ARE WE OVER-MODELLING UNDER-INFORMATIVE DATA?

A research overview with implications for drinking water risk assessment

WATER Science, Technology & Policy Group
UNIVERSITY OF WATERLOO

KEY MESSAGES

- Poorly designed experiments can lead to structural nonidentifiability, which results in model parameters that can never be estimated from this type of data
- Bayesian methods effectively fabricate information about nonidentifiable model parameters and mask uninformative data
- Nonidentifiability can lead to spurious models that fit data well but yield unreliable mechanistic inferences and predictions

Concentration is nonidentifiable if volume plated or dilution factor were not recorded

AUDIENCE

This research overview is designed for:

- Modellers & water scientists using models
- Water and public health officials who develop policy or make decisions based on models

> "In problems of statistical inference, estimation of a parameter is not meaningful unless it is identifiable"
>
> – **B.L.S. Prakasa Rao** (1992)

WHY WAS THIS DONE?

Structural nonidentifiability means that several alternative mechanisms explain the data equally well, and additional data of the same type can never provide meaningful information about nonidentifiable model parameters. As a result, when fitting statistical models to data, several sets of parameter values share the best fit to the data.

Some types of data and experimental designs are inherently not good enough to allow estimation of all the mechanistically important parameters in a model. The purpose of this study was to explore implications of structural nonidentifiability, particularly in quantitative microbial risk assessment (QMRA) for drinking water.

APPROACH

Three QMRA-related models were used to explore implications of nonidentifiability. Structural nonidentifiability was proven algebraically and also illustrated graphically using profile likelihood analysis and Bayesian analysis with uniform priors. Two examples arise from fitting a two-parameter model to essentially one datum: an example related to repeated presence-absence analyses and a published *E. coli* O157:H7 dose-response model fit to outbreak data. The last example arises from a published norovirus dose-response model fit to data from an experiment in which uncontrolled virus aggregation precludes estimation of important model parameters.

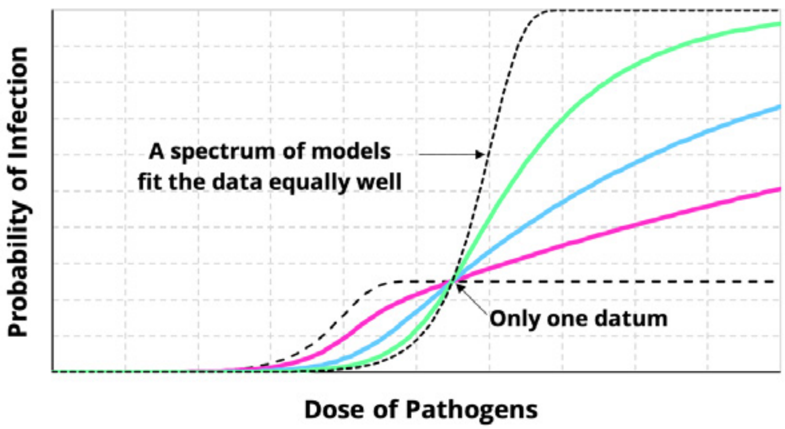
FINDINGS

1 Evaluation of parameter identifiability using simulated data can help to ensure that experimental designs are capable of yielding informative data

2 Even relatively uninformative priors have undue influence upon Bayesian inferences about nonidentifiable parameters

3 It is possible to assume an erroneous value of a nonidentifiable parameter with no loss of fit of the model to the data, leading to a spurious model

Fig. 1: Example of nonidentifiability



> In cases of nonidentifiability, “it will not be possible to use the model to predict, with reasonable accuracy an unobserved, physically meaningful, model output”
>
> – **Guillaume *et al.*** (2019)

IMPLICATIONS

For Experimental Design: Identifiability analysis can avert wasteful experimentation leading to data that are uninformative about important model parameters. Such analysis should be required for experiments needing ethics approval to ensure generated data can be informative.

For Bayesian Analysis: These analyses should include discussion of parameter identifiability and greater justification of the prior used. Such analysis of nonidentifiable models is unduly influenced and potentially biased by the subjective prior and can conceal uninformative supporting data.

For Model Selection: Soundness of mechanistic assumptions should be central to modelling because basing model selection on fit alone can lead to spurious models. For example, some dose-response models fit data well but could overestimate risks from pathogens in drinking water by orders of magnitude.

WATER Science, Technology & Policy Group

About the Authors

Philip J. Schmidt, PhD, AStat
Monica B. Emelko, PhD, Canada Research Chair
Water Science, Technology & Policy Group
Department of Civil & Environmental Engineering
University of Waterloo, Ontario, Canada

Mary E. Thompson, PhD
Department of Statistics & Actuarial Science
University of Waterloo, Ontario, Canada

This overview is adapted from: Schmidt PJ, Emelko MB, and Thompson ME., Recognizing Structural Nonidentifiability: When Experiments Don’t Provide Information about Important Parameters and Misleading Models Can Still Have Great Fit. Risk Analysis (2020), https://doi.org/10.1111/risa.13386.