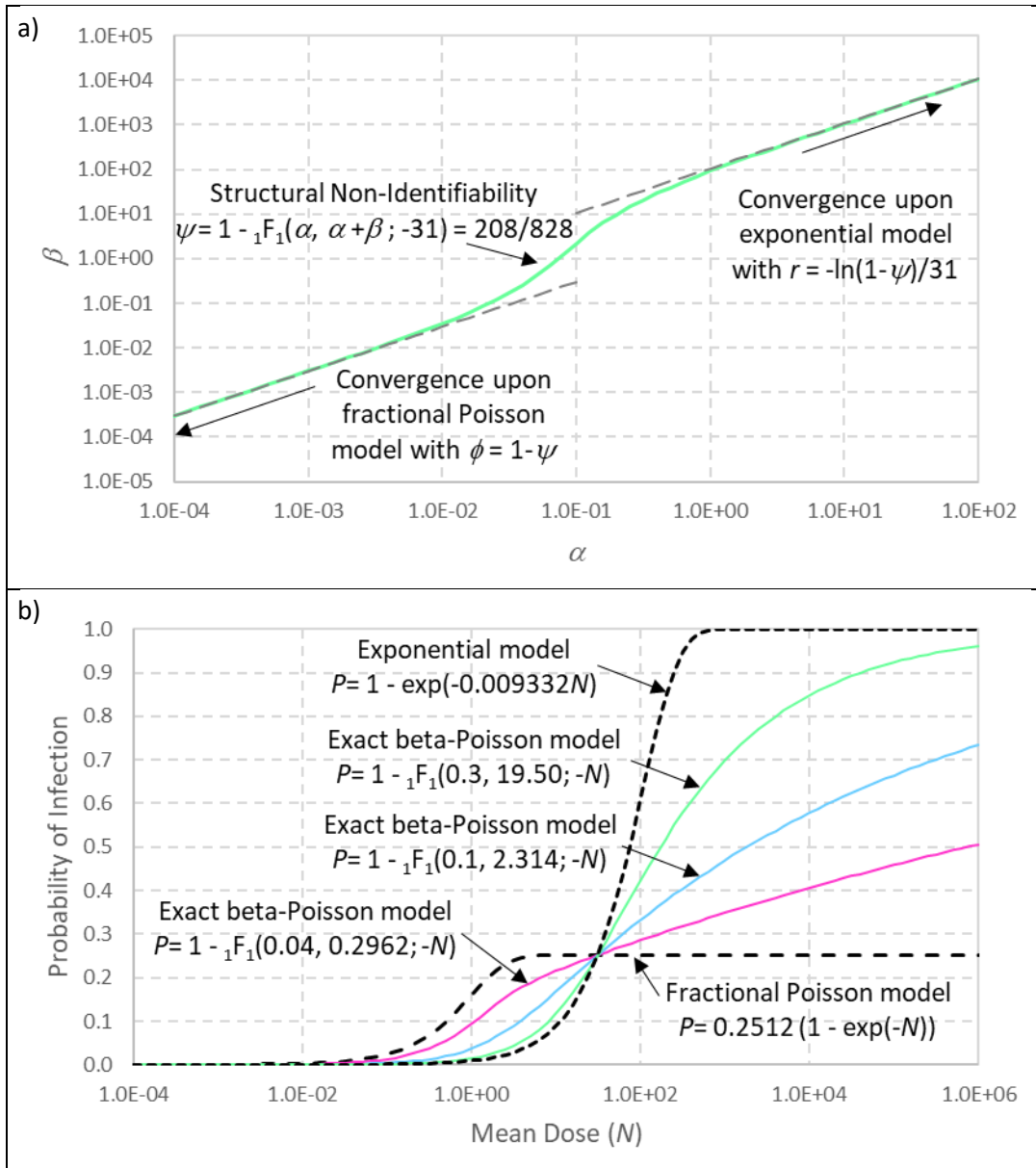
Fig. 5. Structural non-identifiability (a) results from fitting an exact beta-Poisson model to just the outbreak-based result that 208 of 828 pupils exposed to a mean dose of 31 *E. coli* O157:H7 became infected. A spectrum of models (b) passing through the point with a mean dose of 31 *E. coli* and a probability of infection of 208/828 are all equally supported by the data. The exponential and fractional Poisson dose-response models are extreme limiting cases of this spectrum (one with no immunity and constant host susceptibility, the other with immunity and complete susceptibility of non-immune hosts). Determining the optimal fit of the exact beta-Poisson dose-response model requires additional data (i.e. results for another value of the mean dose) or subjective information such as an assumed value of one of the parameters or an informative Bayesian prior.
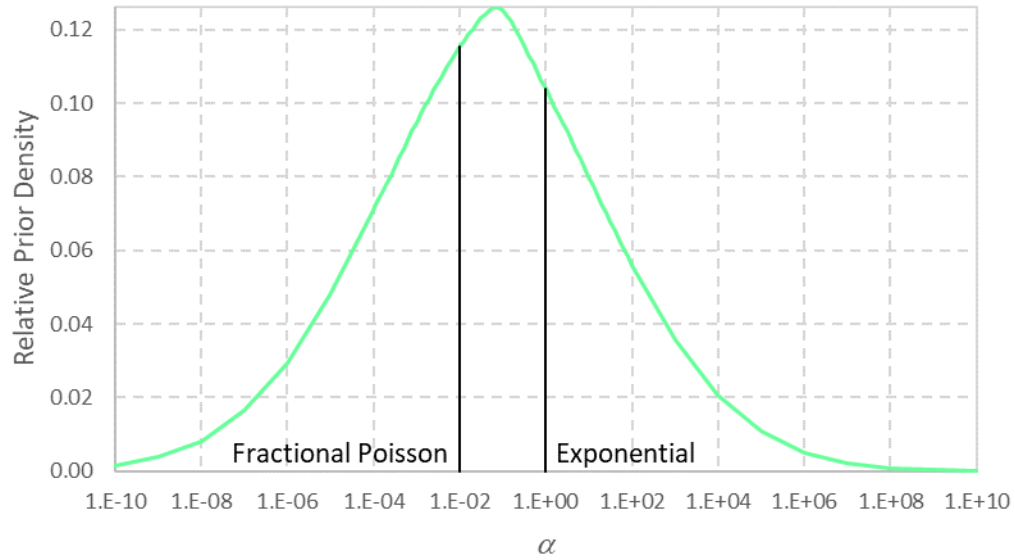
Fig. 6. The relative weighting of points along the structural non-identifiability curve $1 - {}_1F_1(\alpha, \alpha + \beta; -31) = 208/828$, given only the outbreak-based result that 208 of 828 pupils exposed to a mean dose of 31 *E. coli* O157:H7 became infected and the purportedly noninformative priors $\alpha/(\alpha + \beta) \sim uniform(0,1)$ and $log(\alpha + \beta) \sim normal(0,10)$, is shown. Also shown are the regions where the exact beta-Poisson model converges upon a fractional Poisson model (with $\phi = 0.2512$) or an exponential model (with $r = 0.009332$). The available data provide categorically no information to facilitate estimation of either α or β due to structural non-identifiability, and the apparent information illustrated in this plot arises exclusively from the informative prior.
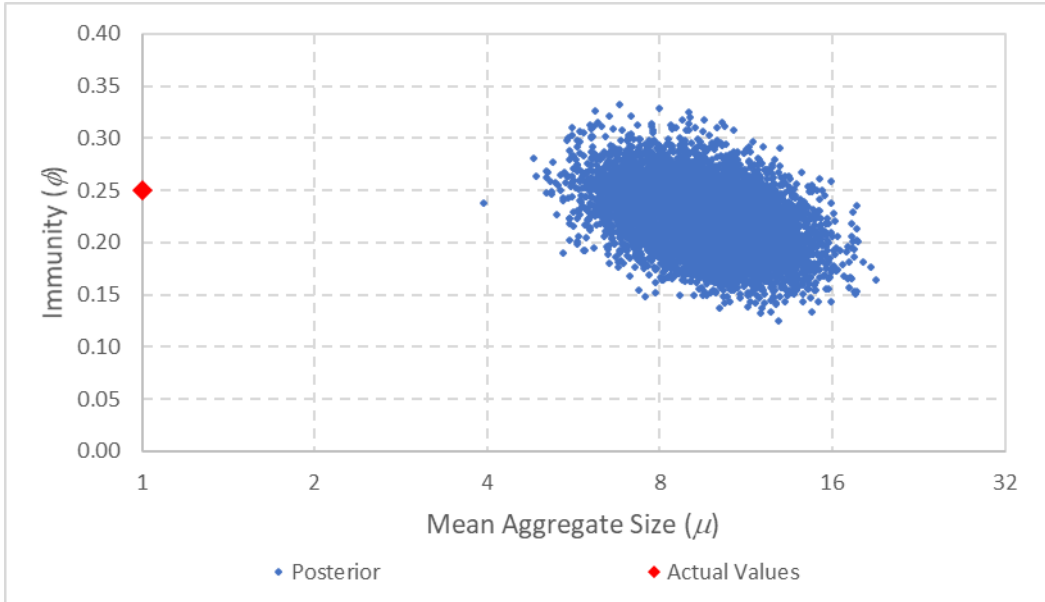
Fig. 7. The scatter plot obtained by MCMC is shown for Bayesian analysis of the aggregated fractional Poisson dose-response model using uniform priors and the simulated data in Table 1. The model fit is spurious because the region of interest of the posterior surface is shifted far away from the true parameters. This occurs because aggregation (represented by the mean aggregate size μ) and host susceptibility of the non-immune are structurally non-identifiable in the more general aggregated exponential with immunity dose-response model. Thus, the aggregation parameter behaves as a tuning parameter to allow the model to contort itself to fit the data despite an erroneous assumed value of the host susceptibility parameter (e.g. $r = 1$).

Table 1. Simulated Aggregated Exponential with Immunity Dose-Response Data

Mean Dose	Actual Probability of Infection[1]	Number of Subjects	Number Infected	Percent Infected
1	0.071372	50	7	14%
3	0.194386	50	8	16%
10	0.474090	50	27	54%
30	0.712660	50	34	68%
100	0.749966	50	41	82%
300	0.750000	50	40	80%
1000	0.750000	50	37	74%
3000	0.750000	50	40	80%

[1] Calculated using $(\phi = 0.25, a \rightarrow 0, r = 0.1)$

Table 2. Alternative Experimental Designs of Simulated Presence/Absence Data

Scenario	Sampling Events	Experimental Design	Notes
A1	5	3x1 mL, 3x0.1 mL, 3x0.01 mL for each sampling event	Parameters identifiable
B1	45	One analysis (1 mL, 0.1 mL, or 0.01 mL) for each sampling event	Parameters weakly identifiable
C1	45	One analysis (0.1mL) for each sampling event	Parameters structurally non-identifiable
A2	10	3x1 mL, 3x0.1 mL, 3x0.01 mL for each sampling event	Parameters identifiable Includes data from scenario A1
B2	90	One analysis (1 mL, 0.1 mL, or 0.01 mL) for each sampling event	Parameters weakly identifiable Includes data from scenario B1
C2	90	One analysis (0.1 mL) for each sampling event	Parameters structurally non-identifiable Includes data from scenario C1

Table 3. Alternative Models Fit to Table 1 Dose-Response Data

Model Name	Immunity (ϕ)	Aggregation (a)	Host Susceptibility (r)	Fitted Model	log-Likelihood
Fractional Poisson	36.46%	disaggregated[1] (known)	1 (assumed)	$P(N) = 0.6354 \times \left(1 - exp(-N)\right)$	-58.03
Exponential with Immunity	22.18%	disaggregated[1] (known)	0.1074	$P(N) = 0.7782 \times \left(1 - exp(-0.1074N)\right)$	-18.71
Aggregated Fractional Poisson	22.18%	0.970395 (μ=9.312)[2]	1 (assumed)	$P(N) = 0.7782 \times \left(1 - exp(-N/9.312)\right)$	-18.71
Aggregated Exponential with Immunity	22.18%	Non-identifiable: $\hat{\psi}_{MLE} = \frac{1-a}{a} ln\left(1 + \frac{ar}{1-a}\right) = 0.1074$			-18.71

[1] The special case of disaggregated microorganisms in the family of aggregated exponential with immunity dose-response models is mathematically represented by a limit as $a \rightarrow 0$

[2] $\mu = -a/[(1-a)ln(1-a)]$