Chadwick, Fergus J. (2023) *Complex observation processes in ecology and epidemiology: general theory and specific examples.* PhD thesis.

https://theses.gla.ac.uk/83512/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk/
research-enlighten@glasgow.ac.uk

Complex Observation Processes in Ecology and Epidemiology

General Theory and Specific Examples

Fergus J Chadwick

A thesis presented for the degree of
Doctor of Philosophy

School of Biodiversity, One Health and Veterinary Medicine
University of Glasgow
UK
November 2022

For Bonnie, the best dog.

You rescued me at the start of my PhD. Stress, deadlines, even the pandemic were manageable with you by my side. Losing you was harder for me than the rest combined.

Abstract

Complex observation processes abound in ecology and epidemiology. In order to answer the large-scale, urgent questions that are the focus of modern research, we must rely on indirect and opportunistic observation. Relating these data to the biological processes we are interested in is challenging. Statisticians working in this area need an understanding of both state-of-the-art modelling techniques and the field-specific nuances of how the data were generated. As a result, many methods to deal with complex observation processes are highly bespoke. Bespoke models are hard to translate between contexts and, because they are often presented in field-specific language, hard to learn from. Modelling of observation processes is thus a fractured area of study, leading to duplication of research effort and limiting the rate at which we can make progress.

In this thesis, I aim to provide a road-map to how we might achieve some unification in this area. I begin by establishing a conceptual framework that can be used to describe observation processes and identify methods to address them. The framework defines all observation processes as a combination of issues of latency, identifiability, effort or scaling (L.I.E.S.). I illustrate the framework using motivating examples from ecology and epidemiology. The risk with conceptual frameworks is that they can be over-fitted to existing data and may fail when faced with new, real-world problems.

To address this, I also approach the problem from a bottom-up perspective by tackling a series of ecological and epidemiological case studies. Each case study requires novel statistical methods to deal with the observation process. By developing new methods, I explore the world of observation processes potentially not well-captured in the literature. I then explore whether these case studies motivate revision or reassessment of my conceptual framework.

While the case studies were chosen to challenge the L.I.E.S. framework, I find that they mutually reinforce each other. The framework provides a helpful scaffolding with which to describe the problems in the case studies. The case studies provide useful examples of more complex observation processes and how the four issues encoded in L.I.E.S. interact with one another. These findings illustrate the value of a framework for unifying approaches to observation processes.

Contents

List of Tables

List of Figures

Acknowledgments

Before arriving at the University of Glasgow to do my Master's, I had spent the year as a boots-on-the-ground field scientist picking flowers in the glorious Burren region of the Republic of Ireland. During this time, I realised that I could not answer the questions I wanted to without a better understanding of statistics. My friend Aaron Westmoreland told me about the incredible Master's course he had taken in Glasgow and that he thought I would get a lot of out of it. In doing so, he changed the entire course of my life and career, and I could not be more grateful. During my Master's, I went from having no knowledge of statistics to it being my passion. I will be forever grateful to Roman Biek and Barbara Mable who organised and administered the degree.

Through this programme, I met my future supervisors Dan Haydon and Jason Matthiopoulos. Dan, you ignited the statistical spark for me. The course you ran is the gold-standard in my head for teaching. Indeed, with your compassion and ethics as a scientist, supervisor and leader you set the gold-standard far beyond teaching. Jason, after Dan lit the spark you, more than anyone, helped it and me grow. Your patience, faith and example helped me develop as a scientist and gave me the courage to take risks I could not have asked for a better mentor or a better friend. I must also thank your lovely family, Valia and Spyros, who have provided me with delicious food and great company so many times. At the start of my PhD, I also picked up two more supervisors, Dirk Husmeier and Frances Mair. Dirk, you held me to a high standard and pushed me to push myself exactly when I needed it. Your willingness to stick with me until I understand has helped me grasp concepts I thought were far beyond my capabilities. Model development sessions with you and Jason remain my favourite part of the PhD. Frances, thank you for giving us such an insight into the world of medical research.

I would also like to thank Jason's research group and the many members of the School of Biodiversity, One Health and Veterinary Medicine and the School of Mathematics and Statistics who provided a supportive environment to develop as a researcher. In particular, without my friends Claire Harris, Crinan Jarrett, Andy Seaton, Sara Gandy, Peter Mortensen, David Pascall, Yacob Haddou, Luca Nelli, Halfan Ngowo, Cat Swedberg and Heather McDevitt the PhD would have been a lot harder and infinitely less fun. I have also been lucky to collaborate with several incredible scientists who have each broadened my perspective, notably Katie Hampson, Otso Ovaskainen, and David McAllister. Finally, I would not have made it to even starting a PhD without my family who supported and encouraged me. Thank you for sticking with me and helping me get to where I am today.

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, this thesis is the result of my own work and has not been submitted for any other degree or professional qualification at the University of Glasgow or any other institution.

Fergus J Chadwick
November 2022

Chapter 1

Introduction

1.1 Background

The urgency and scale of modern research questions in ecology and epidemiology has never been greater. Problems like how to mitigate climate change, biodiversity loss, COVID-19, stretch our research capacity to breaking point while demanding rapid and reliable solutions. Yet our resources remain finite and our ability to deploy the expensive, high quality, systematic survey that research has historically relied upon has not, and cannot, grow to meet these challenges. The modern scientist must thus choose wisely where and when to deploy structured data collection, and supplement these with opportunistically collected data. By definition, these data (such as citizen science records [1], hospital admission records [2], eDNA [3]) are indirect routes to answering the questions we are interested in. However, the indirect route is our only option to stretch our limited resources far enough (we hope) to address the problems we face.

Observation problems are not unique to ecology and epidemiology but these two subject areas are rich grounds for their study. In both fields, observational studies have been commonplace throughout their history, particularly for applied inference [4; 5]. In conservation management and public health, for example, the systems being studied are hard to distill into controlled experiments. Indirect observation (relative to our inferential target) may be less invasive and thus reduce the risk of disturbance, or, as is in the case with GPS tagging, allow us to follow portions of a process that would not be visible through conventional methods [6]. In the case of using citizen science observations, there are additional societal benefits from involving non-professional scientists in data collection, from increasing awareness of a particular problem to the general scientific literacy of the public [7]. As a result, researchers in these fields have been dealing with such data for some time and are familiar with their strengths and weaknesses.

In statistics, we often structure our inferential models to reflect what we believe to be the data-generating process. The biological process model is often the centre of these models and contain the parameters we are interested in interpreting. For complex data, the biological process model must be combined with an observation process model which describes how we believe the biological process has been distorted. Observation process models are relatively new. Systematic survey/experimental design exists to identify and minimise the distorting effect of observation processes, and hence generate data as close to the biological phenomenon of interest as possible, opportunistic data sources provide no such guarantees. Opportunistic data may have only loose relevance to the biological phenomenon and even direct observations may be unreliable when filtered, for example, through the eyes of a citizen scientist or an insensitive testing method. These complex issues need to be addressed during inference [8]. Therefore, while the modern scientist has limited power to increase data collection they must instead make up for the shortfall with more sophisticated data analysis.

Fortunately, data analysis is easier to scale than systematic data collection. Once a class of problem has been identified and solutions developed, it is relatively easy to apply the solutions to new instances of the problem. A serious impediment in this process of scaling and generalisation is that observation processes are not currently well-defined. As a result, even when two observation processes are closely related, the solutions are often developed in parallel and

rarely shared.

So what can we do to address this methodological fragmentation? We can take a top-down approach, one that establishes an overarching framework of ideas, a typology of problems that seeks to organise both our methodology and terminology along a minimal set of canonical, pure-form problems, each with its own toolbox of solutions. Alternatively, we might take a bottom-up approach in which we seek out a diverse collection of case studies and make particular efforts to recognise their commonalities and differences. As discussed below, both approaches have their strengths and neither is without weaknesses. As a result, in this thesis I attempt to develop both approaches simultaneously, using each to inform the other.

1.1.1 Top-down approaches: Sorting through a mixed bag

The development of an overarching framework or theory for understanding observation processes is an appealing but challenging goal [9]. The appeal is easy to see: a unified typology of problems that seeks to organise both our methodology and terminology along a minimal set of canonical, pure-form problems, each with its own toolbox of solutions would allow us to break down the existing methodological siloes. More challenging combinations of these archetypal problems might then also become more tractable, by breaking solutions down to individual components, and thus allow progress to accelerate. However, research into observation processes has not yet, to my knowledge focused on developing such a framework. There are several possible reasons for this.

Firstly, existing overviews of observation processes tend to have different but important aims. Many treatments of are not aimed at producing statistical tools to tackle them. For example, many classic treatments on the topic of observation processes are philosophical in nature, rather than methodological. These treatments are useful and inspiring but offer few practical solutions [10; 11]. Others are targeted at justifying and standardising the use of non-systematic data and offer some practical solutions but do not attempt to be comprehensive [12]. Some treatments have the opposite aim and focus on how broad a range of observation processes a given class of methods can be applied to, e.g. [13; 14].

Secondly, the consequences of the problems caused by observation processes are not well-understood or appreciated. Researchers need to be convinced that observation processes are a key issue that needs to be addressed. Often, this work focuses on highlighting examples in a key application area such as citizen science [15; 1], fisheries [16] or observational medical studies [17]. These studies are insightful, but generally keep solutions siloed within the area of question. Fundamentally, concerns about observation processes require ecologists and epidemiologists to weigh bias and uncertainty against expediency [18; 19], a difficult and demanding task that we are not always well equipped to tackle [20; 21].

Although these works do not attempt to create the typology described above, they do help us understand what a successful typology may look like. The focus of the synthesis must be well chosen: a philosophy of science approach is perhaps too broad to be directly applied while more focused treatments do not, by their very nature, help us break down methodological siloes. Ecology and epidemiology are both similar enough to share lessons and broad enough to contain

many existing siloes. The synthesis must also build upon the large body of literature from both the statistical and applied perspective. The large body of existing work in this area means that researchers are interested in this topic, and by engaging with that literature, we provide a ready bridge to bring them on board. Finally, the impact of observation processes on inference must be made clear. The framework will only be used if people believe it helps solve a genuine problem.

I build on these lessons in Chapter 2, "L.I.E.S. of Omission: Complex observation processes in Ecology and Epidemiology". There, I develop a synthesis of observation processes in which I posit that all observation processes in ecology and epidemiology can be described in terms of four constituent problems: latency, identifiability, effort and scale. I demonstrate each issue with pure-form motivating examples drawn from the applied literature, review the statistical methodology for tackling these problems, and highlight the issues of ignoring each. I also outline several miniature case studies of how to apply this framework to real-world problems.

While I believe the synthesis I have proposed overcomes the main challenges in developing a framework, it cannot overcome the fundamental limitations of a top-down approach. The primary disadvantages of this approaches are that any framework must be conceived of *a priori* from known observation processes and known approaches to tackling them. The framework must, therefore, be under constant review and checked against emerging case studies. In essence, the top-down approach is not sufficient and must be paired and tested against the bottom-up approach.

1.1.2 Bottom-up approaches: Proceeding by Case Studies

In the bottom-up approach, a series of case studies are explored in which observation processes pay a key role. In contrast to the top-down approach, the risk of this approach is that we do not improve upon the methodological fragmentation that already exists in the literature. By developing the case studies alongside the L.I.E.S. framework, the two can feed into and inform one another. In the conclusions, I discuss whether pursuing the two approaches leads to the realisation that the case study belongs to a recognisable class of broader problems, or to the refinement of the broader framework by counterexample.

The choice of case studies within the confines of a thesis were important. To challenge the framework robustly, I chose case studies which needed new methodological development to accommodate the observation process. While the framework uses issues of latency (the relevance and relatedness of the data collected and the biological process), identifiability (the reliability of the parameters inferred), effort (the reliability of the observations made) and scaling (the relevance of the parameters inferred to the biological process) as the four key tenets of observation processes, I selected case studies that had at least a component of latency (although this does not preclude them from having any of the other issues, or in having previously unconsidered issues). Latency is where the observations are of variables that are only indirectly linked to the phenomenon of interest. With issues of latency, the key is in modelling the transformation of the observed variable(s) onto the biological variable of interest.

The motivation for focusing primarily on issues of latency is to maximise the

impact with which we engage the research community. The key to effective modelling of observation processes, and in communicating the risks of not modelling them, is to thoroughly engage both field scientists and statisticians in the process. I believe that issues of latency are the key to doing so at this early stage. Firstly, I believe that latency is the most accessible of these concepts for non-statistical scientists. Proposing mechanistic links between processes is a core skill that many scientists find engaging and intuitive. Secondly, research in latent variable modelling is well-developed in the statistical literature, with entire textbooks devoted to the topic [22; 23; 24]. The key to effective modelling of observation processes is in bridging the gap between statistical and non-statistical scientists, and focusing on areas where they are both already comfortable makes this easier.

1.1.3 Toolbox

Benefits of the Bayesian Paradigm for Observation Process Modelling

One of the key aims of observation process modelling is to ensure propagation of uncertainty (particularly when the uncertainty is heterogeneous). While this is possible under most statistical paradigms, it is most natural using Bayesian methods. In propagating uncertainty naturally, the construction of models with multiple components becomes much easier. For example, in many models, the observation process will be a sub-model that adjusts the biological process model. These models are both important and they both inform each other, so the ability to fit these models jointly with shared uncertainty is extremely beneficial.

The use of priors, unique to the Bayesian paradigm, is both conceptually insightful and practically useful in modelling observation processes. When modelling observation processes, we have data which are informative about the biological process we are interested in but contain some structured uncertainty. Using these data to construct priors is both natural and internally consistent, allowing the observation component of the model to be interpreted easily alongside the biological model.

Practically, priors are also useful tools to improve model identifiability and sampling. Many of the quantities that we are interested in inferring, such as effort, do not necessarily have natural units of measurement. In this context, priors can be extremely useful for constraining parameter space in a reasonable way and making inference possible.

Univariate vs Multivariate Modelling

One of the most common issues with opportunistic data is that there is not necessarily a single, clear response variable. Particularly in latent models, we are often trying to impute our biological variables of interest from a series of other response variables. For example, we might try to understand the abundance of one species that we have only partially observed using the abundance of other species that were observed more thoroughly. Each species abundance is a function of the environment and of interactions with other species [25]. The same problems crop up in missed diagnoses in patients with multiple chronic diseases [26], spatial smoothing problems [27], or in trying to complete a difficult-to-make measurement, such as biomass, based on correlated easy-to-make measures, such as plant

heigh and diameter [28]. Sometimes researchers choose to ignore these collinearities to simplify their models (often to their detriment [29]) but with complex observation processes we generally want to exploit collinearity to help inform missing or partially observed data.

We therefore need modelling structures that can accommodate both covariates and multiple, interacting response variables [30]. There are broadly two classes of models that we can use in a GLM-type framework: univariate models in which the response distribution has a single dimension and multivariate where it has multiple dimensions. In both responses, the inclusion of covariates is relatively trivial through the use of a linear predictor. The two classes diverge in terms of how they allow the response dimensions to interact.

In the univariate approach, multiple models are fit simultaneously with the responses from each model being incorporated in the linear predictor of the others. In this way, each response variable is able to influence the others. This approach quickly increases in complexity, however, when we allow multiple responses to interact with each other in their effect on another response. For example, if we have a patient with multiple chronic diseases, the likelihood of them having a given condition does not increase linearly with the presence of other conditions. Some other conditions will facilitate each other, some will be protective against additional conditions. It is therefore insufficient to have each response affecting the others additively, the responses interact. This can be accommodated in the univariate approach through the use of interaction terms but this quickly becomes unwieldy, especially as the number of interacting variables increases.

In the multivariate approach, the response variables do not interact with each other in the linear predictor (which now just contains covariates) but through a covariance matrix. The covariance matrix determines how the multiple response dimensions are oriented with respect to one another. As the covariance matrix contains all the other response dimensions, they can be evaluated simultaneously. To draw an analogy with the univariate approach, the inverse of the covariance matrix (the precision matrix) gives the partial covariances (i.e. the covariance between two dimensions independent of the others). This is equivalent to including the dimensions alone in the linear predictor in the univariate setting. The covariance matrix contains the full covariance in which each dimension can influence the others. In essence, the covariance structure bakes in the higher order interaction terms in a single structure.

An additional advantage to linking interactions using covariance matrices is that they have well-known properties, both algebraically and in terms of expert knowledge. As a result, they are often much easier to sample and set priors for than the equivalently specified univariate models. For example, although relationships between response dimensions do not need to be symmetrical, there are impossible combinations of relationships. In univariate higher order interactions, this is hard to enforce so must be learnt through sampling which is expensive. In the multivariate approach, this is achieved by ensuring the covariance matrix is positive definite. Our understanding of the covariance matrix means there are efficient sampling methods (for example using Cholesky factors and improved algorithms [31]) and useful summary statistics (for example, matrix sparsity can be summarised using its determinant).

Software Implementation

In most cases, observation process models will be coupled with biological process models (whose parameters are of scientific interest). A joint modelling framework is therefore a natural way to share uncertainty and information between the two processed. To implement efficient sampling algorithms such complex models, it is necessary to have a flexible probabilistic programming language (PPL) with which to specify them. There are many PPLs (Stan [32], Nimble [33], JAGS [34], PyMC [35], and Turing [36], to name a few) available with different advantages and disadvantages.

First, the different languages have very different attitudes to error checking and messaging. Computational faithfulness of MCMC sampling is not a given for any model and many traditionally used diagnostics (e.g. r-hat) have been found to be insensitive to numerous sampling problems [37] and some MCMC algorithms, e.g. those based on Hamiltonian Monte Carlo, have unique warnings, such as divergent transitions [38]. For this reason, opinionated PPLs that deploy extensive diagnostics and error messaging are invaluable. Second, not all PPLs can be easily accessed by all programming languages. While R [39] is probably the most-used programming language in ecology and epidemiology, many researchers use Python [40] or Julia [41] too. Thirdly, Bayesian statistics and programming are difficulty and having a healthy community built around your chosen PPL is a huge boon.

For these reasons, I use Stan for the bulk of my analyses as: it can be called from most major programming languages, it has excellent sampling diagnostics and messaging, and the community is large, friendly and active. One limitation of Stan is that it is not possible to sample discrete parameters directly. While this can be overcome by marginalising discrete parameters, this is not always feasible.

Chosen Case Studies

The three case studies chosen reflect different degrees of understanding of the observation process, from Chapter 3 in which the observation process is fully known to Chapter 5 in which there are multiple competing, plausible processes. While two of these chapters address problems in COVID-19, the observation process in each is quite different, and the third is drawn from the ecological literature.

In Chapter 3, "COVID-19-exploring the implications of long-term condition type and extent of multimorbidity on years of life lost: A modelling study", the observation process is fully known: records of correlated binary covariates have been summarised as marginal counts. Data aggregations like this are particularly common in the medical literature, where individual-level records are sensitive so summaries are produced to guarantee patient privacy. The context is also important here. The data were collected at the start of the COVID-19 pandemic and the focal population were individuals who had died of the disease in Italy. I could be confident that I had almost census-level coverage of the biological process I was interested in (as all deaths in Italy are recorded and causes of death attributed) and full knowledge of the observation process (the data aggregation scheme).

In Chapter 4, "Combining rapid antigen testing and syndromic surveillance improves community-based COVID-19 detection in a low-income country", I integrate two separate observation processes to make predictions about the biolog-

ical process. Although the mechanism by which each observation process occurs here, the degree of overlap between the two is not well understood so the model needed to be flexible in how the two data sources are integrated. The application of these methods to disease detection in Bangladesh underscores the value and potential of imperfect data sources for public health, particularly in such resource-limited settings.

In Chapter 5, "Do identification guides hold the key to species misclassification by citizen scientists?", there are many, non-exclusive mechanisms by which latency could plausibly be introduced in the observation process. In this case study, rather than modelling each plausible mechanism individually, I compare several flexible models that can accommodate the different mechanisms. The key innovation in this chapter is the use of an informative prior on the observation process in the form of the species identification guide used by the citizen scientists.

References

[1] A. Johnston, E. Matechou, and E. B. Dennis, "Outstanding challenges and future directions for biodiversity monitoring using citizen science data," *Methods in Ecology and Evolution*, 2022.

[2] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nature Reviews Genetics*, vol. 13, no. 6, pp. 395–405, 2012.

[3] M. E. Cristescu and P. D. Hebert, "Uses and misuses of environmental dna in biodiversity science and conservation," *Annual Review of Ecology, Evolution, and Systematics*, vol. 49, no. 1, pp. 209–230, 2018.

[4] J. P. Vandenbroucke, "When are observational studies as credible as randomised trials?," *The Lancet*, vol. 363, no. 9422, pp. 1728–1731, 2004.

[5] S. C. Anderson, P. R. Elsen, B. B. Hughes, R. K. Tonietto, M. C. Bletz, D. A. Gill, M. A. Holgerson, S. E. Kuebbing, C. McDonough MacKenzie, M. H. Meek, *et al.*, "Trends in ecology and conservation over eight decades," *Frontiers in Ecology and the Environment*, vol. 19, no. 5, 2021.

[6] G. L. Shillinger, H. Bailey, S. J. Bograd, E. L. Hazen, M. Hamann, P. Gaspar, B. J. Godley, R. P. Wilson, and J. R. Spotila, "Tagging through the stages: technical and ecological challenges in observing life histories through biologging," *Marine Ecology Progress Series*, vol. 457, pp. 165–170, 2012.

[7] R. C. Jordan, H. L. Ballard, and T. B. Phillips, "Key issues and new approaches for evaluating citizen-science learning outcomes," *Frontiers in Ecology and the Environment*, vol. 10, no. 6, pp. 307–309, 2012.

[8] R. King, "Statistical ecology," *Annual Review of Statistics and its Application*, vol. 1, pp. 401–426, 2014.

[9] S. P. Otto and A. Rosales, "Theory in service of narratives in evolution and ecology," *The American Naturalist*, vol. 195, no. 2, pp. 290–299, 2020.

[10] N. R. Hanson, *Patterns of discovery: An inquiry into the conceptual foundations of science*. CUP Archive, 1965.

[11] P. Feyerabend *et al.*, *Against method*. Verso, 1993.

[12] A. Underwood, M. Chapman, and S. Connell, "Observations in ecology: you can't make progress on processes without understanding the patterns," *Journal of Experimental Marine Biology and Ecology*, vol. 250, no. 1-2, pp. 97–115, 2000.

[13] B. T. McClintock, R. Langrock, O. Gimenez, E. Cam, D. L. Borchers, R. Glennie, and T. A. Patterson, "Uncovering ecological state dynamics with hidden markov models," *Ecology letters*, vol. 23, no. 12, pp. 1878–1903, 2020.

[14] M. Auger-Méthé, K. Newman, D. Cole, F. Empacher, R. Gryba, A. A. King, V. Leos-Barajas, J. Mills Flemming, A. Nielsen, G. Petris, *et al.*, "A guide to state–space modeling of ecological time series," *Ecological Monographs*, vol. 91, no. 4, p. e01470, 2021.

[15] N. J. Isaac, A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy, "Statistics for citizen science: extracting signals of change from noisy ecological data," *Methods in Ecology and Evolution*, vol. 5, no. 10, pp. 1052–1060, 2014.

[16] J. Harwood and K. Stokes, "Coping with uncertainty in ecological advice: lessons from fisheries," *Trends in Ecology and Evolution*, vol. 18, no. 12, pp. 617–622, 2003.

[17] E. Von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, J. P. Vandenbroucke, and S. Initiative, "The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies," *Annals of Internal Medicine*, vol. 147, no. 8, 2007.

[18] S. Canessa, G. Guillera-Arroita, J. J. Lahoz-Monfort, D. M. Southwell, D. P. Armstrong, I. Chadès, R. C. Lacy, and S. J. Converse, "When do we need more data? a primer on calculating the value of information for applied ecologists," *Methods in Ecology and Evolution*, vol. 6, no. 10, pp. 1219–1228, 2015.

[19] S. L. Maxwell, J. R. Rhodes, M. C. Runge, H. P. Possingham, C. F. Ng, and E. McDonald-Madden, "How much is new information worth? evaluating the financial benefit of resolving management uncertainty," *Journal of Applied Ecology*, vol. 52, no. 1, pp. 12–20, 2015.

[20] D. D. Murphy and B. D. Noon, "Coping with uncertainty in wildlife biology," *The Journal of Wildlife Management*, pp. 773–782, 1991.

[21] E. Milner-Gulland and K. Shea, "Embracing uncertainty in applied ecology," *The Journal of Applied Ecology*, vol. 54, no. 6, p. 2063, 2017.

[22] A. A. Beaujean, *Latent variable modeling using R: A step-by-step guide*. Routledge, 2014.

[23] W. H. Finch and B. F. French, *Latent variable modeling with R*. Routledge, 2015.

[24] J. Loehlin and A. Beaujean, *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis, Fifth Edition (5th ed.).* Routledge, 2017.

[25] O. Ovaskainen, G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego, "How to make more out of community data? a conceptual framework and its implementation as models and software," *Ecology Letters*, vol. 20, no. 5, pp. 561–576, 2017.

[26] C. J. Whitty and F. M. Watt, "Map clusters of diseases to tackle multimorbidity," 2020.

[27] A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes, *Handbook of Spatial Statistics*. CRC press, 2010.

[28] S. Roxburgh, K. Paul, D. Clifford, J. England, and R. Raison, "Guidelines for constructing allometric models for the prediction of woody biomass: how many individuals to harvest?," *Ecosphere*, vol. 6, no. 3, pp. 1–27, 2015.

[29] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, *et al.*, "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 36, no. 1, pp. 27–46, 2013.

[30] M. M. Mayfield and D. B. Stouffer, "Higher-order interactions capture unexplained complexity in diverse communities," *Nature Ecology and Evolution*, vol. 1, no. 3, pp. 1–7, 2017.

[31] D. Lewandowski, D. Kurowicka, and H. Joe, "Generating random correlation matrices based on vines and extended onion method," *Journal of Multivariate Analysis*, vol. 100, no. 9, pp. 1989–2001, 2009.

[32] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of Statistical Software*, vol. 76, no. 1, 2017.

[33] P. de Valpine, D. Turek, C. J. Paciorek, C. Anderson-Bergman, D. T. Lang, and R. Bodik, "Programming with models: writing statistical algorithms for general model structures with nimble," *Journal of Computational and Graphical Statistics*, vol. 26, no. 2, pp. 403–413, 2017.

[34] M. Plummer *et al.*, "Jags: A program for analysis of bayesian graphical models using gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, vol. 124, pp. 1–10, Vienna, Austria., 2003.

[35] A. Patil, D. Huard, and C. Fonnesbeck, "Pymc: Bayesian stochastic modelling in python," *Journal of Statistical Software*, vol. 35, no. 4, 2010.

[36] H. Ge, K. Xu, and Z. Ghahramani, "Turing: a language for flexible probabilistic inference," in *International Conference on Artificial Intelligence and Statistics*, pp. 1682–1690, PMLR, 2018.

[37] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner, "Rank-normalization, folding, and localization: an improved r for assessing convergence of mcmc," *Bayesian Analysis*, vol. 16, no. 2, 2021.

[38] M. Betancourt, "Diagnosing suboptimal cotangent disintegrations in hamiltonian monte carlo," *arXiv preprint arXiv:1604.00695*, 2016.

[39] R. C. Team, "R: A language and environment for statistical computing (version 4.0. 5)[computer software]," *R Foundation for Statistical Computing*, 2021.

[40] G. Van Rossum and F. L. Drake Jr, *Python Reference Manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[41] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017.

Chapter 2

L.I.E.S. of Omission: Complex observation processes in Ecology and Epidemiology

2.1 Highlights

- In ecology and epidemiology, the observation process (how we collect data) can be as complex as the biological process we are investigating.
- Complex observation processes require us to make inference using data that can be highly abstracted from or limited in their coverage of the biological process.
- Observation processes are often described in application-specific language - this makes it challenging for field-scientists to communicate problems to statisticians and for statisticians to identify pre-existing solutions.
- These challenges often lead to the observation process not being modelled at all (resulting in misleading inference), or to inadvertent re-invention of a pre-existing solution.
- We propose a typology that unifies how we describe both the observation processes and the statistical machinery used to address them.

2.2 Abstract

Advances in statistics mean that it is now possible to tackle increasingly complex observation processes. Advances in data collection techniques mean that this is now essential. Methodological research to make inference about the biological process while accounting for the observation process has expanded dramatically, but solutions are often presented in field-specific terms, limiting our ability to identify commonalities between methods. We suggest a typology of observation processes that could improve translation between fields and aid methodological synthesis that is comprehensive, orthogonal, intuitive, grounded and memorable. We propose the LIES framework (defining observation processes in terms of issues of latency, identifiability, effort and scale) and illustrate its use with both canonical, simple examples and more complex case studies.

2.3 Glossary

biological process: The target of inference for the ecologist, encompassing all topics of ecological study.

controlled experiment: An experiment focusing on a particular relationship between response and explanatory variables, where as many as possible of the confounding (nuisance) variables are kept constant.

ecological phenomena: See *biological process*.

effort (issues of): The amount and distribution of observations of the phenomenon of interest.

functional forms: The mathematical specification of a model. Any difference between the output of models with the same functional forms are due to different parameterisation, different initial conditions, or stochasticity.

generative modelling: a model which is meaningfully decomposed into interpretable parameters and sub-models, and from which data can be simulated.

hidden state/variable: See latent state/variable.

identifiability (issues of): The ability of a model to make inference about the relationships between its components.

latency (issues of): Where some or all parts of the *biological process* are not directly observed, and thus inference must be made indirectly through its impact on observable parts of the system.

latent state/variable: A state or variable that is not directly observed but must be inferred using observable parts of the system (i.e. *manifest variables*).

manifest state/variable: A state or variable which can be directly observed and measured.

non-transferability: A situation where a model fits observations well in the "here-and-now" but predicts poorly in novel situations. The problem can result from under- or overfitting, or from the violation of stationarity assumptions.

observation processes: The method by which an *ecological phenomenon* is recorded as data.

proxies: A *manifest state/variable* that has a well-defined functional relationship with a *latent state/variable*.

scaling (issues of): Any discrepancy between the resolution or extent (of, e.g. space, time, or taxonomy) at which the data are collected and the process of inferential interest occurs.

sensitivity analysis: A mapping between the magnitude of perturbations to a model's inputs and the consequent disturbances produced to its outputs.

typology: In general, refers to the classification of observations according to their characteristics. Here, it refers to a minimal set of characteristics that can be used to describe any observation problem.

2.4 The Increasing Complexity of Observation Processes in Ecology and Epidemiology

Modern ecologists are called upon to tackle crises in the environment, as well as deal with ongoing scientific tasks of data collection and analysis. Technological advances in our ability to collect and analyse observations should give us

unparalleled capacity to address emerging crises, but, instead, we are frequently stymied by the overwhelming scope and complexity of analysing our ever-more complex and multifaceted data. Techniques for collecting data have become almost as complex as the underlying **biological processes** we are trying to understand via these observations. Environmental DNA [1], remote sensing [2], biologging [3] and citizen science [4] data all help get us closer to the spatial, temporal and taxonomic coverage we need to meet contemporary ecological challenges. However, they also introduce complexities which need to be addressed through sophisticated statistical analyses that are often devised as dedicated solutions to particular data sets. Therefore, a counterpart analytical crisis results from the fact that statistical methods that pay proper attention to these difficulties can appear disconnected, specialist and overly technical. As a result, advanced methods are rarely shared between fields, leading to needless duplication of solutions and inhibiting us from identifying methodological gaps that could benefit many fields.

Current solutions to these crises are thin on the ground. As ecology transmutes into a "hard science" [5], part of the solution is to encourage ecologists to become more quantitative [6; 7]. While statistical literacy is arguably higher than ever amongst applied ecologists, we must still rely on close collaboration with statisticians for method development. Alternatively, the analytical crisis can be circumvented by relying more heavily on experimental design. Many classical statistical techniques were developed for designed experiments, involving careful controls of confounders, a high number of replicates and unbiased measurements. Unfortunately, the nature and scale of ecological questions in the 21st century are not always amenable to experimental design. GPS-tagged animals do not remain within pre-defined study areas, citizen scientists have to reconcile their observation efforts with their day-jobs and, crucially, there is no Latin square for climate change. The focus on experimental design and user-friendly statistical methods can lead researchers to, assume-away the more challenging features of their data, to analyse them as if they were gathered in a designed experiment, yielding conclusions that are neither robust nor reproducible (see statistical golems from [8]).

Realistically, therefore, we cannot simplify the methods or the data needed. However, we believe that we can simplify **observation process** modelling by developing a shared **typology** of associated problems. A **typology** will provide a basis for discussions between statisticians and field scientists to elicit what **observation processes** might occur or have occurred, and it creates a set of axes onto which existing methods can be placed. The latter makes it easier to see which methods are closely and distantly related in the problem space, rather than as discrete classes of statistical models. By presenting the problem space in this way, it becomes easier to iterate between model types during development and makes methodological synthesis much easier. If there are large methodological gaps in the problem space, areas ripe for the development of new statistical techniques are revealed. If methods occupy the same point in problem space, this highlights methods which are known by multiple names allowing associated research to be unified. Unification often leads to rapid progress as previously disjointed efforts become focused, techniques shared and crucial gaps identified, as seen in the synthesis of biodiversity metrics under Hill numbers [9], MaxEnt [10] and presence-only models [11] under point processes [12], and

dependent mixture/latent Markov/Markov-switching/regime-switching/state-switching/multi-state models under Hidden Markov Models [13]. A successful typology, therefore, helps identify the observation processes at play, navigate the possible solutions, and direct methods development to where it will be most productive.

2.5 The L.I.E.S. Framework

A shared **typology** for observation problems needs to meet the following criteria. It must be sufficient to describe all observation problems in ecology (*comprehensive*). To make the **typology** efficient, each problem type must exist independently and be able to be used in combination to describe more complex types of problems (*orthogonal*). The problem types need to be understandable to both field scientists and statisticians (*intuitive*) and should be rooted in the existing methodological literature where possible (*grounded*). Finally, the framework will be most effective if it is widely adopted which requires friendly packaging (*memorable*).

Observation problems can be introduced during either data collection or analysis. There are are two sides to these problems: the relevance of the observation to the biological process and the reliability of the the observations made. This motivates a (*comprehensive*) typology of four core concepts:

	Relevance	**Reliability**
Data Collection	*Latency* - what is the relationship between the variables collected and those in the biological process?	*Effort* - what is the breadth and depth of the observations made?
Data Analysis	*Scaling* - what does the scale at which the parameter is estimated mean biologically?	*Identifiability* - what is the uncertainty in our parameter estimates (finite or otherwise)?

Below, we define each of the concepts in non-technical language (*intuitive*). We illustrate each with pure-form motivating examples (*orthogonal*) rooted in the statistical literature (*grounded*). We make these canonical examples simple but realistic and present the concepts using the moniker "L.I.E.S. of Omission", reminding us that failure to model **observation processes** correctly is to risk dishonesty in our analyses (*memorable*). Finally, we draw from publications across the literature to demonstrate that the framework can describe real-world observation problems as either elemental of compound manifestations of these primitive types (*comprehensive* and *grounded*, see Boxes 1-3).

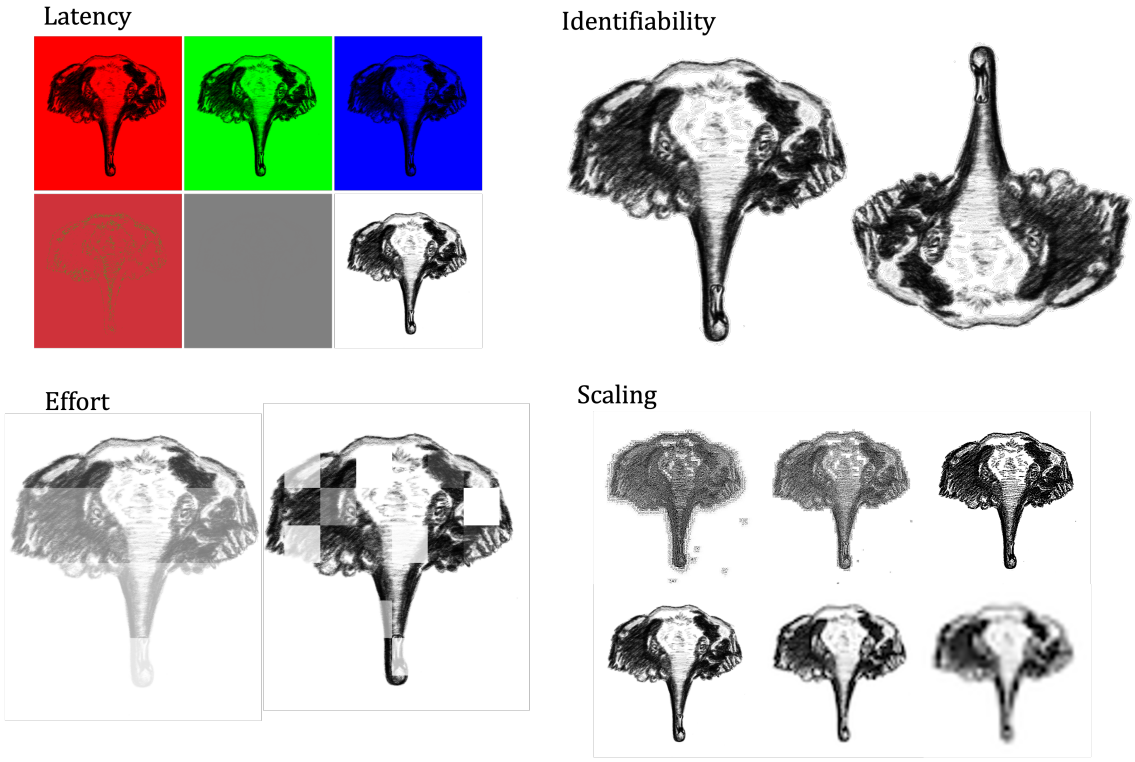

Figure 2.1: The Elephant in the Room: These panels illustrate how the four types of observation process can affect the same image of an elephant. **Latency**: The elephant is recorded as six manifest variables, namely the red, green, blue, hue, saturation and light layers of the image. The latent variable, the elephant, is a combination of either the first three or the second three colour layers. **Identifiability**: There are equally plausible views of this image as either an elephant or a swan, illustrating mathematical identifiability. **Effort**: The grid cells of the elephant picture are observed with different degrees of effort, giving us a clearer idea of some parts of the picture to others. In the left-hand image, there is a bias top to bottom of the image, with very little effort in observing the tip of the trunk. In the right-hand image, effort is less obviously structured, with seemingly better observations of the ears and trunk tip. **Scaling**: The relevant scale for biological inference may be different to the scale at which data were collected. We can think of the scaling process as a process of data aggregation (for a coarser scale) or disaggregation (for a finer scale). A pixel in an image can only have a single value, so splitting up pixels produces a set of pixels whose mean value is that of the original pixel. Aggregating pixels is the reverse process, where pixels with distinct values are combined into a single pixel with the mean value of the constituent pixels. In the Scaling panel, the top left image has the smallest pixel size and the bottom right has the largest pixel size. When the pixel size is too small (top left of panel), noise is introduced and the image becomes harder to recognise. When the pixel is too large (bottom right of panel), the pixels homogenise key details, again making the image harder to recognise.

2.6 Latency - What the Observer Sees

2.6.1 Motivation

Biological phenomena are often hard to observe directly. Sometimes this is due to practical constraints. In principle, it may be possible to weigh the dry biomass of a group of organisms in a habitat, however, it is often more feasible (and less destructive) to measure a related variable such as the dimensions of the organisms while alive. In other cases, the phenomenon we are interested in is not directly observable, perhaps because it is conceptual in nature (such as an ecosystem equilibrium or autocorrelation) or has ceased to be observable (for example, historical species abundance). In such cases, we need to infer the relevant **latent** quantity via its causal or correlational links to **manifest** (observable) quantities. For example, an equilibrium might be inferred from the direction and speed with which an ecosystem might be observed to be moving towards or away from it [14; 15]. Similarly, historical abundances might have dynamical consequences that persist into the observable present day [16]. As illustrated, **latency** is a continuum from small to large degrees of discrepancy between the **manifest** and relevant **latent** variables.

Where the degree of **latency** is small (e.g. the dimensions in the biomass example above), the **manifest** variable can be thought of as a **proxy** variable. **Proxies** can be mapped onto **latent** variables linearly or with known **functional forms**. **Functional forms** can often be motivated from biological understanding and quickly out-perform models that assume linear relationships. For example, the trophic connection between predator and prey often does not depend only on the density of the prey [17], but also that of the predator [18]. To quantify predator intake, therefore, we need to use prey numbers and our understanding of how the two interact. Other examples of such **proxies** abound in the allometry literature where scaling laws are known not only functionally, but also parametrically.

The most challenging situation occurs when the **latent variable** represents a **hidden state** which needs to be inferred from multiple **manifest variables** simultaneously. For example, multiple covariates might determine a single **hidden variable**, which may have a complex relationship with one or more response variables [19; 20]. Alternatively, a **hidden state** might only be inferable via successive observations of **manifest states** in a dynamical system, as is often the case with state-space models [21]. For example, in a time series of biologging observations we can rarely observe an individual animal's internal intentions (the **hidden state** here might be behaviour, see Box 2)[22]. Instead, we must rely on inferring these based on environmental context (the covariates) and the geometry of movement (the **manifest** response variable). Crucially, a model that does not explicitly include the **hidden state** may be poorer, or simply unable to capture the process. For example, disease eradication programs often need to account for contributions to transmission from hidden disease reservoirs [23].

2.6.2 Existing Statistical Methods

Entire statistical textbooks are written on **latent variable** models [24; 25; 26]. Fundamentally, the aim of any **latent variable** method in ecology is to map that which is easily observed into a space that is biologically meaningful. It is, therefore, useful to think of both **latency** and the models to tackle **latency** existing on a continuum. In the case of **proxies** it may be possible to accommodate these relationships using simple generalised linear models (GLMs). Coefficients can help transform scalar **proxies** to the **latent variable** they represent. **Functional forms** may be approximated with link functions or, where they are well defined mathematically from biological first principles, with the known functional response.

While **hidden states** often require sophisticated modelling structures, it is useful to start from the simplest form: the generalised mixed effects model (GLMM) or hierarchical model. Random effect structures in GLMMs correspond to distributional assumptions about complex latent phenomena for different groupings of the data [27]. Stepping up in complexity slightly, multi-level hierarchical models (nested GLMMs) use information from different levels of the data to constrain the **latent variables**. For example, state-space like hidden Markov models (see Box 2) have been developed for the reconstruction of stochastic time series of **hidden states** [13].

The key to effectively tackling latency is to improve our biological understanding of the **latent** phenomenon. **Latent variables** are often hardest to estimate and interpret when they are only weakly constrained by prior knowledge and model structure. By imposing boundaries informed by, for example, expert prior elicitation [28], we can often improve both computation and inference.

2.7 Identifiability - What Signal the Model Detects

2.7.1 Motivation

We build statistical models to identify relationships. The richness with which we can describe these relationships by our models will depend on the model definition. If the model is well-defined and the data contain sufficient signal, the parameters will both capture the real relationships and exclude alternative explanations. Advances in statistical computing have removed many constraints on model specification which makes specifying interesting, biologically relevant models easier (often by thinking of the problem **generatively**). Even then, however, not all relationships are identifiable by all models (**mathematical identifiability**) and many relationships will only be identifiable with sufficient data (**practical identifiability**) [29].

Mathematical identifiability issues can arise even in simple situations. In single-species population ecology, we know that population growth can be written unambiguously as a balance equation between birth and death rates. However, even with unlimited data sets on growth it is not possible to obtain estimates of births and deaths since there are infinite plausible combinations that are consistent with any given growth rate. As model complexity grows, **mathematical identifiability** problems can be much more subtle (see discussion of multi-collinearity in [30]).

Practical identifiability problems result from trying to make inference from data. Even mathematically identifiable models may be unable to estimate relationships with precision if the noise-to-signal ratio is high for the number of available data points, there is strong collinearity in the data or there are large degrees of separation between the data and the parameters being estimated.

The severity of identifiability issues may depend on the model's purpose. When interpreting parameters to make inference, identifiability is essential. When a model is purely for prediction, the identifiability of an individual parameter may not matter so long as the effect of that parameter is identifiable [31]. Similarly, sometimes a parameter may be identified when normalised or transformed. For example, a covariance matrix may not be identifiable but the corresponding correlation matrix is [32].

2.7.2 Existing Statistical Methods

The relationship between the model definition and the quantity of interest defines both types of **identifiability** problem. We can think of the models working in two directions. In the forward direction, we simulate from the model. In the inverse direction, we estimate model parameters using data. We can use the forward direction to identify issues of **mathematical identifiability** by testing whether simulated quantities are affected by the specific model parameters [33]. If changing the model parameter values does not affect the quantities generated, there are **mathematical identifiability** issues.

Once we have ruled out **mathematical identifiability** issues, we can explore the inverse direction. Here, we use data on the quantity of interest to estimate the model parameters. If many parameterisations are equally plausible given the data, we have high uncertainty and **practical identifiability** issues. Methods to assess these problems have been unified under the topics of **sensitivity analysis** [34] and **uncertainty quantification** [35; 36; 37], respectively.

Sensitivity analysis is solely a function of the model definition (i.e. is not affected by data), and there are a plethora of available diagnostics (both analytic and simulation-based). Directed acyclic graphs (DAGs) are an increasingly popular method for finding which relationships are identifiable (particularly within the causal inference literature [38]). Similarly, writing the formal mathematical definition of a model can help highlight the conditions under which a model is identifiable. Simulation-based methods are often more approachable and generalisable, particularly if the underlying model is **generative** in nature [39]. Although **mathematical identifiability** problems are data-invariant, they are often found when fitting to data if the model is not checked before (see Box 2). For example, in MCMC-based algorithms, correlations between parameter samples, slow sampling and chains failing to converge are often indicators of **mathematical identifiability** problems [40].

Uncertainty quantification depends on both the model definition and the noise-to-signal ratio in the data. If the model is over-parameterised or the data are uninformative then there will be high uncertainty in the parameter estimates. The concept of power analysis exists in both the Frequentist [41] and Bayesian [42] paradigms, albeit with very different motivations, interpretations and implementations. In both cases, we are interested in using simulated data to estimate

the nature and amount of data required to identify a relationship to a given precision. These simulations may use the exact model definition, but it is important to also assess how uncertainty changes under model-mispecification (for instance, by using surrogate models for simulation)[43].

Model mis-specification is almost certain when addressing real world problems, so model and variable selection methods are also key tools for addressing **practical identifiability** [44]. Aiming to choose the most parsimonious model that, for example, optimises an objective function (e.g. model likelihood, goodness of fit, prediction error) also tells us which relationships are not being estimated with precision. The literature for model and variable selection is large and contentious but broadly comes down to treating model-space as continuous or discrete. Continuous model-space methods carry out variable selection parametrically as part of the model-fitting process, for example, penalised complexity [45] or slab-and-spike priors in the Bayesian paradigm and ridge regression or LASSO in the Frequentist [46]. Discrete model-space methods involve fitting candidate models independently and choosing a preferred model based on a separate criterion to that used within model fitting. The most commonly used metrics for discrete model-space selection are the information criteria [47; 44], although other metrics or objective functions are also widely used. Continuous model-space approaches benefit from internal logical consistency but can be computationally burdensome and challenging to implement for non-nested models, particularly where models take different **functional forms**.

2.8 Effort - Where the Observation Happens

2.8.1 Motivation

When planning data collection, the aim is to try to gather as information rich observations of the **biological process** as possible with the minimum bias and maximum precision. As such, a key tenet of traditional experimental design is to spread observation effort evenly among sampling units. In doing so, the **observation process** does not distort the underlying process. Outwith controlled conditions, an even distribution of effort is almost impossible, leading to over-recording of some parts of the system (e.g., seasons, years, spatial regions, individuals, population classes) and under-representation of others.

Uneven **effort** often arises from practical constraints. There are limits to where observers can be sent for safety reasons or due to administrative boundaries. Sometimes unevenness is deliberate. For example, data collected alongside a rabies vaccination campaign will generally be targeted towards rabies hotspots. In these cases, stratified effort is uneven but its distribution is known and can be accounted for in the analysis.

The situation is more complex in platform-of-opportunity data. The distribution of citizen scientists (see Box 1)[48], fisheries by-catch surveys [49] or deer-vehicle collisions [50] are all driven by processes that are rarely measured directly and are often driven by multiple other processes. Sometimes these drivers are spatial (e.g. deer-vehicle collisions often depend on traffic flow; fishing boats need to minimise their travel time to fish stocks; and citizen scientists like to record near to where they live and in attractive locations). Often, there are also

cultural drivers of what is reported - legal penalties may reduce reporting of deer-vehicle collisions or of fisheries by-catch, while citizen scientists are sometimes only keen to report ”interesting” findings such as rare or invasive species. As a result, we frequently need to analyse data where the distribution of effort is not only uneven but also unknown.

2.8.2 Existing Methods

Uneven effort can be accounted for statistically, after the collection of data. In principle, to retrieve the biological process from our data we simply need to consider observations per unit effort as the response variable. We can think of this as an offsetting exercise [51], however, first we need to quantify effort across different sample units. The challenge of this grows with the degree to which effort is unknown. Where effort is fully known, the offset can be incorporated into the model as data.

Where effort is in any way unknown, it must be inferred and the degree to which it is unknown determines the complexity of the modelling required to infer it [48]. Here, effort becomes a latent variable. In the section above, we discuss **issues of latency** between the **observation process** and the **biological process**. Here, we have latency between different parts of the observation process. We may be able to use similar modelling techniques to tackle latent biology and latent effort. However, while we often have a good understanding of biological mechanisms with which to model latent **biological processes**, latent effort models require an understanding of human behaviour and (observer perceptions, group dynamics and economics, see Box 1).

In parallel with the methods to address **issues of latency**, there are three levels of complexity when trying to infer effort. The first is to use a **proxy** variable for effort based on an assumed **functional form**. For example, in amateur wildlife recording, researchers often use the frequency of a focal species or recorder's list length for a given site-visit as a measure of recording effort [48], however, this makes strong assumptions about how the focal species and biodiversity are distributed. A relaxation of this relationship is to assume a particular functional form linking effort to the variable. Distance sampling is perhaps the most obvious use of this technique, where effort (the detection function) decays with distance from the observer.

The most complex method for inferring effort relies on multiple **manifest variables** or known relationships. One approach is to use validation data collected with known effort. For the range of the validation data, the **biological process** is well characterised, meaning that where the validation and heterogeneous-effort data overlap, the differences can be attributed to effort, and thus used to train an effort model. Most effort models use covariates to predict effort but some use properties of how effort is distributed, such as self-excitement [52] or spatial autocorrelation to account for recording bias [53]. For example, effort in recording life history data is biased taxonomically but life histories are patterned phylogenetically allowing the **biological process** for taxa with low recording effort to borrow information from those with high recording effort [54; 55].

2.9 Scaling - Where the Model Finds Signal

2.9.1 Motivation

Determining the relevant **scale** for analysis is challenging and often overlooked [56; 57]. For statistical models to be ecologically relevant, the signal detected needs to have a biological interpretation. And yet, frequently, our models are designed to look for signal in raw data where the **scale** is determined by equipment precision and encoding, leading to a discrepancy between the **scale** of our inference and that of the **biological process**. Indeed, the **scale** that variables impact on the same **biological process** may vary, for example, the distance an organism travels to mate vs to feed. A single variable may impact at multiple **scales**, for example, phylogenetic distance may lead to trait autocorrelation at a large scale (organisms within an order are more similar than those in different orders) but negative correlation at a small scale (closely related species within a genus may be more different than more distantly related species in the genus). To reach the **scale** relevant to the **biological process**, therefore, our model needs to be able to change how neighbouring regions in the data are grouped together or divided. To do so, we need to understand what *proximity* means in variables like space, time and taxonomy, and how these units can be sensibly **aggregated** (or disaggregated).

Proximity needs to be defined in biologically sensible ways that may be non-linear and possibly directional. For example, geographic *proximity* might be defined in terms of landscape resistance to a particular organism [58], but also in terms of that organism's mobility (a distance traversed daily by a hare might be a life-time trajectory for a tortoise) [59]. Temporal *proximity* may be determined by latitude with rapid seasonal changes in temperature and weather towards the poles and more smooth transitions in the tropics. Similarly, taxonomic *proximity* can be defined by a combination of morphometrics, genomics and functional traits.

Aggregation operations often make an implicit mean-field assumption: that a system's behaviour is defined by the average value (e.g. of a covariate) across the system, so combining small units into larger units will lead to the same inference. However, **aggregation** of fine-**scale** processes into coarser **scale** observations can eliminate our ability to detect signals (see Box 3) [60]. For instance, a forager can be more efficient if all the prey in its home-range is concentrated at one known location and it may not matter if weather conditions are generally clement if a single day's storm can ruin a season's breeding chances [59]. In niche space, aggregating environments into coarse habitat classes might enclose under a single label distinct habitats that are recognized very differently by a species [61]. On the other hand, using very fine-**scale** data may lose the signal by obscuring the environmental context within which the important biology is unfolding (see Box 2). Different **biological process** may interact with the same covariates at different scales. For example, where a wolf moves in the next minute may be best predicted by habitat composition within 200m, whereas where a wolf establishes its home-range may be best predicted by habitat composition within 20km.

2.9.2 Existing Methods

Both too much and too little **aggregation** can lead to discrepancies between our data and the **biological process** making us vulnerable to over-and under-fitting issues [60]. Statistical diagnostics for these issues are common, but finding the appropriate **scale** is more challenging. One option is to fit models at multiple scales and compare using model-selection procedures [62]. A more sophisticated approach is to treat **scale** (or **scales**) as a parameter to be estimated [63] or to model **scales** hierarchically [64].

While conceptually simple, these approaches can be computationally prohibitive or limited by data availability. When aggregating at a particular **scale**, it is necessary to perform relatively costly numerical integration for each candidate scale. The cost of integration can be reduced using analytical tricks such as Fast Fourier Transformation algorithms. Another common method is to use a distance decay-kernel [65], such that distant observations bear lower importance. The **scale** parameter is then the decay coefficient [66; 67]. Estimating non-linear effects is becoming easier thanks to packages like INLAbru [68], which extends fast approximate Bayesian methods [69] in a user-friendly way to accommodate more complex models.

2.10 Concluding Remarks

Field scientists and statisticians face an ongoing challenge of how to tackle urgent complex questions with complex data sources. Eliciting the observation processes requires field science and statistical teams that work closely together and are motivated to understand one another. Where these teams do not exist, observation processes go unaccounted for, and any inference and policies made as a result are compromised. Where these teams succeed, they generate methodological advances, but advances which are often siloed due to field-specific language. Without breaking down these siloes, we stifle our progress. The typology we propose above is one route through this impasse. However, we believe that it already offers a fresh perspective on observation processes that can lead to methodological synthesis, innovation and insight as well as provide a mental roadmap through challenging terrain.

2.11 Outstanding Questions

- What data integration methods are missing? Many problems can be overcome by integrating complementary data types (e.g. combining fine scale data at a few locations with coarser data across a larger area to overcome issues of scale), however, the key is in identifying them.

- How do we incorporate observation process modelling into teaching? Complex observation processes are rarely emphasised in statistics courses but most students will need to tackle them. Would emphasising the observation process guard against defaulting to interpretation of patterns in data as biological signal?

- Can we link observation processes to experimental design techniques? How can simulating from models with observation processes improve data collection? Can we think of experimental design as a set of techniques to minimise **identifiability** issues while focusing **effort** on a small part of the biological process?
- Can the LIES framework be utilised beyond the fields of ecology and evolution?

2.12 Boxes

2.12.1 Box 1: Citizen Science Case Study

[70] identified four key challenges in analysing citizen science data caused by observer behaviour: spatial bias, observer differences, reporting preferences and false-positive errors. By linking these descriptions of the observation process to the LIES framework, we can better view them in their statistical context and find methodological commonalities between them, and across other fields of application.

Challenge	**Latency**	**Identifiability**	**Effort**	**Scale**
Spatial Bias	**X**	**X**	**X**	
Observer Differences		**X**	**X**	
Reporting Preference	**X**	**X**	**X**	
False +ve Errors	**X**	**X**		

Spatial Bias and Reporting Preferences

Using the LIES framework we found commonalities between Spatial Bias and Reporting Preferences. Both are issues of heterogeneous **effort** (across space and taxonomy, respectively). Citizen scientists are motivated to record by convenience (site accessibility and ease-of-identification) and ecological interest (site biodiversity and species interest, e.g., rarity status). Convenience can sometimes be predicted using covariates as **effort proxies**. While this is common for Spatial Bias, Reporting Preferences are less predictable. Targeting ecologically interesting sites and species leads to **practical identifiability** problems in distinguishing between observational and ecological parameters. A frequent solution is to integrate additional data that is either systematically collected or biased differently, and treat the **latent biological process** as a **hidden state** shared by the different datasets.

Observer Differences

Citizen scientists vary in skill so, even if they spend the same time in the field, their effective **effort** in terms of information gathered, will differ. Observer-level random effects and skill-scores can be used to estimate effective **effort** but these methods also need to account for skill improving with experience. A common solution is to use time as a **proxy** for **effort** changing within an individual.

These methods rely on labelling individual-level observers but, to ensure privacy, records are increasingly anonymised to prevent **mathematical identifiability** of individual observers. In these cases, a **latent variable** of **effort** can be used instead to estimate the combined effective **effort** of recorders across space and time. Analagous problems exist in the field of survey science leading to potential overlap between methods to address observer differences and, for example, participation and response biases.

False-Positive Error

Species misclassifications, where Species A is observed but recorded as Species B, are common in citizen science data and can lead to major **practical identifiability** problems when estimating species-habitat associations. If the two species co-exist in the habitat, then the degree of association may be overstated for Species B. If the record is a false-positive and two species do not co-exist at that location, the habitat association for Species B will be completely incorrect. Many methods have been developed for dealing with false-positive errors, but they often have **mathematical identifiability** issues due to equal likelihood support for the species being present-and-correctly-identified or absent-and-falsely-reported. Alternatively, we could frame the true species identity as a **latent** variable and infer the correct classification by, for example, linking with habitat data from studies with low error rates or, because species are not confused at random, using the citizen scientist's suggested label as a **proxy** variable.

2.12.2 Box 2: Animal Behaviour Case Study

Inferring animal behaviour from telemetry data is an exercise in extreme **latency**, using path geometry to deduce animal moods and motivations. Hidden Markov models (HMMs) have become a popular method for dealing with this problem. HMMs identify **hidden states** based on movement signatures (e.g., specific combinations of speed and tortuosity). The popularity of these models has led to several user-friendly implementations, allowing non-expert users to fit HMMs. However, HMMs hide several pitfalls, even for experienced statisticians. Below, we use the LIES framework to highlight problems that arise when using HMMs to infer behaviours using telemetry data that may not otherwise be apparent to such users.

HMMs suffer from **identifiability** problems in both their statistical machinery and interpretation. **Practical identifiability** of biologically meaningful **hidden states** is challenging. Determining the number of **hidden states** to estimate can be challenging even with prior knowledge of the number of expected behaviours, as the ideal number of statistical states may differ from the number of behaviours. With increasing numbers of and degrees of overlap between the **hidden states** comes considerable computational burden. Then, the statistical labelling of each state is not an intrinsic property and can change during model fitting (a **mathematical identifiability** problem known as label-switching). Even then, there may be no justifiable biological interpretation for the inferred **hidden states**, or ground-truthing (e.g. using video footage) may be required to find it.

The **scale** and homogeneity of recording **effort** can violate the key assumptions of HMMs that 1) switches between **hidden states** happen in discrete time and 2) the state at a given time is only dependent on the state at the previous time (the Markov property). For the **hidden states** identified by the HMM to make sense biologically, the temporal **scale** of observations needs to match that of behaviour switching in the animals. At finer **scales**, the behavioural states are likely to be strongly autocorrelated, while at coarser **scales** the switches may be missed entirely. Even at the right **scale**, **effort** heterogeneities can cause issues. While missing-at-random observations can be easily tolerated in the fitting process, extremely heterogeneous **effort** across time (i.e. all observations unevenly spaced) make the discrete-time assumption hard to justify and continuous-time models, though more challenging, may generate better results. Heterogeneous **effort** in space (e.g. signal problems), behaviour (e.g. recording fails during diving) or life-stage (e.g. only breeding individuals get tagged) will require more sophisticated imputation and interpretation.

2.12.3 Box 3: Biomonitoring By DNA Barcoding

Poorly known and highly species rich organism groups such as arthropods, fungi, bacteria, and protists are increasingly surveyed with DNA-metabarcoding. These methods can be applied either to bulk samples of study organisms (say, arthropods accumulated in a Malaise trap) or environmental samples (say, soil-, air- or water samples). With DNA-metabarcoding, species composition in a sample is a **latent** variable, whereas the DNA-barcode sequences are **manifest** variables. While the DNA barcodes have been specifically selected to be informative for species identification, they involve both **mathematical identifiability** problems (DNA barcodes can be identical for some closely related species) as well as **practical identifiability** problems (similar DNA-barcodes can be difficult to disentangle from noisy sequencing data). A more fundamental issue of **latency** is that most species of, for example, arthropods and fungi are still unknown to science, or known to science but missing from DNA barcode refence databases. In such cases, the sequences are often clustered into operational taxonomical units (OTUs) that can be viewed as **proxies** of taxonomically described species. Alternatively, the taxonomic **scale** is often selected for convenience reasons (rather than ecological reasons) from either the species **scale** or whatever **scale** allows for reliable taxonomical placement, such as the genus or order **scale**.

DNA-metabarcoding data are generally sample-based and thus the **effort** is typically highly standardized: a sample may represent, for example, arthropods that accumulated in a Malaise trap during a given time period, fungal spores sieved from a given volume of air, or microbial communities present in a given volume of soil or water. However, there are two layers of **latency** in converting the resulting sequence counts into species abundances. First, concerning the unit at which abundance is measured, it is to some extent possible to quantify the amount of DNA belonging to each focal species, e.g., by spiking the sample with controlled amount of synthetic DNA and by deriving species-specific DNA amplification factors. Unfortunately, one is seldom interested in species abundance

in units of ng of DNA (the **proxy variable**), but in units of counts of individuals, dry biomass, or so on (the **latent variables**). Second, concerning the spatial and temporal scales at which abundance is measured, one is seldom interested in species abundances in a specific sample, but in an ecologically more relevant unit such as a forest patch. How to convert the from the scale of a sample into ecologically relevant **spatial and temporal scale** is a major challenge that typically requires additional knowledge on, for example, the movement behavior of the organisms.

References

[1] M. E. Cristescu and P. D. Hebert, "Uses and misuses of environmental dna in biodiversity science and conservation," *Annual Review of Ecology, Evolution, and Systematics*, vol. 49, no. 1, pp. 209–230, 2018.

[2] J. Cavender-Bares, F. D. Schneider, M. J. Santos, A. Armstrong, A. Carnaval, K. M. Dahlin, L. Fatoyinbo, G. C. Hurtt, D. Schimel, P. A. Townsend, *et al.*, "Integrating remote sensing with ecology and evolution to advance biodiversity conservation," *Nature Ecology and Evolution*, pp. 1–14, 2022.

[3] A. M. Sequeira, M. O'Toole, T. R. Keates, L. H. McDonnell, C. D. Braun, X. Hoenner, F. R. Jaine, I. D. Jonsen, P. Newman, J. Pye, *et al.*, "A standardisation framework for bio-logging data to advance ecological research and conservation," *Methods in Ecology and Evolution*, vol. 12, no. 6, pp. 996–1007, 2021.

[4] E. D. Brown and B. K. Williams, "The potential for citizen science to produce reliable and useful information in ecology," *Conservation Biology*, vol. 33, no. 3, pp. 561–569, 2019.

[5] J. R. Platt, "Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others.," *Science*, vol. 146, no. 3642, pp. 347–353, 1964.

[6] J. S. Clark, "Why environmental scientists are becoming bayesians," *Ecology Letters*, vol. 8, no. 1, pp. 2–14, 2005.

[7] A. M. Ellison and B. Dennis, "Paths to statistical fluency for ecologists," *Frontiers in Ecology and the Environment*, vol. 8, no. 7, pp. 362–370, 2010.

[8] R. McElreath, *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC, 2020.

[9] A. Chao, C.-H. Chiu, and L. Jost, "Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through hill numbers," *Annual Review of Ecology, Evolution, and Systematics*, vol. 45, pp. 297–324, 2014.

[10] I. W. Renner and D. I. Warton, "Equivalence of maxent and poisson point process models for species distribution modeling in ecology," *Biometrics*, vol. 69, no. 1, pp. 274–281, 2013.

[11] I. W. Renner, J. Elith, A. Baddeley, W. Fithian, T. Hastie, S. J. Phillips, G. Popovic, and D. I. Warton, "Point process models for presence-only analysis," *Methods in Ecology and Evolution*, vol. 6, no. 4, pp. 366–379, 2015.

[12] G. Aarts, J. Fieberg, and J. Matthiopoulos, "Comparative interpretation of count, presence–absence and point methods for species distribution models," *Methods in Ecology and Evolution*, vol. 3, no. 1, pp. 177–187, 2012.

[13] B. T. McClintock, R. Langrock, O. Gimenez, E. Cam, D. L. Borchers, R. Glennie, and T. A. Patterson, "Uncovering ecological state dynamics with hidden markov models," *Ecology letters*, vol. 23, no. 12, pp. 1878–1903, 2020.

[14] M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, and B. Walker, "Catastrophic shifts in ecosystems," *Nature*, vol. 413, no. 6856, pp. 591–596, 2001.

[15] E. A. Bender, T. J. Case, and M. E. Gilpin, "Perturbation experiments in community ecology: theory and practice," *Ecology*, vol. 65, no. 1, pp. 1–13, 1984.

[16] T. Royama, *Analytical population dynamics*, vol. 10. Springer Science and Business Media, 2012.

[17] W. W. Murdoch, C. J. Briggs, and R. M. Nisbet, "Consumer-resource dynamics (mpb-36)," in *Consumer-Resource Dynamics (MPB-36)*, Princeton University Press, 2013.

[18] P. A. Abrams and L. R. Ginzburg, "The nature of predation: prey dependent, ratio dependent or neither?," *Trends in Ecology and Evolution*, vol. 15, no. 8, pp. 337–341, 2000.

[19] O. Ovaskainen, G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego, "How to make more out of community data? a conceptual framework and its implementation as models and software," *Ecology Letters*, vol. 20, no. 5, pp. 561–576, 2017.

[20] J. Niku, F. K. Hui, S. Taskinen, and D. I. Warton, "gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r," *Methods in Ecology and Evolution*, vol. 10, no. 12, pp. 2173–2182, 2019.

[21] M. Auger-Méthé, K. Newman, D. Cole, F. Empacher, R. Gryba, A. A. King, V. Leos-Barajas, J. Mills Flemming, A. Nielsen, G. Petris, *et al.*, "A guide to state–space modeling of ecological time series," *Ecological Monographs*, vol. 91, no. 4, p. e01470, 2021.

[22] R. Langrock, R. King, J. Matthiopoulos, L. Thomas, D. Fortin, and J. M. Morales, "Flexible and practical modeling of animal telemetry data: hidden markov models and extensions," *Ecology*, vol. 93, no. 11, pp. 2336–2342, 2012.

[23] P. Büscher, J.-M. Bart, M. Boelaert, B. Bucheton, G. Cecchi, N. Chitnis, D. Courtin, L. M. Figueiredo, J.-R. Franco, P. Grébaut, *et al.*, "Do cryptic reservoirs threaten gambiense-sleeping sickness elimination?," *Trends in Parasitology*, vol. 34, no. 3, pp. 197–207, 2018.

[24] A. A. Beaujean, *Latent variable modeling using R: A step-by-step guide*. Routledge, 2014.

[25] W. H. Finch and B. F. French, *Latent variable modeling with R*. Routledge, 2015.

[26] J. Loehlin and A. Beaujean, *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis, Fifth Edition (5th ed.)*. Routledge, 2017.

[27] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White, "Generalized linear mixed models: a practical guide for ecology and evolution," *Trends in ecology & evolution*, vol. 24, no. 3, pp. 127–135, 2009.

[28] V. Hemming, A. M. Hanea, T. Walshe, and M. A. Burgman, "Weighting and aggregating expert ecological judgments," *Ecological Applications*, vol. 30, no. 4, p. e02075, 2020.

[29] K. Ogle and J. J. Barber, "Ensuring identifiability in hierarchical mixed effects bayesian models," *Ecological Applications*, vol. 30, no. 7, p. e02159, 2020.

[30] C. F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, *et al.*, "Collinearity: a review of methods to deal with it and a simulation study evaluating their performance," *Ecography*, vol. 36, no. 1, pp. 27–46, 2013.

[31] F.-G. Wieland, A. L. Hauber, M. Rosenblatt, C. Tönsing, and J. Timmer, "On structural and practical identifiability," *Current Opinion in Systems Biology*, vol. 25, pp. 60–69, 2021.

[32] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[33] M. Modrák, A. H. Moon, S. Kim, P. Bürkner, N. Huurre, K. Faltejsková, A. Gelman, and A. Vehtari, "Simulation-based calibration checking for bayesian computation: The choice of test quantities shapes sensitivity," *arXiv preprint arXiv:2211.02383*, 2022.

[34] C. Xu, Y. Hu, Y. Chang, Y. Jiang, X. Li, R. Bu, and H. He, "Sensitivity analysis in ecological modeling," *Ying yong sheng tai xue bao= The Journal of Applied Ecology*, vol. 15, no. 6, pp. 1056–1062, 2004.

[35] C. Soize, *Uncertainty quantification*. Springer, 2017.

[36] T. J. Sullivan, *Introduction to uncertainty quantification*, vol. 63. Springer, 2015.

[37] J. R. Reimer, F. R. Adler, K. M. Golden, and A. Narayan, "Uncertainty quantification for ecological models with random parameters," *Ecology Letters*, vol. 25, no. 10, pp. 2232–2244, 2022.

[38] Z. M. Laubach, E. J. Murray, K. L. Hoke, R. J. Safran, and W. Perng, "A biologist's guide to model selection and causal inference," *Proceedings of the Royal Society B*, vol. 288, no. 1943, p. 20202815, 2021.

[39] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák, "Bayesian workflow," *arXiv preprint arXiv:2011.01808*, 2020.

[40] A. Gelman, J. B. Carlin, H. S. Stern, D. Duncan, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. Chapman and Hall/CRC, 2014.

[41] P. C. Johnson, S. J. Barry, H. M. Ferguson, and P. Müller, "Power analysis for generalized linear mixed models in ecology and evolution," *Methods in Ecology and Evolution*, vol. 6, no. 2, pp. 133–142, 2015.

[42] J. K. Kruschke and T. M. Liddell, "The bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective," *Psychonomic Bulletin and Review*, vol. 25, no. 1, pp. 178–206, 2018.

[43] R. B. Gramacy, *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020.

[44] M. J. Brewer, A. Butler, and S. L. Cooksley, "The relative performance of aic, aicc and bic in the presence of unobserved heterogeneity," *Methods in Ecology and Evolution*, vol. 7, no. 6, pp. 679–692, 2016.

[45] D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye, "Penalising model component complexity: A principled, practical approach to constructing priors," *Statistical Science*, vol. 32, no. 1, pp. 1–28, 2017.

[46] A. T. Tredennick, G. Hooker, S. P. Ellner, and P. B. Adler, "A practical guide to selecting models for exploration, inference, and prediction in ecology," *Ecology*, vol. 102, no. 6, p. e03336, 2021.

[47] A. Vehtari, A. Gelman, and J. Gabry, "Practical bayesian model evaluation using leave-one-out cross-validation and waic," *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[48] N. J. Isaac, A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy, "Statistics for citizen science: extracting signals of change from noisy ecological data," *Methods in Ecology and Evolution*, vol. 5, no. 10, pp. 1052–1060, 2014.

[49] T. Mendo, J. Mendo, J. Ransijn, I. Gomez, P. Gil-Kodaka, J. Fernández, R. Delgado, A. Travezaño, R. Arroyo, K. Loza, *et al.*, "Assessing discards in an illegal small-scale fishery using fisher-led reporting," *Reviews in Fish Biology and Fisheries*, pp. 1–12, 2022.

[50] L. Nelli, J. Langbein, P. Watson, and R. Putman, "Mapping risk: Quantifying and predicting the risk of deer-vehicle collisions on major roads in england," *Mammalian Biology*, vol. 91, no. 1, pp. 71–78, 2018.

[51] J. Matthiopoulos, E. Wakefield, J. W. Jeglinski, R. W. Furness, M. Trinder, G. Tyler, A. Mccluskie, S. Allen, J. Braithwaite, and T. Evans, "Integrated modelling of seabird-habitat associations from multi-platform data: A review," *Journal of Applied Ecology*, 2022.

[52] P. Clayton, D. Murthy, and F. Lago, "A framework for harnessing citizen scientists and journalist networks for post-disaster reconnaissance," in *2019 Natural Hazards Workshop—Researchers Meeting*, 2019.

[53] E. Browning, R. Freeman, K. L. Boughey, N. J. Isaac, and K. E. Jones, "Accounting for spatial autocorrelation and environment are important to derive robust bat population trends from citizen science data," *Ecological Indicators*, vol. 136, p. 108719, 2022.

[54] H. K. Kindsvater, N. K. Dulvy, C. Horswill, M.-J. Juan-Jordá, M. Mangel, and J. Matthiopoulos, "Overcoming the data crisis in biodiversity conservation," *Trends in Ecology and Evolution*, vol. 33, no. 9, pp. 676–688, 2018.

[55] T. F. Johnson, N. J. Isaac, A. Paviolo, and M. González-Suárez, "Handling missing values in trait data," *Global Ecology and Biogeography*, vol. 30, no. 1, pp. 51–62, 2021.

[56] S. A. Levin, "The problem of pattern and scale in ecology: the robert h. macarthur award lecture," *Ecology*, vol. 73, no. 6, pp. 1943–1967, 1992.

[57] J. Chave, "The problem of pattern and scale in ecology: what have we learned in 20 years?," *Ecology letters*, vol. 16, pp. 4–16, 2013.

[58] P. E. Howell, E. Muths, B. R. Hossack, B. H. Sigafus, and R. B. Chandler, "Increasing connectivity between metapopulation ecology and landscape ecology," *Ecology*, vol. 99, no. 5, pp. 1119–1128, 2018.

[59] J. Matthiopoulos, J. Fieberg, G. Aarts, F. Barraquand, and B. E. Kendall, "Within reach? habitat availability as a function of individual mobility and spatial structuring," *The American Naturalist*, vol. 195, no. 6, pp. 1009–1026, 2020.

[60] R. S. Paton and J. Matthiopoulos, "Defining the scale of habitat availability for models of habitat selection," *Ecology*, vol. 97, no. 5, pp. 1113–1122, 2016.

[61] J. Matthiopoulos, "How animals endure by bending environmental space: Redefining the fundamental niche," *bioRxiv*, 2021.

[62] R. Mancy, M. Rajeev, A. Lugelo, K. Brunker, S. Cleaveland, E. A. Ferguson, K. Hotopp, R. Kazwala, M. Magoto, K. Rysava, *et al.*, "Rabies shows how scale of transmission can enable acute infections to persist at low prevalence," *Science*, vol. 376, no. 6592, pp. 512–516, 2022.

[63] Y. Haddou, R. Mancy, J. Matthiopoulos, S. Spatharis, and D. M. Dominoni, "Widespread extinction debts and colonization credits in united states breeding bird communities," *Nature Ecology and Evolution*, vol. 6, no. 3, pp. 324–331, 2022.

[64] N. Abrego, C. Bässler, M. Christensen, and J. Heilmann-Clausen, "Traits and phylogenies modulate the environmental responses of wood-inhabiting fungal communities across spatial scales," *Journal of Ecology*, vol. 110, no. 4, pp. 784–798, 2022.

[65] B. Aue, K. Ekschmitt, S. Hotes, and V. Wolters, "Distance weighting avoids erroneous scale effects in species-habitat models," *Methods in Ecology and Evolution*, vol. 3, no. 1, pp. 102–111, 2012.

[66] R. Chandler and J. Hepinstall-Cymerman, "Estimating the spatial scales of landscape effects on abundance," *Landscape Ecology*, vol. 31, no. 6, pp. 1383–1394, 2016.

[67] F. Carpentier and O. Martin, "Siland a r package for estimating the spatial influence of landscape," *Scientific Reports*, vol. 11, no. 1, pp. 1–6, 2021.

[68] F. E. Bachl, F. Lindgren, D. L. Borchers, and J. B. Illian, "inlabru: an r package for bayesian spatial modelling from ecological survey data," *Methods in Ecology and Evolution*, vol. 10, no. 6, pp. 760–766, 2019.

[69] F. Lindgren and H. Rue, "Bayesian spatial modelling with r-inla," *Journal of Statistical Software*, vol. 63, pp. 1–25, 2015.

[70] A. Johnston, E. Matechou, and E. B. Dennis, "Outstanding challenges and future directions for biodiversity monitoring using citizen science data," *Methods in Ecology and Evolution*, 2022.

Chapter 3

COVID-19 – exploring the implications of long-term condition type and extent of multimorbidity on years of life lost: a modelling study

This chapter has been published in: Hanlon P, Chadwick F, Shah A, Wood R, Minton J, McCartney G, Fischbacher C, Mair FS, Husmeier D, Matthiopoulos J, McAllister DA. COVID-19 - exploring the implications of long-term condition type and extent of multimorbidity on years of life lost: a modelling study. Wellcome Open Res. 2021 Mar 1;5:75. doi: 10.12688/wellcomeopenres.15849.3. PMID: 33709037; PMCID: PMC7927210.

My role in this paper was to lead the development of the "Long-term condition prevalence and correlation models". Understanding the correlations between long-term conditions was crucial in addressing the central aims of this project and required entirely novel methodology. The outputs of these models were then analysed using established techniques in the health economics literature. A mathematical description of this model is available in Appendix A.

The L.I.E.S. Framework

The observation processes in the following case study include:

Latency: the diseases in this case study are observed as aggregations (counts) of binary (present/absent) but exist and are correlated in continuous states.

Identifiability: (*mathematical*) the count data are sums of binarised (present/absent) diseases, however there are many possible patient-level combinations which are consistent with the count data. The disease correlations are modelled using a multivariate probit model in which the mean and variance are not identifiable.

Identifiability: (*practical*) information will be lost with each degree of latency, there will likely be large uncertainty in the resulting inference.

The application of the framework to the case study will be discussed in more detail in the conclusion.

3.1 Abstract

Background: COVID-19 is responsible for increasing deaths globally. As most people dying with COVID-19 are older with underlying long-term conditions (LTCs), some speculate that YLL are low. We aim to estimate YLL attributable to COVID-19, before and after adjustment for number/type of LTCs, using the limited data available early in the pandemic.

Methods: We first estimated YLL from COVID-19 using WHO life tables, based on published age/sex data from COVID-19 deaths in Italy. We then used aggregate data on number/type of LTCs in a Bayesian model to estimate likely combinations of LTCs among people dying with COVID-19. We used routine UK healthcare data from Scotland and Wales to estimate life expectancy based on age/sex/these combinations of LTCs using Gompertz models from which we then estimate YLL.

Results: Using the standard WHO life tables, YLL per COVID-19 death was 14 for men and 12 for women. After adjustment for number and type of LTCs, the mean YLL was slightly lower, but remained high (11.6 and 9.4 years for men and women, respectively). The number and type of LTCs led to wide variability in the estimated YLL at a given age (e.g. at $\geq$80 years, YLL was $>$10 years for people with 0 LTCs, and $<$3 years for people with $\geq$6).

Conclusions: Deaths from COVID-19 represent a substantial burden in terms of per-person YLL, more than a decade, even after adjusting for the typical number and type of LTCs found in people dying of COVID-19. The extent of multimorbidity heavily influences the estimated YLL at a given age. More comprehensive and standardised collection of data (including LTC type, severity, and potential confounders such as socioeconomic-deprivation and care-home status) is needed to optimise YLL estimates for specific populations, and to understand the global burden of COVID-19, and guide policy-making and interventions.

3.2 Introduction

The SARS-CoV-2 pandemic, the virus causing COVID-19, emerged in late 2019 and continues to have substantial impact on populations and healthcare systems throughout the world. This manuscript presents a revised version of an analysis initially conducted in March 2020, at which time Italy, the first European nation to experience a major outbreak of COVID-19, was seeing rapidly escalating numbers of cases and deaths. In the UK, at that time, the initially small number of hospitalisations and deaths were beginning to rise. The analysis sought to estimate the burden of COVID-19 deaths in terms of potential years of life lost (YLL), at a time when individual-level data on COVID-19 deaths was scarce.

When severe, coronavirus disease 2019 (COVID-19) causes acute respiratory failure, often requiring mechanical ventilation [1]. At the beginning of April 2020, more than 1,200,000 confirmed cases have been reported globally, including 67,000 deaths [2]. In response to this threat, governments introduced non - pharmaceutical interventions such as physical distancing and the delivery of health services has radically changed, with resources diverted towards the management of COVID-19 and away from their usual activities [3]. These measures

have aimed to limit a surge in cases that risks overwhelming healthcare services [4], and have continued and repeated in various forms throughout the world.

Since few health care systems could have responded adequately to the increased need for acute care without these changes, these decisions were in some ways inevitable. However, as societies seek to "return to normal", decisions about the extent and nature of ongoing measures to limit spread of COVID-19 will be more difficult. These choices will require balancing the likely direct effects on mortality from COVID-19 against the likely indirect impacts on mortality for other conditions – due, for example, to inadequate access to necessary services for many people with long-term conditions (LTCs), potential reluctance of the public to attend for acute events such as myocardial infarction, or impacts from forced unemployment, loss of income and social isolation. The indirect effects are likely to be complex, most will be downstream, and will require extensive research to be better understood. However, we need to capture the direct effects of COVID-19 as accurately as possible now, via currently available data and methodologies.

In April 2020, most reports of COVID-19 deaths used raw counts [2]. This may give a distorting picture of the mortality burden, however, as it does not consider how long someone who died from COVID-19 might otherwise have been expected to live. As people dying from COVID-19 are predominantly older and have pre-existing LTCs [5; 6; 7], some have speculated that many of these people would have soon died of other causes and that life expectancy may therefore not being greatly impacted [8; 9]. While multimorbidity, the presence of multiple LTCs, is known to be associated with increased mortality [10], people with multimorbidity nonetheless can be expected to live for many years [11]. Raw counts of deaths may therefore mislead policy-makers and the public, causing them to either over- or under-estimate the total impact of COVID-19 related deaths.

Within epidemiology, there is a standard measure used to account for this difficulty, the years of potential life lost (YLL) [12]. YLL can be expressed per-capita as the average number of years an individual would have been expected to live had they not died of a given cause. The conventional approach to YLL uses data on the age at which deaths occurred combined with typical life expectancy at a given age, to estimate a weighted average of the number of years lost. YLL is used to allow fair comparisons of the health impact of different policies, such as different measures to address the pandemic. However, given the controversial role of multimorbidity in COVID-19 deaths it is also important to calculate YLL additionally considering the effects of the presence of a single LTC or multimorbidity.

Therefore, we propose to quantify the burden of mortality related to COVID-19, both using the conventional age-based YLL measure, and YLL additionally accounting for type and number of underlying LTCs. We draw upon data sources available in April 2020, as this modelling study aimed to estimate the potential YLL at an early stage in the pandemic, when the impact was emerging. It should be noted, however, that events unfolding throughout the pandemic are likely to impact the YLL. Any estimate, particularly in the context of a pandemic, is dependent on what populations are exposed, and to what extent. Updated estimates, taking account of events which transpired in the UK and beyond, are the subject of ongoing collaborative efforts and we have not attempted to model these. Rather, this manuscript provides a detailed and reproducible quantification of

YLL using techniques targeting the specific challenges of estimation at the early stages of the pandemic.

3.3 Methods

3.3.1 WHO standard YLL approach

The standard approach for calculating years of life lost is to apply the distribution of ages among those who died from a specific cause to a standard life-table. For the purposes of international comparison, we opted to use the WHO 2010 Global Burden of Diseases table as the reference [13], which presents YLL by age, but not by sex or extent of multimorbidity. This method involves summing the expected years of life remaining from the table according to the number (or for the mean YLL the proportion) of people dying within each age-band. We applied the age distribution of COVID-19 deaths in Italy from published data to estimate the YLL [14].

We chose the WHO life tables to allow comparison of the burden of COVID-19 deaths with other conditions in an international context. However these, unlike many national-level life tables, do not stratify by sex. Furthermore our subsequent modelling draws upon data from specific setting based on availability early in the pandemic (namely data on COVID-19 deaths from Italy, and life-expectancy estimates based on data from Wales). Therefore, following comments from academic colleagues via social media, we performed sensitivity analyses using life tables from Italy (2017), United Kingdom (2016–2018) and, for comparison, the United States (2017).

3.3.2 Overview of modelling to accommodate long-term conditions and multimorbidity

The remainder of the methods describes our approach to estimating YLL accounting for number and type of underlying LTC, along with age and sex. Our modelling comprised three main components: (i) estimating the prevalence of, and correlations between, LTCs among people dying with COVID-19; (ii) modelling UK life expectancy based on age, sex, and each combination of these LTCs separately; and (iii) combining these models to calculate the estimated YLL per death with COVID-19. These are summarised by age-group, sex, and multimorbidity counts (that take into account different combinations of LTCs).

The data sources used for each of these stages of modelling are summarised in Figure 3.1.

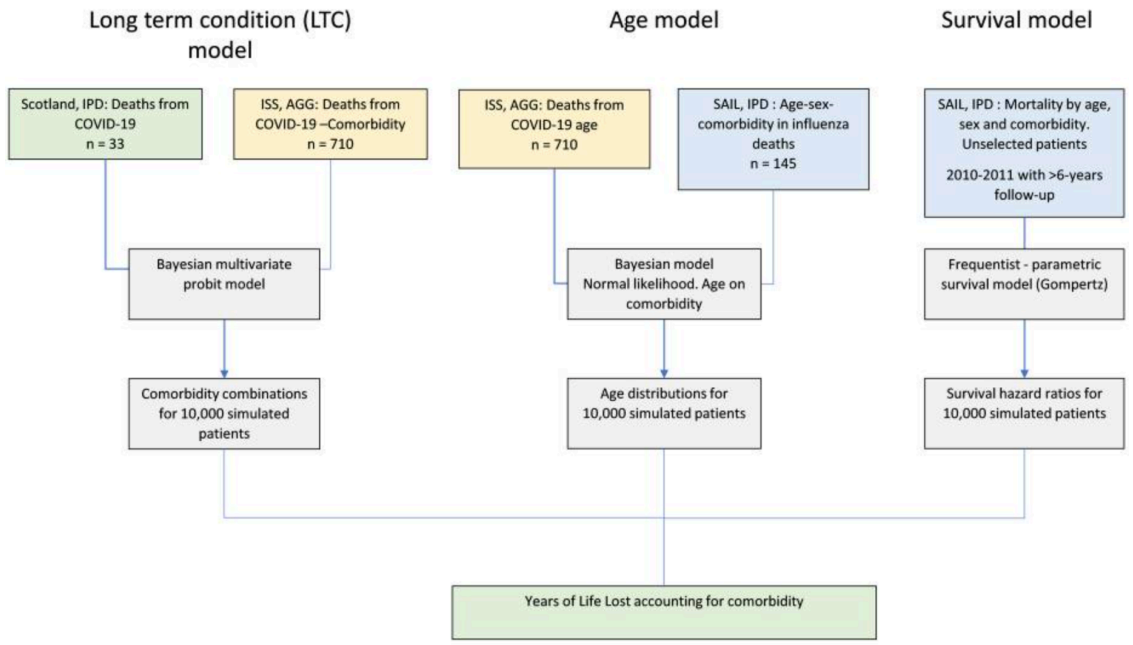

Figure 3.1: **Overview of Components of Models.** Green boxes indicate source of data or final outputs. Yellow boxes indicate Istituto Superiore di Sanità (ISS) data and blue boxes indicate Secure Anonymised Record Linkage (SAIL) data. White boxes indicate each model used to inform the final analysis. AGG - aggregate. IPD - individual level patient data.

3.3.3 Rapid review

To inform our estimates of number and type of LTCs, we first sought to identify the most detailed data available for underlying long-term conditions among people dying of COVID-19. We performed a rapid review to identify data on underlying conditions for people dying with COVID-19. We searched the WHO repository of COVID-19 studies on 24th March 2020. To identify studies reporting data on LTCs among people who had died from Covid-19, we screened titles and abstracts of all epidemiological, clinical, case-series and review articles (n=1685). We identified and screened 77 potentially relevant full-text articles, of which four reported aggregate data on LTCs among people who had died of COVID-19. Three were small studies (32, 44, and 54 deaths, respectively) based in Wuhan, China [5; 6; 7]. However, the fourth was a comprehensive report from the Istituto Superiore di Sanità (ISS) (published each Tuesday and Wednesday) including data on 11 common LTCs (ischaemic heart disease, atrial fibrillation, heart failure, stroke, hypertension, diabetes, dementia, chronic obstructive pulmonary disease, active cancer in the past 5 years, chronic liver disease and chronic renal failure), as well as the number of patients who had 0, 1, 2 or $\geq$3 LTCs for 701 of the 6801 people who died with COVID-19 in Italy [14]. In view of the smaller sizes of the Chinese studies, and the greater dissimilarity of these populations

with the UK relative to the Italian data, we opted not to include these in the analysis. These data were used to construct a plausible scenario for the prevalence of combinations of LTCs among people who died from COVID-19 for the modelling presented here.

Long-term condition prevalence and correlation models. This first stage of our modelling aimed to estimate the prevalence and correlation between specific LTCs among people dying with COVID-19.

We utilised aggregate data on COVID-19 deaths from the Istituto Superiore di Sanità in Italy. Since we were unable to obtain individual patient data for the Italian case-series of deaths from COVID-19, we had to infer the joint prevalence of LTCs from the summarised information available, i.e. the marginal distribution of multimorbidity counts (the row sums, or total number of diseases for each patient, wherein counts of $\geq$3 LTCs were collapsed into the single category of 3+) and the marginal distributions of LTC frequency (the columns sums, or the total number of patients with each LTC). To that end, we developed a Bayesian latent process model of disease prevalence and correlation and fitted it using Markov chain Monte Carlo (MCMC) to both elements in the published data. This analysis was applied jointly to the small number of deaths that had occurred in Scotland, primarily to aid convergence in Bayesian model fitting by providing some information about the correlation between LTCs [15]. The Scottish subset of the data contained a partial record of known LTCs for individual patients, but the multimorbidity count per patient, as well as the marginal frequency of each LTC, were missing (hence, modelled as latent). Bayesian priors for the correlations between diseases were specified with a tendency to zero (shrinkage). Numerical investigations indicated little sensitivity of convergence to the strength of shrinkage, so we opted for weak shrinkage as a precautionary approach. This model gave us the full matrix of correlations between every combination of LTCs at the level of individuals, therefore providing us with a complete dependence structure of LTCs presented within the sample of COVID-19 mortalities. In order to propagate uncertainty through the analysis, from this fitted model (effective sample size of MCMC 410) we simulated 10,000 notionally "typical" patients, with plausible combinations of LTCs (under the combined Italian and Scottish data).

To test the sensitivity of our findings to the estimated correlations, we also estimated the YLL under two opposite extremes (i) that LTCs were independent and (ii) that LTCs were highly correlated. Unlike the Bayesian LTC mode, these sensitivity analyses did not use the information on the multimorbidity counts from the ISS report, but only the proportion of patients with each of the eleven comorbidities. For the "independent" scenario we created 11 vectors comprising 1s and 0s (respectively with and without the long term condition) corresponding in length to the number of patients. We then sampled from these vectors with replacement to obtain 10,000 simulated patients. For the "highly correlated" scenario we first sorted each vector, then combined them to form a 710x11 matrix, then sampled each row with replacement to obtain 10,000 simulated patients. This generated a dataset where individuals with one comorbidity which reduces life expectancy were more likely to have other comorbidities which reduce life expectancy (and vice versa).

Age models. Next, we modelled the relationship between age and multimorbidity counts among people dying with COVID-19. We were unable to ob-

tain direct estimates of the association between age and extent of multimorbidity among patients who had died from COVID-19. Therefore, we modelled two scenarios: independence between age and multimorbidity count (i.e. no correlation between age and multimorbidity count among people dying of COVID-19), and a positive association between age and multimorbidity count. To inform the latter, we examined data within the Secure Anonymised Information Linkage (SAIL) databank for 145 patients who had influenza recorded as the cause of death in their death certificate in 2011. SAIL is a repository of routinely collected healthcare data (including primary care, hospital episodes, and mortality data) from a representative sample covering approximately 70% of the population of Wales. While influenza is a different condition, these data were used for the sole purpose of estimating correlations between age and multimorbidity counts (conditioning on death), and did not inform the model in any other way. We found that for men, age increased by 4.7 years per unit increase in the number of LTCs until the count reached 6 after which there was no evidence of further increase. For women, the figure was 2.6. Therefore, we performed the modelling assuming that for COVID-19 the mean age increased by 5 years per unit increase in multimorbidity count across the range from 0 to 6 LTCs in men. To allow for some degree of uncertainty around this estimate by sampling from a normal distribution. We arbitrarily chose a standard deviation of 0.5. We estimated this similarly for women, but using a mean increase of age of 3 years per increase in multimorbidity count. We incorporated this information in a model fitted to the summary age data provided in the Italian case report. We obtained 10,000 samples from the posterior distribution for inclusion in the YLL calculations. SAIL analyses were approved by SAIL Information Governance Review Panel (Project 0830). Approval for the use of individual patient data in the analysis was given by the NHS Public Health Scotland Caldicott officer.

Survival models. For patients aged 50 years or older at death, we estimated mortality according to age, sex and combinations of each LTC using the SAIL. From these data, we identified all participants aged over 49 years who were registered with a participating practice for the duration of 2011 (approximately 0.85 million people). This period was selected as electronic coding of diagnoses was well established, and it allowed >6 years of follow-up. Age and sex were extracted from primary care records. We also identified all LTCs for which we had information of COVID-19 deaths from Italy. LTCs were identified using a combination of primary care data (using Read diagnostic codes) and hospital episodes (using ICD-10 codes). Individuals were considered to have a LTC if they had a relevant diagnostic code entered prior to 31st December 2011. Relevant codes were identified from the Charlson comorbidity index and the Elixhauser comorbidity index [16; 17], which had established algorithms for identification from ICD-10 codes [18], and have been adapted for using Read codes in primary care [19]. Code lists are available in [15].

All-cause mortality was assessed by linkage to national mortality registers from 1st January 2012 until August 2018 (last available data). Participants were censored if they de-registered from a participating SAIL practice. We used the flexsurv package in R (version 1.1.1) to fit a Gompertz model treating age as the timescale [20]. We assessed the fit of this distribution graphically [15]. In models stratified by sex we included all the LTCs as main effects as well as age-LTC inter-

actions that improved the model fit in terms of the Akaike information criterion. In sensitivity analyses we also included two-way (comorbidity-comorbidity) and three-way (comorbidity-comorbidity-age) interaction terms for the four comorbidities with the largest effect measure estimates (COPD, heart failure, liver failure and dementia) requiring 12 additional parameters. To propagate uncertainty from the survival models we obtained 10,000 samples of the coefficient estimates by sampling from a multivariate normal distribution corresponding to the coefficients and variance-covariance matrix from the regression models.

Combination of comorbidity and mortality models. In the final analysis, we combined 10,000 samples from all three sources: LTC combination models, age models and survival models. We used the rate and shape parameters with the cumulative distribution function implemented in the flexsurv package to calculate the survival probabilities at 3-month intervals from aged 50 to 120 (to allow all curves to descend to zero). From these times and survival probabilities we estimated the mean survival, or life expectancy.

Bayesian models were written in the JAGS language [21] and implemented using runjags for R (version 2.0.4) [22], survival models were fit using the flexsurv package in R (version 1.1.1) 20, and for the final analysis the model-outputs were also combined in R (version 3.6.1). The 95% uncertainty intervals were obtained using empirical bootstrapping, with the number of samples in the mean equal to the effective sample size from the LTC correlation model. All code, data (except individual-level data for Scotland), intermediate outputs and diagnostic plots are provided on GitHub (https://github.com/dmcalli2/covid19_yll_final) [15].

3.4 Results

3.4.1 WHO life tables

The proportion of men and women in 10-year age-bands was reported for the 6801 deaths included in the ISS case report. On applying the proportion in each age-band to the WHO Global Burden of Disease 2010 life tables for men, we found that the YLL was 14.4 per person using the whole cohort and 14 after excluding those aged under 50. For women, comparable figures were 12.2 and 11.8 years, respectively. In sensitivity analyses using alternative life tables, life expectancy was lower (particularly for men), however the estimates YLL remained above 10 years for both men and women, regardless of life table used (detailed results shown in https://github.com/dmcalli2/covid19_yll_final/blob/master/Scripts/Addendum.md).

3.4.2 Comorbidity models

For 710 patients who had died with COVID-19 for whom information on LTCs was presented in the ISS report [14], the proportion with each LTC was as follows: ischaemic heart disease 27.8%, atrial fibrillation 23.7%, heart failure 17.1%, stroke 11.3%, hypertension 73 diabetes 31.3%, dementia 14.5%, chronic obstructive pulmonary disease 16.7%, active cancer in the past 5 years 17.3%, chronic liver disease 4.1%, chronic renal failure 22.2%. The ISS report also presented the proportion of patients who died with each of the following multimorbidity

counts: 0 (2.1%), 1 (21.3%), 2 (25.9%) and ≥3 (50.7%). Using these data, alongside individual - level patient data for a small number of patients from Scotland to aid with model fitting, we were able to simulate a set of realistic notional patients with specific combination of LTCs. The correlations between every pair of LTCs are shown in the appendix and the full posterior distributions from the modelling are available at GitHub (https://github.com/dmcalli2/covid19_yll_final) [15].

3.4.3 Age models

Based on the proportions reported for each age-band, for men the mean age for the ISS deaths was 77.9 years when people aged less than 50 were excluded and 77.4 years overall. For women the figure was 81.1 for both. The models we fit to these data to smooth out the distribution and to make it easier to accommodate different scenarios for the association between age and multimorbidity counts comorbidity are shown in Figure 3.2; the distribution of age and multimorbidity counts for men and women are shown under the assumption that these are independent, and under the assumption that multimorbidity is associated with age.

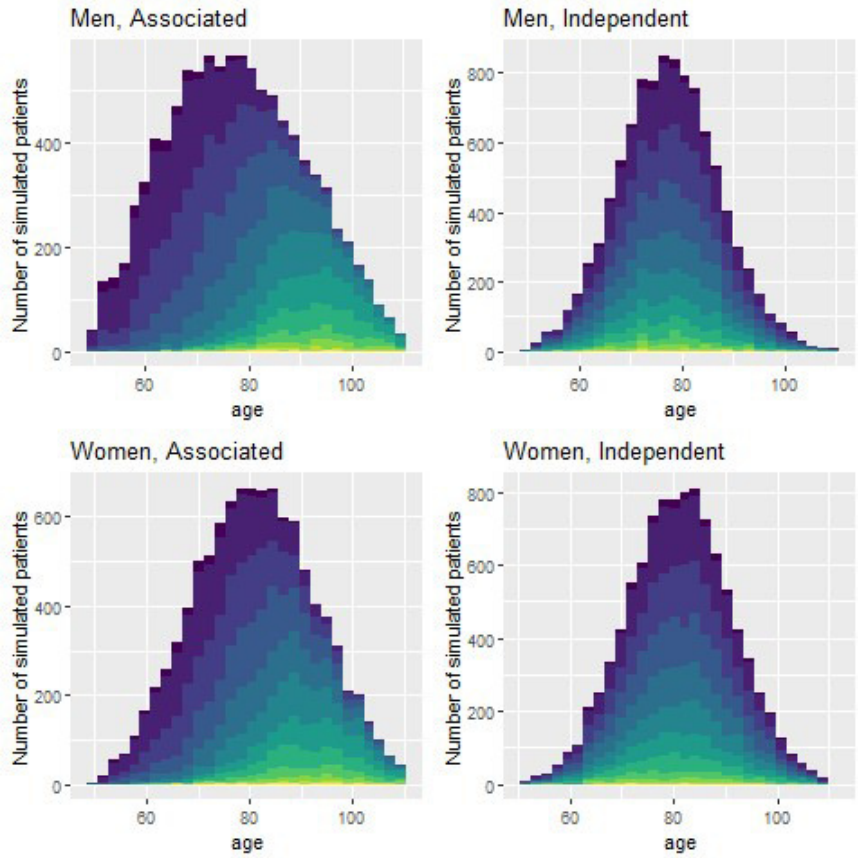

Figure 3.2: **Modelled distribution of age in ISS population, assuming age is associated with comorbidity counts, and assuming age and comorbidity are independent.** Coloured bars indicate the comorbidity count from zero (dark/blue) to 11 (light/yellow).

3.4.4 Survival models

The coefficients for the survival models are shown in the supplementary appendix. Briefly, all LTCs other than hypertension were associated with increased mortality (in a model including 10 other LTCs), and for each LTC the association with mortality was attenuated as the baseline age increased. Figure 3.3 shows the survival curves applied to different age and combinations of LTCs, stratified by age-band and multimorbidity count. This figure shows how these associations and age relate to survival across the age range from 50 to 110 years old.

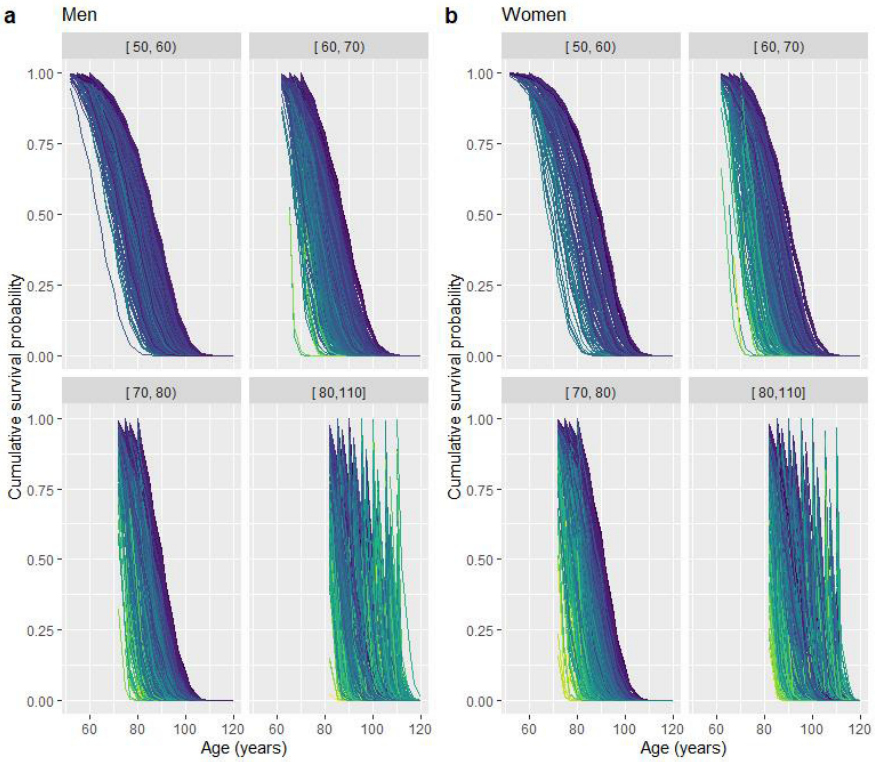

Figure 3.3: **Survival curves for all-cause mortality** Figures are paneled by age and sex. Individual lines represent survival curves for a single simulated patients with a given set of LTCs. From light to dark (yellow to blue) they show decreasing multimorbidity counts (11 to 0). There are 10, 000 lines, one for each notional patient. Lines run from the age at which each simulated patient died (survival probability = 1) to when they would have died under the model (survival probability = 0). Patients with the same age and total multimorbidity count will have a different survival curve if they have a different set of 11 LTCs.

3.4.5 Years of life lost

For men the average YLL on adjusting for number and type of LTC as well as age was 11.6(10.9–12.4). For women this value was 9.4(8.7–10). The results were similar under the different assumptions for the age-multimorbidity association and in both sensitivity analyses, whether assuming strongly correlated or independent LTCs (Table 3.1). For comparison, the YLL based on age alone using the WHO tables was 14.0 and 11.8 for men and women, respectively.

Table 3.1: Years of life lost (YLL) and 95% credible intervals under different modelling assumptions.

LTC-LTC correlation	Age-multimorbidity correlation	Men	Women
Modelled	Associated	11.6 (10.9-12.4)	9.4 (8.7-10)
Modelled	Independent	11.1 (10.4-11.7)	9.2 (8.6-9.8)
Independent	Associated	12 (11.2-12.9)	9.8 (9.2-10.5)
Independent	Independent	11.5 (10.9-12.1)	9.6 (9.1-10.2)
Highly correlated	Associated	13.5 (12.5-14.4)	10.9 (10.1-11.8)
Highly correlated	Independent	12.8 (12.1-13.6)	10.7 (10-11.5)

Across the simulated patients there was substantial variation in YLL adjusted for multimorbidity count (Figure 3.4).

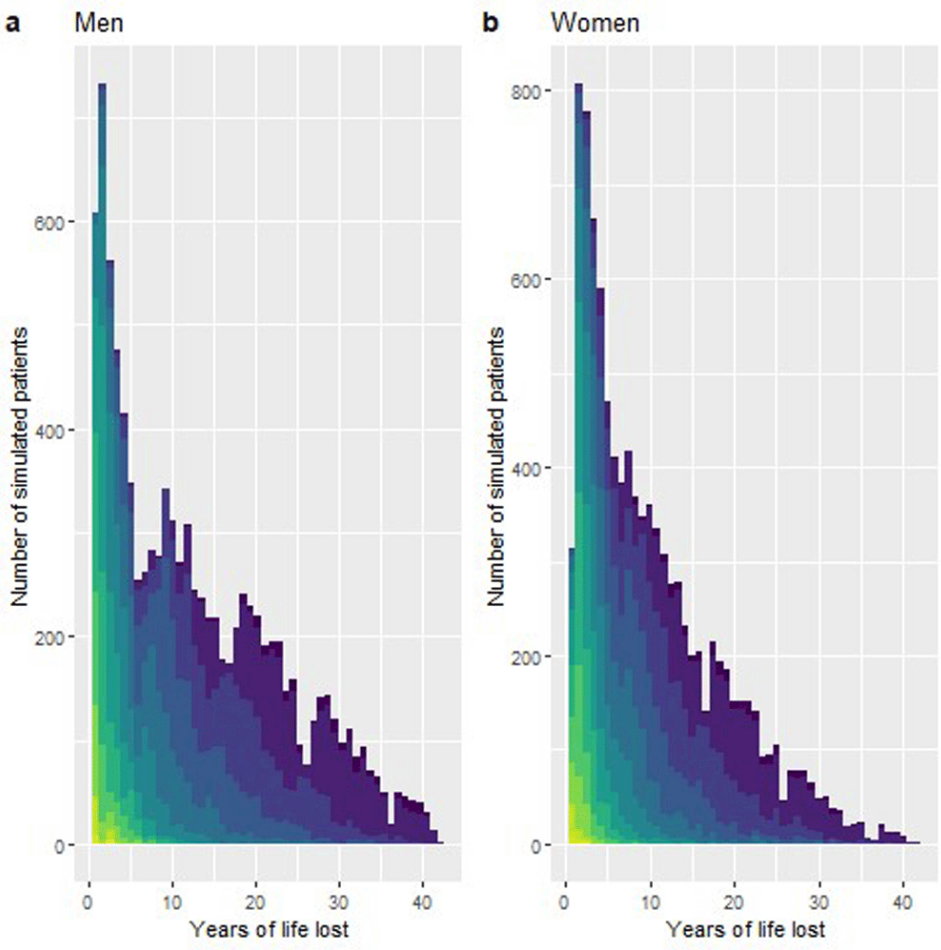

Figure 3.4: **YLL by sex.** Coloured bars indicate the multimorbidity count from zero (dark/blue) to 11 (light/yellow).

On stratifying the YLL estimates by sex, age and multimorbidity count (for the simulated patients) there were clear differences (Figure 3.5, Table 2) with the YLL

ranging from around 2-years per person in men or women aged 80 with large numbers of LTCs, to around 35 years in younger people without any LTCs (Table 3.2). For most age-bands and most multimorbidity counts the YLL per person remained above 5. In sensitivity analyses including the survival models with additional comorbidity-comorbidity and comorbidity-comorbidity-age interaction terms, (despite these models having a better fit based on AIC) than the model presented here, the YLL only changed minimally from that seen in the main analysis. This was true overall YLL for each sex (13.1, 95% CI 12.2–14.0 and 10.5; 95% CI 9.7–11.3 for men and women respectively) and on additionally stratifying on age and multimorbidity count (as shown in Table 3.2). For the latter comparison, the largest difference – 0.7 YLL – was seen in women aged 50–59 with six comorbidities. For most age-comorbidity bands the YLL was the same, to one decimal place, under both survival models.

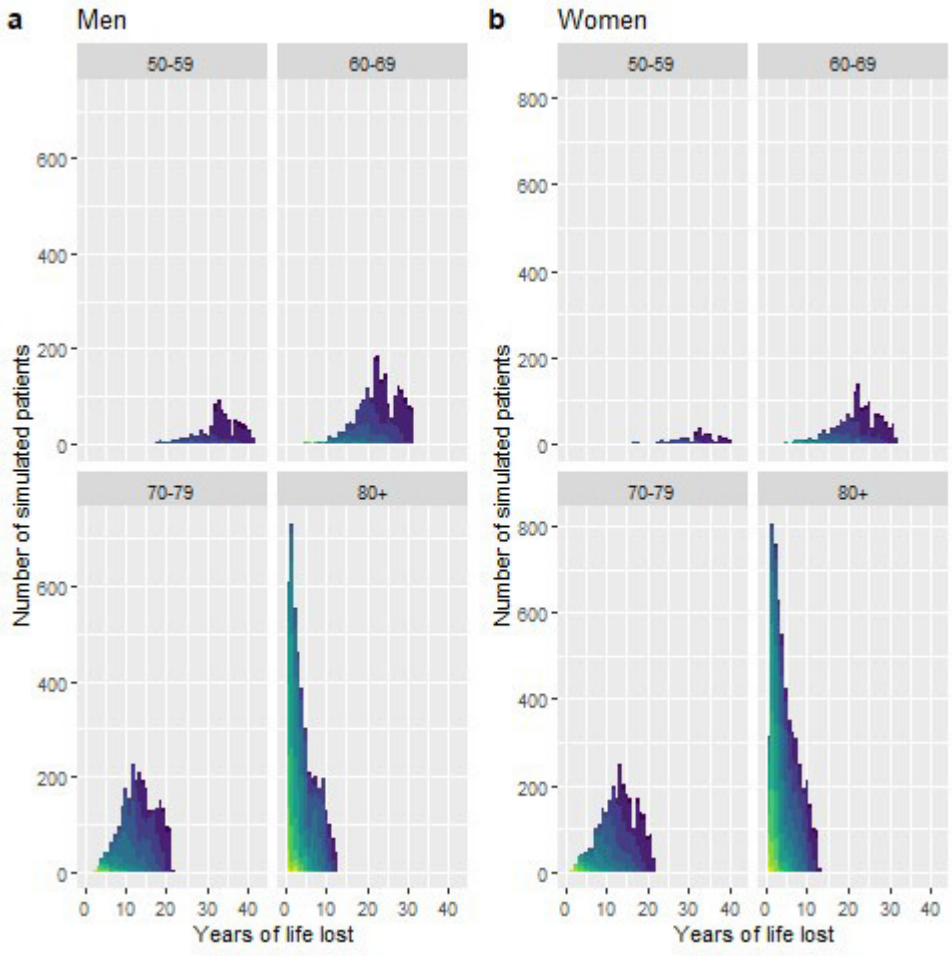

Figure 3.5: **YLL stratified by sex, age and multimorbidity count.** Coloured bars indicate the multimorbidity count from zero (dark/blue) to 11 (light/yellow).

Table 3.2: Mean years of life lost, accounting for type of long-term conditions, by age-band, sex and multimorbidity count. Estimates are based on life-expectancy calculates for specific types and combinations of LTCs, which are then aggregated across LTC counts.

	Men				Women			
Multi-morbidity Count	50-59	60-69	70-79	80+	50-59	60-69	70-79	80+
0	35.37	25.76	16.83	7.29	33.59	26.40	17.00	6.85
1	34.99	25.42	16.73	6.69	35.12	25.51	16.62	6.99
2	30.04	22.36	14.58	5.78	28.76	21.41	14.37	6.09
3	26.49	19.01	12.35	5.14	25.31	18.26	11.94	5.31
4	22.00	15.93	10.64	4.36	20.27	15.27	10.07	4.46
5	18.27	13.79	9.07	3.60	16.63	12.70	8.28	3.84
6	14.63	11.09	7.26	-	11.67	9.61	6.57	3.27
7	11.32	9.44	6.08	2.56	9.82	7.67	5.05	2.76
8	7.68	6.97	4.56	2.03	6.62	5.48	3.88	2.33
9	-	5.81	3.84	1.64	-	3.64	2.80	1.60
10	-	-	4.14	-	-	-	2.71	

3.5 Discussion

3.5.1 Summary of main findings

Using published data on people who have died from COVID-19 and survival models based on age and multimorbidity count in a general population in the United Kingdom, we estimated the burden (years life lost) from COVID-19 related mortality. We make a number of important observations. First, using the WHO GBD 2010 life tables as the reference [13], the estimated YLL was over a decade for COVID-19 deaths with 14 YLL in men and 12 in women. As such, mortality from COVID-19 represents a substantial burden to individuals and comparable to high burden LTCs such as ischaemic heart disease and chronic obstructive pulmonary disease. Second, YLL estimated from models using the prevalence of underlying LTCs based on patients dying from COVID-19 in Italy and age-, sex- and multimorbidity count-specific survival models in the UK did not drastically impact the YLL. Across both men and women, the number of YLL dropped to 11.6 and 9.4 years respectively. Third, across most age and multimorbidity count strata the estimated YLL per person remained substantial and generally above 5 years. This means that even after accounting for multimorbidity count, most individuals lost considerably more than the "1–2 years" suggested by some commentators [23] perhaps [24; 25] reflecting the high prevalence of multimorbidity in this population, especially in those over the age of 50 years [26; 27]. Finally, whilst the YLL remained high across most age- and multimorbidity count strata, the presence of multimorbidity did indeed influence the magnitude of the YLL. For example, in the elderly, over the age of 80, the estimated YLL in people with no LTCs was 7 years falling to less than two years with an increasing multimorbidity count.

YLL is a widely used metric to compare the relative impact of different causes of death and is used to guide policy-making and health service delivery and to prioritise interventions aimed at preventing deaths [28]. Using UK reports for approximate comparisons, the YLL in England and Wales for other conditions ranged, per capita from 8.2 for chronic obstructive pulmonary disease, 11.6 for coronary heart disease, 13.1 for pneumonia, and 21.6 for asthma [29]. Therefore, against these benchmarks, mortality from COVID-19 represents a substantial burden to individuals. It should be noted, however, that YLL for an emergent infection such as COVID-19, particularly in a pandemic, will be sensitive to the specific circumstances of the virus spreading, mitigation strategies, and potential future treatment or vaccines. These estimates, therefore, relate to the specific conditions at the time of modelling and will need to be updated particularly as vaccination or other strategies alter susceptibility or severity of infection. It is important to note, however, that it would be a misuse of any such modelling if it were used to criticise decision-making undertaken at the time.

The estimated YLL can vary substantially depending on the reference population chosen and the age distribution among those who die. Moreover, where attempts are made to account for underlying conditions in those who died, the accuracy will depend on the quality and completeness of data both for those deaths, and in the reference population used to obtain estimates of survival according to those underlying conditions. Nonetheless, although imperfect, we would argue that public health agencies should present estimates of YLL for COVID-19, alongside the more usual counts of deaths. We have already seen that if agencies do not do so, commentators can and will fill this vacuum, sometimes making substantial errors such as using life expectancy at birth to make inferences about the years of life lost by someone who has already lived into later life and thereby considerably underestimating the impact of the disease on individuals [23]. In additional to reporting YLL, metrics such as excess deaths and quality-adjusted life years are important to fully contextualise the loss of life seen in the pandemic.

It should be noted that these estimates were made early in the pandemic and could not account for specific patterns and events which emerged within the UK. For example, these analyses were performed before the impact of COVID-19 in care homes in the UK became apparent. SAIL contains data on all participants registered with a GP (and so would include care-home residents), however our estimates of life expectancy do not distinguish between people who live in care-homes and those who do not. As such our analyses would not reflect the YLL at a population level where care-homes are disproportionately impacted. Our estimates, given the data sources which were available at the time, are more likely to reflect the YLL of COVID-19 deaths among hospitalised patients.

Finally, our estimates of YLL only attempt to quantify the direct effects of COVID-19. Indirect impacts on mortality (e.g. through pressure on healthcare services of unintended consequences of lockdown measures) should also be considered, and are not captures by our YLL calculation.

3.5.2 Strengths and Limitations

Our analysis is novel in that it adjusts YLL for the number and type of underlying LTCs. This is important as people with underlying multimorbidity are recognised

to be more vulnerable to COVID-19. However, although we had data for eleven common and important LTCs, we did not have markers of underlying disease severity among those who died. Severity of the underlying LTC has considerable impact on life expectancy [30]. Moreover, we had no data for rarer severe LTCs, which may nonetheless be common among those who die from COVID-19 at younger ages. As such, the attenuation of YLL following adjustment for LTCs may be an underestimate. However, we think that this effect is unlikely to be substantial enough to reduce YLL to the orders of magnitude suggested by some commentators. Indeed, on stratifying by age and multimorbidity counts, we rarely found average YLLs of below three. Also, we were not able to adjust our estimates for other factors and exposures (such as socioeconomic status, occupation, smoking, health behaviours) which would have given a more accurate representation of life-expectancy in the absence of COVID-19.

Socioeconomic status is a particularly pertinent issue, as it may influence not only outcomes from infection (e.g. through multimorbidity and other risk factors) but also the likelihood of exposure (e.g. higher proportions of occupations for which home-working was not feasible). Since socio-economic status also predicts mortality there is a possibility of residual confounding due to the lack of data on socioeconomic status available for our models. To prevent mean inflation through rare deaths in younger people, who only modelled deaths in people over 50 years, however deaths among younger people may influence estimates YLL.

We did not have access to large quantities of individual-level data with which to estimate the prevalence of different combinations of LTCs. Therefore, we fitted a complex model (which was methodologically innovative and will be the subject of a separate publication) to estimate the joint probabilities, using the overall (marginal) estimates of each LTC, and the overall multimorbidity counts alongside a small amount of individual-level data from Scotland to help with model fitting. This model (i) represents the best estimate for the joint probabilities given the available data and importantly, (ii) the results for overall YLL remained substantially similar in widely different sensitivity analyses assuming either that LTCs are highly correlated among people dying from COVID-19 or that they are entirely independent.

Finally, given the emergent nature of the coronavirus pandemic, this study was conducted rapidly and under pressure of time. We chose the best data for age, sex and prevalence of LTCs that was available to us at the time of our modelling, but better-quality individual-level data specific to individual countries will yield substantially more reliable estimates. We would suggest that each public health agency should produce country-specific estimates, using the same LTC definitions in those who died as in the reference population and ideally to an agreed international protocol. Our study has used complex state-of-the-art statistical modelling and inference techniques, which rely on expensive computer simulations. We have also provided all our data (except individual-level data form the Scottish population, for which we provide a simulated substitute dataset) and code to allow others to check our modelling and correct any errors [15].

Our model, due to limited data available at the time, combined data on Covid-19 deaths and life expectancy data from different countries and contexts. While this synthesis of data sources allowed an estimation to be generated at an early stage, it limits the generalisability to specific contexts. Summaries of YLL relating

to a specific country or context should ideally use data (both life-expectancy and Covid-19 related) from that context. A comparison of such estimates (based on individual-level and country specific data) with our approach (modelling aggregate - and individual - level data from multiple sources early in the pandemic) would be important to test the utility of this approach for future pandemics.

Despite these limitations, our findings do indicate that adjusting for number and type of LTCs does not substantially reduce the estimated YLL compared to the standard approach. Our analysis does not, however, offer a definitive estimation of YLL across all contexts, nor does it necessarily fully adjust for underlying health status. For example, further work based in Scotland has illustrated that the life expectancy in care-home residents, and therefore the estimated YLL, is substantially different from the general population [31]. This is important given the large proportion of COVID-19 deaths that have occurred in care homes [32; 33]. Additionally, it indicates that additional factors are likely to influence underlying health status, life expectancy, likelihood of dying from Covid-19, and by extension YLL. These factors are not fully represented by the presence or absence of specific LTCs. Some of these factors are likely to be challenging to estimate from routine data alone, and producing YLL estimates which account for these factors should be an area of future investigation.

3.6 Conclusion

Among patients dying of COVID-19, there appears to be a considerable burden in terms of years of life lost, commensurate with diseases such as coronary heart disease or pneumonia. While media coverage of the pandemic has focused heavily on COVID-19 affecting people with 'underlying health conditions', and while the number and type of LTCs certainly influence the life expectancy and YLL for individuals, adjustment for number and type of LTCs only modestly reduces the estimated YLL due to COVID-19 compared to estimates based only on age and sex. Public health agencies and governments should report on YLL, ideally adjusting for the presence of underlying LTCs, to allow the public and policymakers to better understand the burden of this disease.

3.7 Data availability

All code, data (except individual-level data for Scotland), intermediate outputs and diagnostic plots are provided on GitHub: https://github.com/dmcalli2/covid19_yll_final.

3.7.1 Source data

Zenodo: Data and Code to support COVID-19 - exploring the implications of long-term condition type and extent of multimorbidity on years of life lost: a modelling study. https://doi.org/10.5281/zenodo.3751561 [34].

This project contains the source data used in performing this modelling study (except individual-level data for Scotland), which are also available via GitHub (https://github.com/dmcalli2/covid19_yll_final/tree/master/Data).

Individual-level data for Scotland are accessible via application to the electronic Data Research and Innovation Service (eDRIS) and the Public Benefit and Privacy Panel (PBPP) (www.isdscotland.org/Products-and-Services). Individual - level data for Wales are available via application to the Secure Anonymised Information Linkage (SAIL) at saildatabank.com. For both eDRIS and SAIL, individuals are required to complete information governance training, be affiliated with an appropriate organisation (e.g. a university, healthcare organisation, etc.) complete an application form, and the analysis must be performed to support research conducted in the public interest.

3.7.2 Extended data

Zenodo: Data to support COVID-19 - exploring the implications of long-term condition type and extent of multimorbidity on years of life lost: a modelling study. http://doi.org/10.5281/zenodo.3751561 34.

This project contains the archived scripts used during this modelling study, which are also available via GitHub (github.com/dmcalli2/covid19_yll_final /tree/master/Scripts).

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

References

[1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china," *The Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[2] W. H. Organization *et al.*, "World health organization coronavirus disease 2019 (covid-19) situation report," *Geneva: Switzerland: World Health Organisation*, vol. 1, no. 77, 2020.

[3] G. UK, "Guidance on social distancing for everyone in the uk," *Archived from the Original on*, vol. 24, 2020.

[4] R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu, *et al.*, "Estimates of the severity of coronavirus disease 2019: a model-based analysis," *The Lancet Infectious Diseases*, vol. 20, no. 6, pp. 669–677, 2020.

[5] C. Wu, X. Chen, Y. Cai, X. Zhou, S. Xu, H. Huang, L. Zhang, X. Zhou, C. Du, Y. Zhang, *et al.*, "Risk factors associated with acute respiratory distress syndrome and death in patients with coronavirus disease 2019 pneumonia in wuhan, china," *JAMA Internal Medicine*, vol. 180, no. 7, pp. 934–943, 2020.

[6] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, *et al.*, "Clinical course and risk factors for mortality of adult inpatients with covid-19 in wuhan, china: a retrospective cohort study," *The Lancet*, vol. 395, no. 10229, pp. 1054–1062, 2020.

[7] X. Yang, Y. Yu, J. Xu, H. Shu, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu, *et al.*, "Clinical course and outcomes of critically ill patients with sars-cov-2 pneumonia in wuhan, china: a single-centered, retrospective, observational study," *The Lancet Respiratory Medicine*, vol. 8, no. 5, pp. 475–481, 2020.

[8] D. Spiegelhalter, "How much'normal'risk does covid-19 represent," *BMJ*, vol. 2020, p. 059068, 2020.

[9] A. Banerjee, L. Pasea, S. Harris, A. Gonzalez-Izquierdo, A. Torralbo, L. Shallcross, M. Noursadeghi, D. Pillay, C. Pagel, W. K. Wong, *et al.*, "Estimating excess 1-year mortality from covid-19 according to underlying conditions and age in england: a rapid analysis using nhs health records in 3.8 million adults," *MedRxiv*, 2020.

[10] B. D. Jani, P. Hanlon, B. I. Nicholl, R. McQueenie, K. I. Gallacher, D. Lee, and F. S. Mair, "Relationship between multimorbidity, demographic factors and mortality: findings from the uk biobank cohort," *BMC Medicine*, vol. 17, no. 1, pp. 1–13, 2019.

[11] M. S. Chan, A. van den Hout, M. Pujades-Rodriguez, M. M. Jones, F. E. Matthews, C. Jagger, R. Raine, and M. Bajekal, "Socio-economic inequalities in life expectancy of older adults with and without multimorbidity: a record linkage study of 1.1 million people in england," *International Journal of Epidemiology*, vol. 48, no. 4, pp. 1340–1351, 2019.

[12] J. W. Gardner and J. S. Sanborn, "Years of potential life lost (ypll) — what does it measure?," *Epidemiology*, pp. 322–329, 1990.

[13] G. WHO, "Who methods and data sources for global burden of disease estimates 2000–2011," *Geneva: Department of Health Statistics and Information Systems*, 2013.

[14] I. Sanità, "Characteristics of sars-cov-2 patients dying in italy," *Report based on available data on March 26th*, 2020.

[15] D. McAllister, "Supplementary material," *github.com/dmcalli2/covid19_yll_final*, 2020.

[16] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," *Medical Care*, pp. 8–27, 1998.

[17] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation," *Journal of Chronic Diseases*, vol. 40, no. 5, pp. 373–383, 1987.

[18] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L. D. Saunders, C. A. Beck, T. E. Feasby, and W. A. Ghali, "Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data," *Medical Care*, pp. 1130–1139, 2005.

[19] D. Metcalfe, J. Masters, A. Delmestri, A. Judge, D. Perry, C. Zogg, B. Gabbe, and M. Costa, "Coding algorithms for defining charlson and elixhauser comorbidities in read-coded databases," *BMC Medical Research Methodology*, vol. 19, no. 1, pp. 1–9, 2019.

[20] C. H. Jackson, "flexsurv: a platform for parametric survival modeling in r," *Journal of Statistical Software*, vol. 70, 2016.

[21] M. Plummer *et al.*, "Jags: A program for analysis of bayesian graphical models using gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, vol. 124, pp. 1–10, Vienna, Austria., 2003.

[22] M. J. Denwood, "runjags: An r package providing interface utilities, model templates, parallel computing methods and additional distributions for mcmc models in jags," *Journal of Statistical Software*, vol. 71, pp. 1–25, 2016.

[23] T. Young, "Has the government overreacted to the coronavirus crisis," *The Critic*, vol. 31, 2020.

[24] "Covid kills, but do we overestimate the risk?," *Financial Times*, 2021.

[25] "Two thirds of coronavirus victims may have died this year anyway, government adviser says," *The Telegraph*, 2020.

[26] K. Barnett, S. W. Mercer, M. Norbury, G. Watt, S. Wyke, and B. Guthrie, "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study," *The Lancet*, vol. 380, no. 9836, pp. 37–43, 2012.

[27] R. A. Payne, S. C. Mendonca, M. N. Elliott, C. L. Saunders, D. A. Edwards, M. Marshall, and M. Roland, "Development and validation of the cambridge multimorbidity score," *Cmaj*, vol. 192, no. 5, pp. E107–E114, 2020.

[28] R. Martinez, P. Soliz, R. Caixeta, and P. Ordunez, "Reflection on modern methods: years of life lost due to premature mortality—a versatile and comprehensive measure for monitoring non-communicable disease mortality," *International Journal of Epidemiology*, vol. 48, no. 4, pp. 1367–1376, 2019.

[29] N. Digital, "Nhs digital: Compendium: Years of life lost," 2019.

[30] R. M. Shavelle, D. R. Paculdo, S. J. Kush, D. M. Mannino, and D. J. Strauss, "Life expectancy and years of life lost in chronic obstructive pulmonary disease: findings from the nhanes iii follow-up study," *International Journal of Chronic Obstructive Pulmonary Disease*, vol. 4, p. 137, 2009.

[31] J. K. Burton, M. Reid, C. Gribben, D. Caldwell, D. N. Clark, P. Hanlon, T. J. Quinn, C. Fischbacher, P. Knight, B. Guthrie, *et al.*, "Impact of covid-19 on care-home mortality and life expectancy in scotland," *Age and Ageing*, vol. 50, no. 4, pp. 1029–1037, 2021.

[32] A. Comas-Herrera, J. Zalakaín, E. Lemmon, D. Henderson, C. Litwin, A. T. Hsu, A. E. Schmidt, G. Arling, F. Kruse, and J.-L. Fernández, "Mortality associated with covid-19 in care homes: international evidence," *Article in LTCcovid. org, international long-term care policy network, CPEC-LSE*, vol. 14, 2020.

[33] M. Morciano, J. Stokes, E. Kontopantelis, I. Hall, and A. J. Turner, "Excess mortality for care home residents during the first 23 weeks of the covid-19 pandemic in england: a national cohort study," *BMC Medicine*, vol. 19, no. 1, pp. 1–11, 2021.

[34] P. Hanlon and D. McAllister, "Data and code to support covid-19 - exploring the implications of long-term condition type and extent of multimorbidity on years of life lost: a modelling study (version 1)," *Zenodo*, 2020.

Chapter 4

Combining rapid antigen testing and syndromic surveillance improves community-based COVID-19 detection in a low-income country

This chapter has been published in: Chadwick, F.J., Clark, J., Chowdhury, S. et al. Combining rapid antigen testing and syndromic surveillance improves community-based COVID-19 detection in a low-income country. Nat Commun 13, 2877 (2022). https://doi.org/10.1038/s41467-022-30640-w. Additional results and methodological details available in Appendix B.

The L.I.E.S. Framework

The observation processes in the following case study include:

Latency: the true COVID-19 status is hidden and must be inferred from observed variables (symptoms and rapid antigen test results). The observed variables are binary realisations (present/absent) of continuous immunological processes.

Identifiability: (*mathematical*) the correlations between the binary observed variables are modelled using a multivariate probit model in which the mean and variance are not identifiable.

Identifiability: (*practical*) the model is purely predictive meaning practical identifiability issues could manifest as poor prediction or classification.

Scaling: for purely predictive models, scaling is only relevant in terms of the cross-validation structure (i.e. do model tests match how they will be implemented). In this case, the data arrive in two-week intervals which provides a natural scale for testing.

The application of the framework to the case study will be discussed in more detail in the conclusion.

4.1 Abstract

Diagnostics for COVID-19 detection are limited in many settings. Syndromic surveillance is often the only means to identify cases but lacks specificity. Rapid antigen testing is inexpensive and easy-to-deploy but can lack sensitivity. We examine how combining these approaches can improve surveillance for guiding interventions in low-income communities in Dhaka, Bangladesh. Rapid-antigen-testing with PCR validation was performed on 1172 symptomatically-identified individuals in their homes. Statistical models were fitted to predict PCR-status using rapid-antigen-test results, syndromic data, and their combination. Under contrasting epidemiological scenarios, the models' predictive and classification performance was evaluated. Models combining rapid-antigen-testing and syndromic data yielded equal-to-better performance to rapid-antigen-test-only models across all scenarios with their best performance in the epidemic growth scenario. These results show that drawing on complementary strengths across rapid diagnostics, improves COVID-19 detection, and reduces false-positive and -negative diagnoses to match local requirements; improvements achievable without additional expense, or changes for patients or practitioners.

4.2 Introduction

Identification and isolation of COVID-19 cases remains key to the pandemic response. The faster and more accurately cases can be identified, the more effectively clinical care can be provided, and transmission reduced through targeted interventions. Real-time PCR has rapidly become the gold-standard test for SARS-CoV-2 detection (although Dramé et al point out that, with less than 100% sensitivity, PCR falls short of being a true gold-standard)[1] due to its high sensitivity and specificity.[2] However, turnaround can be slow and access to laboratory diagnostics is limited in many parts of the world. As such, syndromic surveillance has often been the primary means of case identification for guiding individual and population-wide mitigation measures.[3; 4] Rapid antigen tests are an increasingly popular alternative to PCR as they have high specificity, and are less expensive, easier to perform, and faster, returning results within 20 minutes. Hence, rapid antigen tests have potential to greatly decrease the time and expense associated with case detection, but concerns have been raised that their lower sensitivity leads to unacceptably high false negative diagnoses.[5; 6; 7; 8] Improving COVID-19 diagnosis is a priority and, therefore, requires us to better harness imperfect but fast and inexpensive methods, particularly for individual diagnosis but also for population-level surveillance.[9]

Syndromic surveillance has been used since the start of the pandemic.[10] The COVID-19 case definition was based on early data from clinical cases,[11] but, as the virus has evolved and spread, the clinical picture of COVID-19 has changed. Updated case definitions have improved, though are necessarily non-specific and generate many false positive diagnoses (and ignores asymptomatic cases entirely). [12; 13] A natural extension is syndromic modelling, whereby symptomatic and risk factor data are used to fit a model to allow more accurate prediction of how likely a patient is to have COVID-19. [14] However, disease syndromes change between populations, when new variants emerge, and as

other diseases become more or less common, [12; 15] which can make syndromic models perform poorly in new settings across space and time. This is a particular challenge for seasonal respiratory pathogens, where symptoms often co-occur and are non-specific.[12]

A key limitation of both rapid tests and syndromic surveillance is their low effectiveness at COVID-19 detection in asymptomatic patients. Asymptomatic cases are known to play a role in driving transmission.[16] Resource limitations mean that many health agencies and governments have exclusively or temporarily targeted interventions towards symptomatic individuals to reduce transmission. Asymptomatic cases can still be identified through contact tracing from symptomatic patients. Reliable diagnosis of symptomatic cases of COVID-19, therefore, is a priority in many settings and is the focus of this paper.

Even for symptomatic patients, neither rapid tests nor syndromic surveillance can match PCR in terms of both sensitivity and specificity. However, lower sensitivity and specificity may be admissible depending on the scale and impact of misclassification.[17] Indeed, there are costs to both individuals and societies that must be considered when making policy decisions to the determine the most appropriate approach to testing. Low specificity means more common COVID-19 misdiagnoses (false positives), leading to unnecessary self-isolation, which is expensive to individuals and society.[18] Low sensitivity means COVID-19 cases will be missed (false negatives) and mitigation measures not put in place leading to increased transmission and disease burden.[19] These misclassifications are complementary for a given diagnostic, meaning increasing specificity will lead to decreased sensitivity, and vice versa.

The typical approach is to balance sensitivity and specificity to maximise the number of correct classifications and assume that both misclassification types are equally costly. The costs of false positives and false negatives, however, vary enormously depending on the intersection of perspective, economic and epidemiological concerns. An individual may be motivated to secure a false negative diagnosis if there is insufficient support for self-isolation. In contrast, at the government level, false negatives may be acceptable if the economic cost of supporting those individuals is less than the cost of accelerating case rates. The epidemiological context will also alter the impact of false positives and false negatives. For example, if the disease is prevalent or increasing the priority of both individuals and governments may be to curb transmission and reduce impacts as quickly as possible. In this instance, false negatives have an outsized and costly impact by increasing the number of contact events occurring in the population and delaying control measures by underestimating epidemic size.[19] In contrast, under low prevalence, false negatives will be correspondingly low so even a high false negative rate (low sensitivity) will have modest impact, but small decreases in specificity will lead to a large number of expensive false positives.[20] In practice the situation will be more nuanced and modulated by testing capacity constraints, requiring a balance to be struck.[17]

The best diagnostic approach for surveillance will therefore be one where correct classifications have highest value and misclassifications have lowest cost. Here, we examine the use of rapid antigen testing and syndromic surveillance of COVID-19 in symptomatic patients from low-income communities in Dhaka, Bangladesh, where a large volunteer workforce supports COVID-19 diagnosis,

care and prevention. In this context, community-based workers used a mobile phone-based application to record patient symptoms and provide advice and support services, with a diagnostic algorithm deployed on the app to inform their provisioning. This algorithm could be updated in real-time depending upon the epidemiological context to allow appropriate tailoring of service provision (although was not updated during the study period).

Here, we demonstrate that by combining rapid antigen testing and syndromic surveillance we can draw on their complementary strengths, ameliorate their respective weaknesses, and tune them for different epidemiological scenarios. We compare their performance alone and in combination for general prediction and as diagnostics under three scenarios with different misclassification requirements determined by government policy-makers. Overall, we show that the optimised combined models achieve equal-to-much-lower error rates than the rapid antigen test- or syndromic surveillance-only in all metrics, and how integrating data from multiple rapid testing methods can improve diagnostics, particularly when adapted to local situations.

4.3 Methods

4.3.1 Data Collection

Recruitment took place across low-income communities in Dhaka North Community Corporation between 19 May 2021 and 11 July 2021. Participants were identified for COVID-19 testing by CSTs. CSTs are community-based volunteer health workers trained to identify individuals reporting symptoms suggestive of COVID-19 through hotline calls or community-based reporting channels. Probable cases identified by CSTs are counselled to isolate for 14 days under household quarantine, connected to telemedicine services for home-based COVID-19 management, and provided with over-the-counter medication or medical referrals if the case is severe. CSTs submit surveillance data to a centralised database through a mobile-phone-based application (Supplementary Materials (Data Collection)).

Participants were selected for testing if they were over 15 years old, had a fever (>38°C) at the point of assessment, and one or more of 14 symptoms listed in Table 4.1. CSTs collected the enrolled individual's age and gender, and took two nasal swabs. One swab was used for rapid antigen testing (SD Biosensor STANDARD^{TM} Q COVID-19 Ag Test BioNote) at the household, and the other returned under cold-storage to the Institute of Epidemiology, Disease Control and Research (IEDCR) for PCR testing. The full questionnaire and testing protocols are provided in Supplementary Methods.

Participants provided written informed consent to sample collection and for their results to be analysed in the study. The study protocol was approved by the Institutional Review Board at the IEDCR, Ministry of Health, Bangladesh, IEDCR/IRB/04.

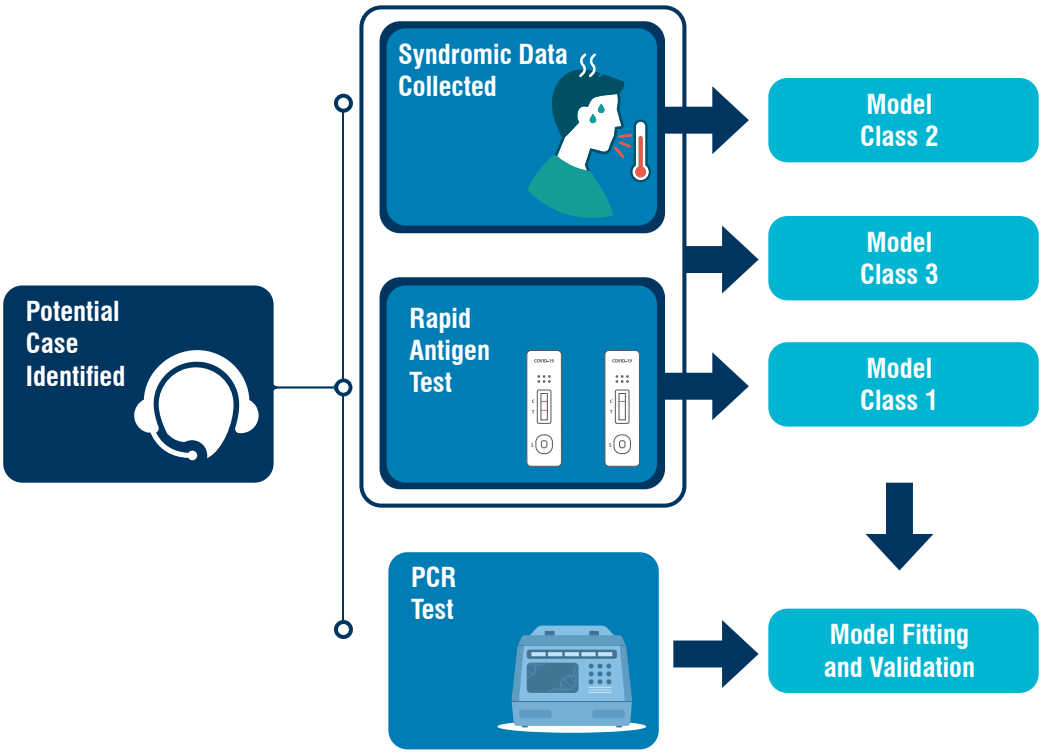

Figure 4.1: **Schematic description of identification of likely COVID-19 cases by community support teams (CSTs) and model definitions.** CSTs collect syndromic data (age, gender and presence/absence of 14 predetermined symptoms), and two sets of naso-pharyngeal swabs (for rapid antigen testing and PCR). We used three model classes: rapid-antigen-test-only in 1, syndromic data only in 2, and both rapid-antigen-test and syndromic data in 3. The PCR result is used to train and test each model using temporal cross-validation.

4.3.2 Statistical Modelling

Structure

We developed three model classes using: 1. the rapid-antigen-test result; 2. the syndromic data, and 3. the two data sources combined (Figure 4.1). We identified cases by PCR. As RAT-only used the rapid-antigen-test result, no statistical model is needed. For Syndromic-only, we used a Bayesian multivariate probit model,[21] with multivariate referring to multiple response variables. The multivariate probit structure allows the model to account for the binary and correlated nature of the symptoms, while conditioning on the risk factors of age and gender, thereby improving over models which implicitly assume independence between symptoms. By using a Bayesian formulation, we generate full posteriors for our parameter estimates, allowing natural quantification of uncertainty. We chose minimally informative priors, with standard normals for the covariates and intercepts and a flat LKJ distribution for the correlation matrix (described in more detail in Supplementary Materials: Statistical Methodology).

For Syndromic - RAT Combined, we use a hurdled multivariate probit. The

approach exploits the specificity of rapid antigen tests by treating rapid test-positives as cases. While this sounds like a strong assumption, this simply translates in practice to telling all rapid test-positive individuals to assume they have COVID-19. Rapid - antigen - test - negative individuals are then modelled using the sensitive syndromic approach of Syndromic - only to capture PCR - positives missed by the rapid antigen test. This approach leverages the potentially different syndromic profiles of PCR-positive patients who are rapid-antigen-test-positive and -negative, allowing the model to adapt solely to the latter. The models were fitted to the data using Bayesian inference techniques based on Hamiltonian Monte Carlo in the Stan programming language.[22] Further technical details and model equations are presented in Supplementary Methods.

Model Selection

For model selection and all measures of performance, we used out-of-sample, temporal cross-validation (Figure 4.2), where training and testing data are separated based on time. We structured the cross-validation temporally to reflect the real-world prediction problem: using recent testing data to predict new cases. Due to the changing nature of the disease and its management over time, using unstructured cross-validation would result in an overstatement of model performance.

We conducted backwards model selection, starting with the most complex biologically plausible model, to identify a subset of models with the highest predictive power. Shrinking the number of possible models was necessary to lower computational demand and reduce the risk of overfitting. The large number of symptoms corresponds to many potential model configurations (>131 000 for 14 symptoms and two covariates) which might perform well on the test sets by chance (even under temporal cross-validation) but lack transferability to novel situations. The Bayesian multivariate probit structure common to these models directly estimates the full posterior correlation matrix for the PCR-status and other symptoms. By first using the strength of the correlation with the PCR-status (coarse selection, Figure 4.2) and general predictive power (fine selection, Figure 4.2) to narrow down the number of candidate models, and then testing those models under the epidemiological scenarios, we are more likely to choose models that generalise well to new data (Supplementary Materials (Statistical Methodology)).

Measuring Model Performance

We assessed models using three sets of increasingly policy-relevant criteria. First, we use predictive performance to measure model performance in a decision -free context (i.e. comparing predicted probabilities of an individual having COVID-19 to their true status). Second, we use receiver operating characteristic (ROC) curves to show generic model classification performance. Finally, we measure classification performance under three epidemiological scenarios (defined in Table 4.2).

We scored the models' predictive power using cross-entropy (defined in Supplementary Methods). Cross-entropy measures the accuracy of predicted probabilities of binary outcomes, rather than making binary classifications, similar in

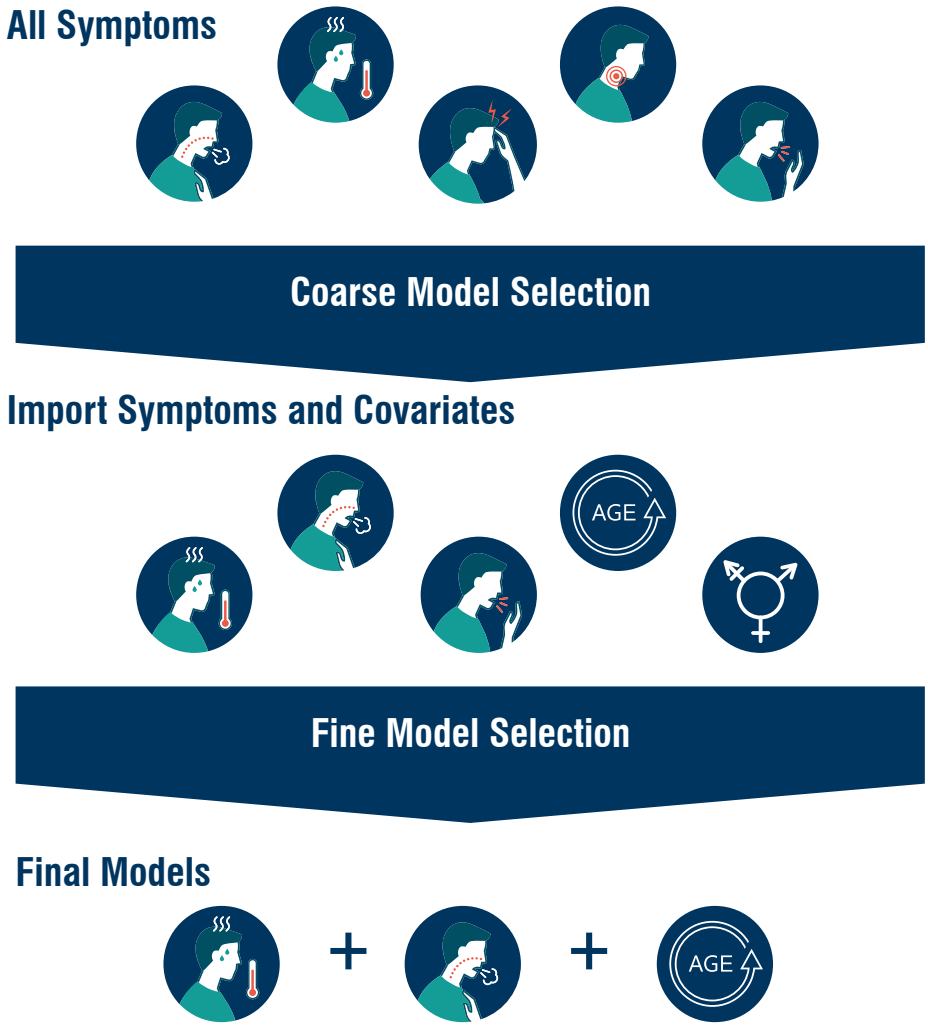

Figure 4.2: **Model selection procedure.** Rounds of model selection in the multivariate probit component of the Syndromic-only and Syndromic-Rapid Antigen Test (RAT) Combined models. With 14 symptoms (5 shown for demonstration purposes) and two covariates there are over 131 000 possible model combinations. To make exploring these models computationally feasible and to reduce the risk of overfitting, we carried out two rounds of model selection. A subset of symptoms are identified using the strength of posterior correlation between each symptom and PCR-status identified by the corresponding model, with the weakest correlated symptoms removed during each round of selection. From this subset of symptoms, a more exhaustive search of potential models is then conducted to identify the best symptom-covariate relationships, using temporal-cross validation to measure model performance. The best model for each level of complexity (i.e. number of symptoms) are then used as our candidate models. Only these final models are used for classification. This reduces the set of models tested as classifiers from >131 000 to just four per model class.

concept to a mean square error for normally-distributed data, but adapted for binary data.[23] A cross-entropy of zero indicates a model that predicts with certainty the correct result each time. A random classifier for the problem scored 11.54

In practice, models are often evaluated on their performance as deterministic classifiers rather than as stochastic prediction engines (i.e. their ability to classify an individual as a COVID-19 case or not, rather than the probability that the individual is a case). Deterministic classification requires that a probability threshold is chosen over which patients are classified as COVID-19 positive. Classifier performance was compared generically (using ROC curves to look at the error rates that can be achieved with each model without specifying a scenario). Generic performance here is only used to show the flexibility of the model classes, i.e. model performance without reference to a specific scenario. The best model for a local situation can only be determined if the relative costs of false positives and negatives are considered.

We compare model performance under three scenarios (using error terms described in Table 4.2) developed for illustrative purposes through discussion with colleagues at IEDCR. In Scenario 1, we do not consider epidemiological context but minimise false negative and false positive rates equally by maximising the correct classification rates individually and in total, as measured by the harmonic mean (not the arithmetic mean which would maximise the rates in total, Supplementary Methods). Scenario 2 corresponds to epidemic growth as experienced during the spread of the Delta variant during the period of data collection. Under these circumstances, false negatives are costly relative to false positives. In Scenario 3, incidence is assumed to be low and relatively stable. In this situation, policy-makers may prioritize keeping false positive diagnoses low to prevent fatigue and to keep the workforce active.

4.4 Results

4.4.1 Population Characteristics

Of 1241 participants enrolled by community support teams across Dhaka, 1172 (94%), had complete data available for analyses. The remainder were removed due to duplicated sample identification codes that prevented reliable matching of test results to symptom metadata. These duplications occur at random, due to human error, and we do not believe they could bias results. Patient summaries by age, gender, case positivity and symptoms are presented in Table 4.1. No participants had been vaccinated as the study pre-dated mass vaccination in low-income communities in Dhaka and only symptomatic patients were included in this study because they were the local government priority for support. Case positivity measured by PCR in Dhaka increased from 15.8% to 23.8% from the first (19th-26th May 2021) to the last week (4th-11th July 2021) of the study, corresponding to prevalence rising from 1.4 to 13.8 confirmed cases per 100 000 people [24].

Table 4.1: Breakdown of patient numbers by age and gender, in relation to case positivity by PCR and reported symptoms (both as % rounded to nearest integer). Although age is binned here, raw age in years was used for analyses. Furthermore, in the survey non-binary genders were permitted but none reported.

				Symptoms (%)														
Age (years)	Gender	Count	Positivity Rate (%)	Breathing Problems	Cough (Any)	Cough (Dry)	Cough (Wet)	Diarrhoea	Ongoing Fever	Headache	Loss of Smell	Loss of Taste	Muscle Pain	Red Eyes	Runny Nose	Sore Throat	Tiredness	Vomiting
16-25	Women	124	19	23	73	69	19	4	94	77	38	51	52	10	49	43	73	19
16-25	Men	157	20	20	74	72	22	5	91	73	44	45	50	10	36	42	62	13
26-35	Women	144	17	25	72	70	19	10	90	75	35	42	51	4	40	43	69	7
26-35	Men	178	26	26	80	78	14	10	89	74	38	38	49	7	38	33	69	16
36-45	Women	101	26	28	79	77	25	4	93	78	38	48	53	5	47	42	72	18
36-45	Men	119	24	23	75	71	18	7	89	71	38	38	55	8	39	41	67	8
46-55	Women	66	20	17	74	74	15	3	86	70	32	32	55	0	35	33	58	15
46-55	Men	58	22	16	55	55	14	2	84	57	34	34	52	10	45	33	69	7
56+	Women	57	23	25	72	68	23	11	84	54	33	30	49	4	32	26	60	14
56+	Men	61	26	30	66	64	15	5	77	59	41	36	49	8	36	23	52	11
All		1065	22	23	74	71	19	7	89	71	38	41	51	7	40	38	66	13

4.4.2 Model Selection

Backwards model selection using strength of posterior correlation with outcome (Methods (Statistical Modelling: Model Selection)) for both the multivariate probit syndromic data only model and the thresholded multivariate probit syndromic data with rapid antigen test result (hereafter the Syndromic-only and Syndromic-RAT combined) models showed a marked decline in predictive power at more than 4 symptoms. The final four symptoms retained in Syndromic-only were loss of smell, ongoing fever, diarrhoea and loss of taste and in Syndromic-RAT combined were ongoing fever, wet cough, loss of smell and dry cough. The symptoms are listed in reverse order of importance as determined by model selection (i.e. all four symptoms were retained in the four symptom model, the first was removed in the three symptom model, the second was also removed in the two symptom model etc.) and the median estimated correlations can be seen in the Supplementary Results (Supplementary Figures 1 and 2). The covariate gender was dropped for both model classes while age was dropped in the Syndromic-RAT combined class but retained in the Syndromic-only class.

4.4.3 Predictive Performance

In the comparison of predictive performance under out-of-sample temporal cross - validation (Methods (Statistical Modelling: Model Performance)), RAT - only (rapid-antigen-test result) performed worst with a cross-entropy of 3.18 (cross-entropy values further from zero correspond to worse predictive performance). The median cross-entropy values were between 2.71 and 2.78 for Syndromic-only models. Syndromic-RAT combined models performed best with cross-entropy values between 1.56 and 1.6 (Figure 4.3).

4.4.4 Classification Performance

Generic model classification performance under out-of-sample temporal cross-validation (Methods (Statistical Modelling: Model Performance)) for the one and four symptom models in the Syndromic-only and Syndromic-RAT Combined classes is shown by their ROC curves (Figure 4.4). The curves for the models of different complexities are extremely similar (as are the two and three symptom model curves, not shown), however, note that the four symptom model has higher precision and granularity across both axes. The RAT-only model is a binary test (rapid-antigen-test positive or negative) and so the ROC is a single value, not a curve, with false positive rate of 0.02 and a false negative rate of 0.45.

4.4.5 Scenario-Specific Performance

Scenario-specific classification performance under out-of-sample temporal cross-validation (Methods (Statistical Modelling: Model Performance)) is shown in Figure 4.5. Across all scenarios (defined in Table 4.2), the best models in Syndromic-RAT Combined that used both the rapid antigen testing and syndromic data performed equally well or better than the other two model classes. In Scenario 1 (“Agnostic”, wherein the correct classification is maximised, assuming equal

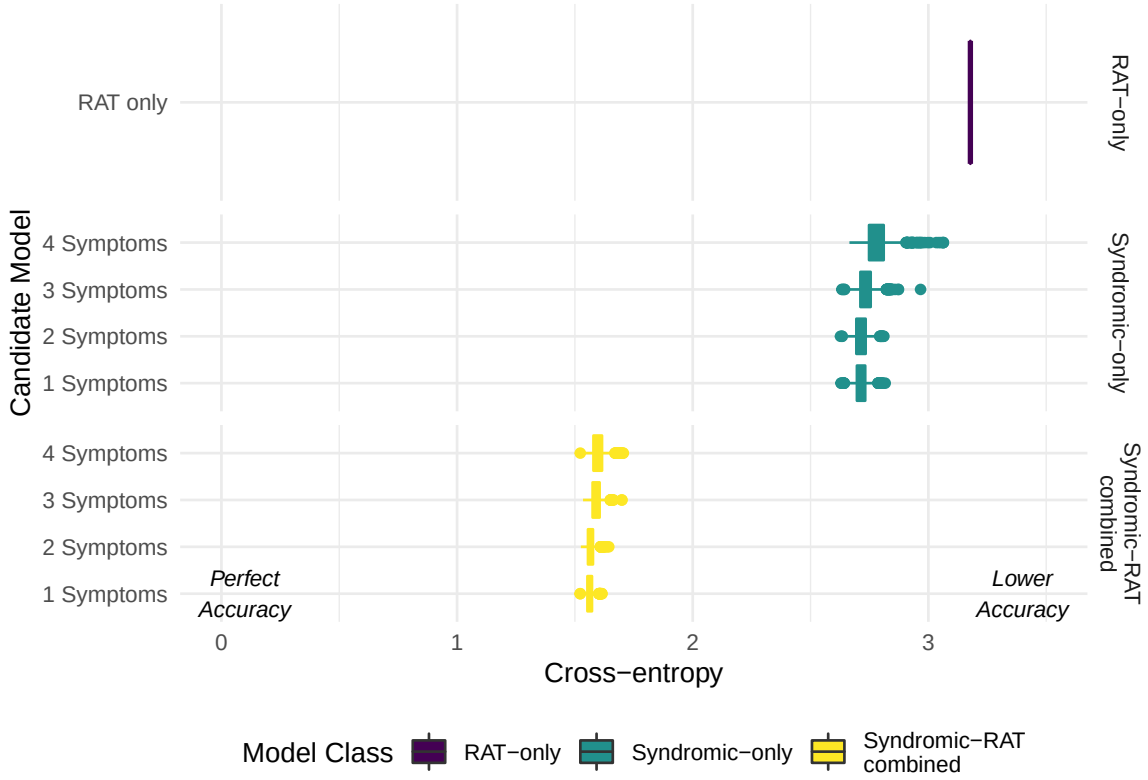

Figure 4.3: **Model Predictive Performance.** Predictive performance of candidate models were measured using out - of - sample cross - entropy. Combined posterior median and interquartile ranges for n = 1172 biologically independent individuals predicted under temporally - structured cross - validation. Cross - entropy shows the most generalised-level of model predictive power, assessing performance in the probability scale without requiring classification threshold decisions. A cross - entropy of zero indicates a model that predicts with certainty the correct result each time. A random classifier for the problem scored 11.54. Interquartile ranges are shown for the posterior cross-entropy of the best candidate models at each level of model complexity tested under temporal cross - validation. The intermediate complexity models perform best at prediction, although performance is similar across all the models within each model class. There was a marked decline in predictive power at more than four symptoms, leading us to choose this as the maximum complexity model in our candidate models. Model classes are colour-coded, the rapid-antigen-test only (RAT-only) model is purple, Syndromic - only model is teal, and the Syndromc - RAT Combined model is yellow.

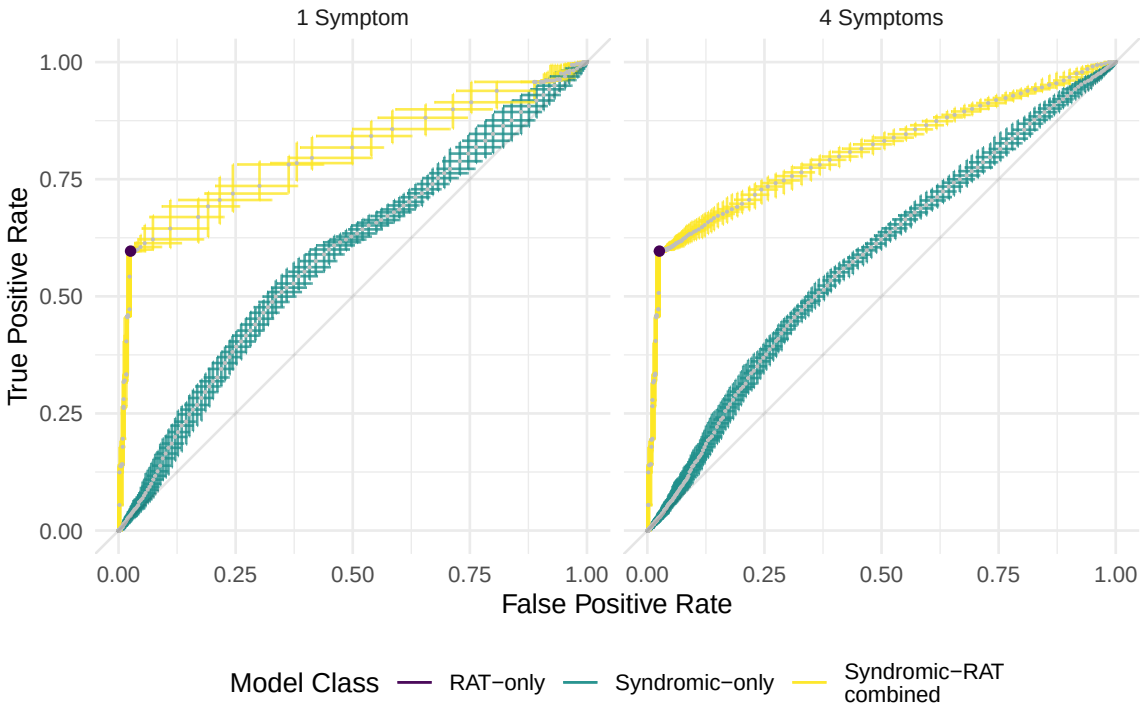

Figure 4.4: **Generic Model Classification Performance.** Median (grey dots) and interquartile ranges for receiver operating characteristics (ROC) for rapid-antigen-testing-only approach (purple) and posterior median and interquartile range ROC for Syndromic - only (teal) and Syndromic - Rapid Antigen Test (RAT) Combined (yellow) models for n = 1172 biologically independent individuals predicted under temporally - structured cross - validation. In the RAT-only model, the ROC is a single value (i.e. a dot rather than a line) as the binary test has a single sensitivity and specificity. In Syndromic - only and Syndromic - RAT Combined classes, the ROC values demonstrate the performance of the model for any hypothetical scenario as defined by the axes (as opposed to Figure 4.5 which demonstrates model performance in specific epidemiological scenarios which are realisations of single points in this space). While ROC plots are often plotted as curves, we do not have continuous probability values due the binary nature of predictor symptoms. This is important as discontinuity in the probabilities impacts the sensitivity of the model to classification thresholds, such as those used in the scenarios below.

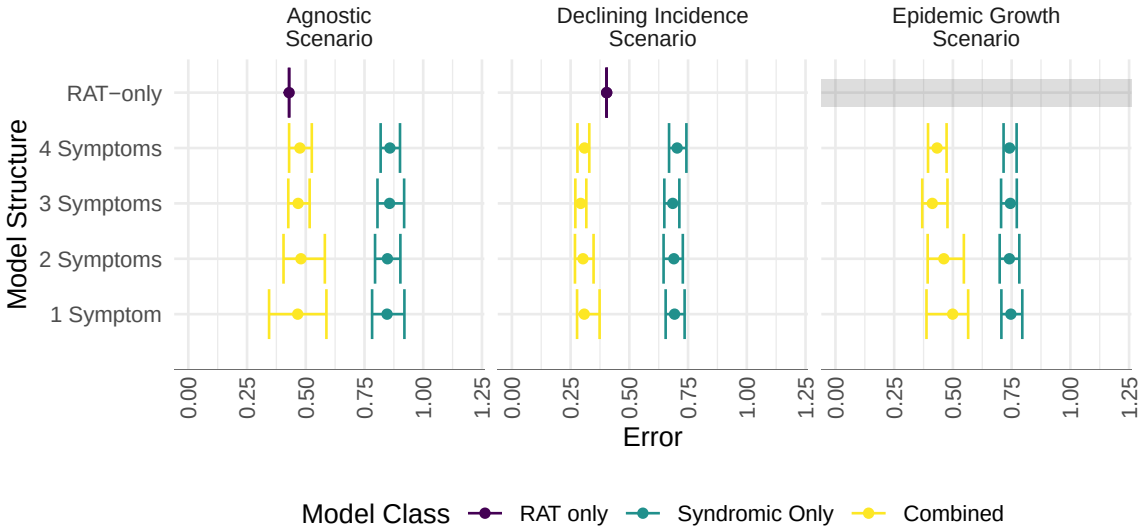

Figure 4.5: **Performance of models under three epidemiological scenarios.** Combined posterior median and interquartile ranges of error rates for n = 1172 biologically independent individuals predicted under temporally - structured cross - validation. In the Agnostic Scenario, the model is optimised to maximise the correct classification rate with error measured as the sum of the false positive and false negative rates. In the Epidemic Growth Scenario, a maximum false negative rate of 20% is permitted, and the error is measured as the false positive rate. In the Declining Incidence scenario, a maximum false positive rate of 20% is permitted, and the error is measured as the false negative rate. These requirements were determined through discussion with colleagues at the Institute of Epidemiology and Disease Control (IEDCR), Bangladesh. The plot shows the posterior median and interquartile range for scenario-specific errors. Lower errors correspond to better model performance. There is no error rate defined for rapid-antigen-testing-only model (RAT-only) in the Epidemic Growth Scenario as the model failed to meet the requirement for that scenario (indicated by grey bar). Model classes are colour-coded, the RAT - only model is purple, Syndromic - only model is teal, and the Syndromc - RAT Combined model is yellow.

costing of false positives and false negatives, Table 4.2), models in RAT-only and Syndromic-RAT Combined classes performed equally well (overlapping posterior interquartile ranges) and distinctly better (no overlap in posterior interquartile range) than models in the Syndromic-only class. The median errors, as defined in Table 4.2, were 0.43 for models in RAT-only and Syndromic-RAT Combined and between 0.85 and 0.86 for Syndromic-only models (Figure 4.5). In Scenario 2 (“Epidemic Growth”, wherein false negative rates must be below 20%, Table 4.2), the RAT-only models failed to meet the scenario-requirement. The median errors were between 0.74 and 0.75 for Syndromic-only models, and 0.41 and 0.5 for Syndromic-RAT Combined models (Figure 4.5).

In Scenario 3 (“Declining Incidence”, wherein false positive rates must be below 20%, Table 4.2), Syndromic-only again performed worst, and Syndromic-RAT Combined achieved the lowest error, with RAT-only falling between the two (closer to Syndromic-RAT Combined than Syndromic-only). The error in RAT-

Table 4.2: Requirements and performance criteria for each epidemiological scenario. The requirement refers to a base level of performance the model must achieve, allowing the more flexible models to be adapted to meet that requirement as closely as possible (e.g. by determining a classification threshold). These requirements were determined through discussion with colleagues at the Institute of Epidemiology and Disease Control (IEDCR), Bangladesh, using internal resource projections. The performance criterion is used to determine which model performs the 'best' given that the requirement has been met.

Scenario Name	Requirement	Performance Criterion (Error)
1 Agnostic	Maximise correct classification rates	Sum of error rates
2 Epidemic Growth	<20% false negative rate	False positive rate
3 Declining Incidence	<20% false positive rate	False negative rate

only was 0.03 and the median errors ranged from 0.19 to 0.2 for Syndromic-only, and 0.19 to 0.2 for Syndromic-RAT Combined (Figure 4.5). The results for each scenario-model combination can be translated into numbers of misclassifications per 1000 tests if the test positivity rate is known. We present this in Supplementary Results (Supplementary Results Table 1) for low- (5%), average- (20%) and high- (35%) test positivity rates in Bangladesh.

The candidate models are chosen as a result of a selection process and performed much better than more complex models (i.e. with 5 or more symptoms) or simpler models (with no symptoms but an intercept and age and gender as covariates) in terms of cross-entropy and ROC. For the models that used syndromic data, across all scenarios, within the final four candidate models the number of symptoms included made relatively little difference in terms of median performance (with respect to error, Figure 4.5 scenario-plot and Table 4.2), although the more complex models have higher precision.

Across all metrics, the rapid antigen test result is the most informative datatype for potential COVID-19 patients. However, incorporation of even one symptom and the use of a modelling framework greatly improves our ability to predict and classify cases, both generically and in specified scenarios. Including additional symptoms and covariates provides further information on the patient's status and greater model flexibility, resulting in higher precision in predictions and classifications.

4.5 Discussion

We have demonstrated that combining rapid antigen tests with syndromic modelling yields better identification of COVID-19 cases than either diagnostic in isolation. These gains in performance are mirrored across metrics of prediction, as well as general and scenario-specific classifications. The biggest improvement is seen under the scenario of "Epidemic Growth" (Table 4.2), and as expected following relaxation of restrictions and with the emergence of new variants. In this

scenario, the combined data model has a false negative rate of 18% (IQR: 21-15), 22 (IQR: 19-25) percentage points lower than the rapid-antigen-test-only model. Although the syndromic model matches the combined model's false negative rate, its false positive rate is 41% (IQR: 47- 37), 33 (IQR: 30- 33) percentage points higher. In real terms, at the end of our study, there was a 20% case positivity rate in Bangladesh. By applying our framework under the "Epidemic Growth" scenario, for every 100 rapid antigen tests, our approach would capture an additional 7 cases. In a country deploying millions of tests per week, this results in catching tens of thousands of cases that would otherwise be missed." Similarly, the combined model class performs equally well or better than the other models for the other scenarios explored (Figure 4.5). These scenarios offer snapshots of performance, while the model prediction and classification metrics provide an indication of how the models perform more generically (Figures 4.3 and 4.4, respectively). The more complex model classes achieve this top performance across all scenarios and metrics measured here thanks to their flexibility (allowing them to be readily adapted to new situations) and their synergistic use of the higher specificity rapid antigen testing and the more sensitive syndromic data.

The final symptoms and covariates chosen through model selection should be interpreted cautiously. Firstly, the power of the models to detect relationships will be partially determined by sample size. Secondly, these models were developed for prediction and classification in a unique sub-population: community support team (CST)-identified, symptomatic patients in low-income communities in Dhaka. From the same symptom and risk factor set, different variables were retained for different model classes, despite data being collected over a short period from the same population. These differences may point to mechanisms by which CST-identified and rapid antigen test positive individuals differ from other groups. They also underline the importance of collecting a relatively broad range of symptom data as the syndromic profile of the disease shifts between populations. Of interest is whether individuals identified by PCR but missed by rapid antigen tests are less infectious and more typical of asymptomatic cases (perhaps due to different lengths of time since symptom onset). This could be examined using viral load measured as Threshold Cycle (Ct) values from PCR and further testing for other illnesses.[25] Our use of PCR as a validation test should also be explored further, as it does not have 100% sensitivity so additional validation tests may be informative. However, finding alterantive gold-standard tests that can be carried out in the community is challenging.[26]

The modelling frameworks allow for the potential inclusion of additional covariates where they are collected reliably. These covariates may define different sub-populations in which we expect the relationships between symptoms and infectious status to differ. For example, vaccinated patients would be expected to exhibit fewer and milder symptoms than unvaccinated patients. By including vaccination status alongside symptoms within the model, the model can share information between the two groups while allowing the relationships to differ where this improves prediction. Similar approaches could be taken to incorporate rapid antigen test manufacturer, recent disease prevalence or time since symptom onset. Furthermore, using a modelling framework allows explicit estimation and exploration of these differences, rather than relying on *post hoc* analysis of misdiagnosis rates (for example, [27]). When a particular data source is found to have

good predictive power, it would be useful to identify whether this could target further data collection. For example, the low false-positive rate of rapid antigen tests means that, if affordable, serial-testing of the same individual could increase true positive detections without a major impact on accuracy.

The boost in diagnostic performance we found was achieved by harnessing data collected by community-based health workers using a mobile-phone based application to record patient symptoms and test results. These data were already being collected in Bangladesh and similar methods are being rolled out in other Low- and Middle-Income Countries.[28; 29] We ensured our method is scalable by developing it using a large community-based sample and with input from the CST program organisers. As CST data are collected via a mobile phone application the diagnostic model can be updated in real-time. The algorithm of the app could therefore be modified to reflect local epidemiological requirements, local case rates and the considered cost/benefits of misdiagnosis, thereby facilitating adaptation to new variants or even new diseases. Similarly, if a source of data becomes unavailable then the underlying model can be changed to reflect this. For example, if there are rapid antigen test supply problems, the app could deploy Syndromic-only which uses the same data as Syndromic-RAT Combined, without relying on the rapid antigen test, and the combined model could be retrained on tests from different manufacturers with different performance characteristics.

One of the key innovations of this framework is the ability to adapt the diagnostic to local populations and their needs. To achieve this, we need good quality, local data collection and to understand the costs of sensitivity and specificity. The costs of false negatives and false positives vary greatly depending on epidemic context, and balancing the treatment of individuals with control of the health burden at a societal level.[30] Similarly, the market price of interventions can fluctuate depending on demand, aid funding and global trends.[31] In practice, the costs of rapid antigen tests are likely to be up to an order of magnitude lower than PCR when considering the additional infrastructure and personnel. Access to testing (RAT or PCR) needs to be considered as part of weighing up the costs and benefits of surveillance approaches.[32] Understanding how to measure and balance these demands requires insights from economists, epidemiologists, social scientists and policy-makers, and is an area of active research. [33] Given the degree of complexity, it is tempting to rely on methods that do not openly require a decision to be made about the relative costs of the different misclassification types. However, rather than removing the complex cost structures involved, such methods simply hide them. All methods place a balance on false positives and negatives implicitly, our hope is that by requiring these decisions to be made explicitly, they are more readily challenged, researched and improved upon. Similarly, the need for local data collection should not be seen as a weakness of the method, but rather a welcome requirement that allows us to directly assess intervention success and biases.

Pandemic management can only be done with testing at scale. The combined syndromic and rapid antigen testing approach that we report is promising for large-scale COVID-19 testing in low-income communities. Moreover, our framework is adaptable, including for many other infectious diseases where strict adherence to gold-standard laboratory diagnostics greatly limits testing capacity. Imperfect diagnostics are frequently imperfect in different ways, and these dif-

ferences are ripe for statistical treatment. These methods are often more agile than gold-standard diagnostics in changing situations as experienced during the pandemic, when fast responses are essential. Overall, our approach shows that by understanding how to utilise the complementary strengths of imperfect but rapid diagnostics (and deploying the more limited gold-standard testing for validation), good quality large-scale testing can be achieved even in low-income communities.

4.6 Code Availability

The statistical code used in this study are available in a GitHub repository at https://github.com/fergusjchadwick/COVID19_SyndromicRAT_public.

References

[1] M. Dramé, M. T. Teguo, E. Proye, F. Hequet, M. Hentzien, L. Kanagaratnam, and L. Godaert, "Should rt-pcr be considered a gold standard in the diagnosis of covid-19?," *Journal of Medical Virology*, 2020.

[2] V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, *et al.*, "Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr," *Eurosurveillance*, vol. 25, no. 3, p. 2000045, 2020.

[3] R. Chowdhury, S. Luhar, N. Khan, S. R. Choudhury, I. Matin, and O. H. Franco, "Long-term strategies to control covid-19 in low and middle-income countries: an options overview of community-based, non-pharmacological interventions," *European Journal of Epidemiology*, vol. 35, no. 8, pp. 743–748, 2020.

[4] O. Vandenberg, D. Martiny, O. Rochas, A. van Belkum, and Z. Kozlakidis, "Considerations for diagnostic covid-19 tests," *Nature Reviews Microbiology*, vol. 19, no. 3, pp. 171–183, 2021.

[5] J. Dinnes, J. J. Deeks, S. Berhane, M. Taylor, A. Adriano, C. Davenport, S. Dittrich, D. Emperador, Y. Takwoingi, J. Cunningham, *et al.*, "Rapid, point-of-care antigen and molecular-based tests for diagnosis of sars-cov-2 infection," *Cochrane Database of Systematic Reviews*, no. 3, 2021.

[6] Y. Boum, K. N. Fai, B. Nikolay, A. B. Mboringong, L. M. Bebell, M. Ndifon, A. Abbah, R. Essaka, L. Eteki, F. Luquero, *et al.*, "Performance and operational feasibility of antigen and antibody rapid diagnostic tests for covid-19 in symptomatic and asymptomatic patients in cameroon: a clinical, prospective, diagnostic accuracy study," *The Lancet Infectious Diseases*, 2021.

[7] G. C. Mak, P. K. Cheng, S. S. Lau, K. K. Wong, C. Lau, E. T. Lam, R. C. Chan, and D. N. Tsang, "Evaluation of rapid antigen test for detection of sars-cov-2 virus," *Journal of Clinical Virology*, vol. 129, p. 104500, 2020.

[8] S. Muhi, N. Tayler, T. Hoang, S. A. Ballard, M. Graham, A. Rojek, J. C. Kwong, J. A. Trubiano, O. Smibert, G. Drewett, *et al.*, "Multi-site assessment of rapid, point-of-care antigen testing for the diagnosis of sars-cov-2 infection in a low-prevalence setting: A validation and implementation study," *The Lancet Regional Health-Western Pacific*, vol. 9, p. 100115, 2021.

[9] D. B. Larremore, B. Wilder, E. Lester, S. Shehata, J. M. Burke, J. A. Hay, M. Tambe, M. J. Mina, and R. Parker, "Test sensitivity is secondary to frequency and turnaround time for covid-19 screening," *Science Advances*, vol. 7, no. 1, p. eabd5393, 2021.

[10] Y.-H. Jin, L. Cai, Z.-S. Cheng, H. Cheng, T. Deng, Y.-P. Fan, C. Fang, D. Huang, L.-Q. Huang, Q. Huang, *et al.*, "A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-ncov) infected pneumonia (standard version)," *Military Medical Research*, vol. 7, no. 1, pp. 1–23, 2020.

[11] W. H. Organization *et al.*, "Who covid-19 case definition," tech. rep., World Health Organization, 2020.

[12] A. S. Maharaj, J. Parker, J. P. Hopkins, E. Gournis, I. I. Bogoch, B. Rader, C. M. Astley, N. Ivers, J. B. Hawkins, N. VanStone, *et al.*, "The effect of seasonal respiratory virus transmission on syndromic surveillance for covid-19 in ontario, canada," *The Lancet Infectious Diseases*, vol. 21, no. 5, pp. 593–594, 2021.

[13] T. Struyf, J. J. Deeks, J. Dinnes, Y. Takwoingi, C. Davenport, M. M. Leeflang, R. Spijker, L. Hooft, D. Emperador, J. Domen, *et al.*, "Signs and symptoms to determine if a patient presenting in primary care or hospital outpatient settings has covid-19," *Cochrane Database of Systematic Reviews*, no. 2, 2021.

[14] C. Menni, A. M. Valdes, M. B. Freidin, C. H. Sudre, L. H. Nguyen, D. A. Drew, S. Ganesh, T. Varsavsky, M. J. Cardoso, J. S. E.-S. Moustafa, *et al.*, "Real-time tracking of self-reported symptoms to predict potential covid-19," *Nature Medicine*, vol. 26, no. 7, pp. 1037–1040, 2020.

[15] S. Garry, N. Abdelmagid, L. Baxter, N. Roberts, O. l. P. de Waroux, S. Ismail, R. Ratnayake, C. Favas, E. Lewis, and F. Checchi, "Considerations for planning covid-19 treatment services in humanitarian responses," *Conflict and Health*, vol. 14, no. 1, pp. 1–11, 2020.

[16] S. M. Moghadas, M. C. Fitzpatrick, P. Sah, A. Pandey, A. Shoukat, B. H. Singer, and A. P. Galvani, "The implications of silent transmission for the control of covid-19 outbreaks," *Proceedings of the National Academy of Sciences*, vol. 117, no. 30, pp. 17513–17515, 2020.

[17] R. W. Peeling, P. L. Olliaro, D. I. Boeras, and N. Fongwen, "Scaling up covid-19 rapid antigen tests: promises and challenges," *The Lancet Infectious Diseases*, 2021.

[18] E. Surkova, V. Nikolayevskyy, and F. Drobniewski, "False-positive covid-19 results: hidden problems and costs," *The Lancet Respiratory Medicine*, vol. 8, no. 12, pp. 1167–1168, 2020.

[19] C. P. West, V. M. Montori, and P. Sampathkumar, "Covid-19 testing: the threat of false-negative results," in *Mayo Clinic Proceedings*, vol. 95, pp. 1127–1129, Elsevier, 2020.

[20] B. Healy, A. Khan, H. Metezai, I. Blyth, and H. Asad, "The impact of false positive covid-19 results in an area of low prevalence," *Clinical Medicine*, vol. 21, no. 1, p. e54, 2021.

[21] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[22] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of Statistical Software*, vol. 76, no. 1, 2017.

[23] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.

[24] J. Hasell, E. Mathieu, D. Beltekian, B. Macdonald, C. Giattino, E. Ortiz-Ospina, M. Roser, and H. Ritchie, "A cross-country database of covid-19 testing," *Scientific Data*, vol. 7, no. 1, pp. 1–7, 2020.

[25] E. Albert, I. Torres, F. Bueno, D. Huntley, E. Molla, M. A. Fernandez-Fuentes, M. Martínez, S. Poujois, L. Forqué, A. Valdivia, *et al.*, "Field evaluation of a rapid antigen test (panbio™ covid-19 ag rapid test device) for covid-19 diagnosis in primary healthcare centres," *Clinical Microbiology and Infection*, vol. 27, no. 3, pp. 472–e7, 2021.

[26] A. K. Giri and D. R. Rana, "Charting the challenges behind the testing of covid-19 in developing countries: Nepal as a case study," 2020.

[27] G. R. Babu, R. Sundaresan, S. Athreya, J. Akhtar, P. K. Pandey, P. S. Maroor, M. R. Padma, R. Lalitha, M. Shariff, L. Krishnappa, *et al.*, "The burden of active infection and anti-sars-cov-2 igg antibodies in the general population: Results from a statewide sentinel-based population survey in karnataka, india," *International Journal of Infectious Diseases*, vol. 108, pp. 27–36, 2021.

[28] A. B. Aziz, R. Raqib, W. A. Khan, M. Rahman, R. Haque, M. Alam, K. Zaman, and A. G. Ross, "Integrated control of covid-19 in resource poor countries," *International Journal of Infectious Diseases*, 2020.

[29] M. J. Schultz, T. H. Gebremariam, C. Park, L. Pisani, C. Sivakorn, S. Taran, A. Papali, *et al.*, "Pragmatic recommendations for the use of diagnostic testing and prognostic models in hospitalized patients with severe covid-19 in low-and middle-income countries," *The American Journal of Tropical Medicine and Hygiene*, vol. 104, no. 3 Suppl, p. 34, 2021.

[30] F. Toxvaerd and M. Makris, "Introduction: Economic contributions to infection control," *National Institute Economic Review*, vol. 257, p. 9–13, 2021.

[31] E. Mahase, "Covid-19: 120 million rapid tests pledged to low and middle income countries," 2020.

[32] C. Batista, P. Hotez, Y. B. Amor, J. H. Kim, D. Kaslow, B. Lall, O. Ergonul, J. P. Figueroa, M. Gursel, M. Hassanain, *et al.*, "The silent and dangerous inequity around access to covid-19 testing: A call to action," *EClinicalMedicine*, vol. 43, 2022.

[33] L. de Vries, M. Koopmans, A. Morton, and P. van Baal, "The economics of improving global infectious disease surveillance," *BMJ Global Health*, vol. 6, no. 9, p. e006597, 2021.

Chapter 5

Do identification guides hold the key to species misclassification by citizen scientists?

5.1 Abstract

1. Citizen science data often contain high levels of species misclassification that can bias inference and conservation decisions. Current approaches to address mislabelling rely on expert taxonomists validating every record. This approach makes intensive use of a scarce resource and reduces the role of the citizen scientist.

2. Species, however, are not confused at random. If two species appear more similar, it is probable they will be more easily confused than two highly distinctive species. Identification guides are intended to use these patterns to aid correct classification, but misclassifications still occur due to user-error and imperfect guidebook design. Statistical models should be able to exploit this non-randomness to learn confusion patterns from small validation data-sets provided by expert taxonomists, yielding a much-needed reduction in expert workload. Here, we use a variety of Bayesian hierarchical models to probabilistically classify species based on the species-label provided by the citizen scientist. We also explore the utility of guidebooks provided by the citizen science schemes as a prior for species similarity, and hence draw conclusions for their future improvement.

3. We find that the species-label assigned to a record by a citizen scientist, even when incorrect, contains useful information about the true species-identity. The citizen scientists correctly identify the species in around 58% of records. Using models trained on only 10% of these records (validated by experts), we can correctly predict species-identity for 69 (90%CI: 64-73)% of records when the guidebook is used, vs 64 (58-69)% for models that do not use the guidebook. The fact that misclassifications can be predicted systematically indicates that improvements could be made to the guidebook to reduce misclassification.

4. By using Bayesian, hierarchical models we can greatly reduce the workload for experts by providing a probabilistic correction to citizen science records, rather than requiring manual review. This is increasingly important as the number of citizen science schemes grows and the relative number of taxonomists shrinks. By learning confusion patterns statistically, we open up future avenues of research to identify what causes these confusions and how to better address them.

5.2 Introduction

Citizen science has become increasingly important for addressing modern ecological questions. Systematic methods for monitoring biodiversity can rarely achieve the same scale in space or time. Simultaneously, these projects empower members of the public to take an active interest in the natural world and its stewardship [1]. However, these data are challenging to analyse, with heterogeneous recording effort across space, time, and taxa, and frequent misclassification of species [2]. A large number of models have been developed to tackle heterogeneous effort problems in citizen science data [3; 4; 5], however, fewer attempts have been made to address species misclassification [6; 7].

Misclassifications can affect biological inference. For example, in a survey aimed at describing the habitat associations of a focal species, mislabelling of other species as the focal species can increase both bias and uncertainty in the perceived habitat usage. If the habitat usage by the two species overlaps, this may be small, however, if species misclassifications lead to false positive records the resulting inference will be biased [8; 9]. Schemes that acknowledge these problems tend to tackle species misclassification via labour-intensive manual review of each record by experts [10]. Unfortunately, expert reviewers are few [11] and citizen science records are many [12]. We therefore require solutions to the problem that make efficient use of small amounts of expertly reviewed data.

Fortunately, most species are not confused at random [10]. Species are most likely to be confused if they share similar physical traits, such as coloration, size or behaviours. Such non-randomness is amenable to statistical treatment via the development of suitable observation models. Statistically modelling the observation process allows us to learn which species are confused and to make probabilistic reclassifications. These misclassification patterns can be learned from expert-validated data alone [13] but to reduce the workload for experts we can incorporate prior knowledge about species similarity. Crucially, we can incorporate prior knowledge about species similarity *from the perspective of the citizen scientists*. The features that make two species look similar to a citizen scientist, for whom species identification may be a new experience, are likely to differ from those used by experts.

Fortunately, we have expert knowledge on what citizen scientists see codified in identification guides. These guides are developed by scheme organisers and use simple, easy-to-learn features to help citizen scientists distinguish species. Scheme organisers often incorporate knowledge from citizen science workshops and previous schemes when designing guides [14]. This may lead to the traits used in different guides varying for the same species group, however, most guides are characterised by minimal structure (i.e. they do not use keys or only use very coarse keys) and the traits selected tend to be easy to find and distinguish without previous experience. Translating guidebooks into formal priors is challenging due to this lack of imposed structure, as it is unlikely that the traits are weighted equally and citizen scientists will develop heuristic hierarchies of traits and combinations thereof.

Overcoming this challenge and including guidebooks in misclassification models should also allow us to assess guidebook design. Assuming the species are distinguishable, if the models do not find signal in which species are confused

(with or without the guidebook prior), then the guidebook (or some other form of training) is performing well and misclassifications are down to user error alone. If the models do find a signal in which species are confused, then there is structure in the misclassifications not currently addressed by the guidebooks. If the guidebook prior is uninformative, it is not capturing features that lead to confusions. This may arise from the citizen scientists using traits that are not included in the guidebook or making errors that do not correspond to conventional traits, for instance, labelling the record as a species with a similarly spelled name or some species being culturally more important, for example, if they are rare or invasive [15]. If the guidebook prior is informative, it is generating confusions by making species appear overly similar. This may arise from including too many traits and placing insufficient emphasis on the most informative traits.

In this paper, we present a series of observation models that address species misclassifications explicitly. These models include multiple approaches to incorporating guidebooks as priors (including not incorporating the guidebook). The prediction performance of these models is tested under cross-validation with different amounts of training data using both simulated and real-world data from the "Blooms for Bees" citizen science scheme [10]. We discuss the implications of the results for guidebook design and suggest extensions to these models.

5.3 Materials and Methods

5.3.1 Modelling Problem

For a data set of N records we consider two N-length vectors: the record-labels (the species label chosen by the citizen scientist), $\boldsymbol{U}$; and the record-identities (the species that a record truly belongs to), $\boldsymbol{T}$. The elements of both vectors take values from a set of M levels representing the possible species.

$$T_n, U_n \in \{\text{Species 1, ... , Species M}\}$$

We construct our model generatively, beginning with the underlying biological process that determines the relative frequency of each species (the true record-identities). We assume this process is independent of our observation process, is proportionate to the species abundance and will often be the pattern we are interested in reconstructing from our data. Even if we are not interested in the biological process, by jointly modelling it with the observation process we can potentially improve our estimation of the species misclassifications [7].

In principle, any biological process model could be used. To simplify notation, we will assume the mth species' relative frequency, α_m, is a function of K environmental covariates (X, e.g. distribution of temperature, precipitation and wind): $\alpha_m = f\left(\sum_{k=0}^{K} \beta_{k,m} X_{k,m}\right)$

Since the true record-identities, T_n, are non-ordinal, the categorical (or single-

trial multinomial) distribution, Cat, is a natural choice for the likelihood of any given record. The categorical distribution is parameterised with a vector of probabilities, $\boldsymbol{A}$, corresponding to the probability that the record truly belongs to each potential species. We therefore transform our unbounded linear predictor, $\boldsymbol{\alpha}$, into a simplex using the softmax link function (a multivariate generalisation of the logit transformation, also known as the "multi-logit").

$$
\begin{aligned}
T_n &\sim \text{Cat}(\boldsymbol{A}) \\
\boldsymbol{A} : A_m &= \frac{e^{\alpha_m}}{\sum_{i=1}^{M} e^{\alpha_i}}
\end{aligned}
$$

Now that we have a model for generating the true species identities, we need to link these identities to the labels assigned by the citizen scientists (i.e. the observation process model). We want to estimate the probability that a given record-label, U_n, is generated conditional on the underlying record-identity, T_n, Here, too, we will use a categorical distribution parameterised by the rows of an $M \times M$ matrix, $\boldsymbol{C}$ that corresponds to the record-identity, T_n. This matrix encodes the pairwise confusability of T_n with each potential U_n, i.e., the generation of U_n from T_n. The formulation of $\boldsymbol{C}_{T_n}$ could take multiple forms depending on the amount of trust placed in the citizen scientist, whether external information is incorporated, and, if so, how that information is incorporated. Here, we describe the generic structure using the shorthand $\boldsymbol{f}$ for the different modelling frameworks, with $\boldsymbol{f}$ indicating the vector of classification probabilities conditional on the true identity of the record, and f_m showing the m^{th} element of that vector. Below, we describe the different variants of the modelling framework in more detail (see also Figure 5.1)

$$
\begin{aligned}
U_n &\sim \text{Cat}(\boldsymbol{f}(\boldsymbol{C}_{T_n})) \\
P(U_n = \text{Species } m | T_n) &= f_m(\boldsymbol{C}_{T_n})
\end{aligned}
$$

5.3.2 Model 0: Trust User

Our null "model" assumes that the citizen scientists are 100% correct in their labels (i.e. the label always matches the species identity). This model is extreme since even experts are rarely 100% correct but, implicitly, this is the model assumed by any analysis of citizen science scheme that uses the data without correction or mediation via calibration datasets and strong priors.

$$
\begin{aligned}
P(U_n = i | T_n = j) &= \delta_{i,j} \\
\text{where } \delta_{i,j} &= \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

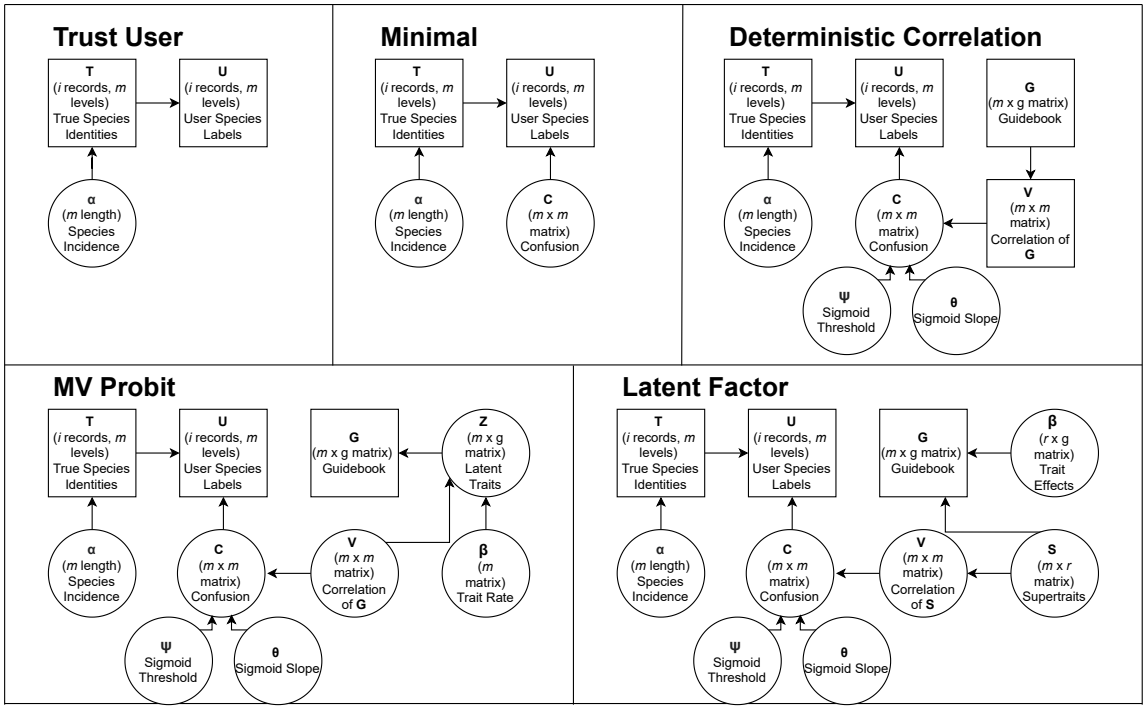

Figure 5.1: **Causal Structures of Candidate Models** We define five modelling frameworks of different degrees of complexity. All of the frameworks contain a "Species Incidence" term, $\boldsymbol{\alpha}$, which corresponds to the function that allows the estimation of (and adjustment for) relative species abundance. The "Trust User" framework assumes that the record-label matches the record-identity. The "Minimal" framework incorporates an unstructured confusion matrix, $\boldsymbol{C}$, allowing the record-label to differ from the record-identity. The "Deterministic Correlation", "MV Probit" and "Latent Factor" frameworks all use the citizen science scheme's guidebook, $\boldsymbol{G}$, to estimate correlations between species, $\boldsymbol{V}$, to inform $\boldsymbol{C}$. The "Deterministic Correlation" uses the empirical correlation between the species in the guidebook as data to inform $\boldsymbol{C}$. The "MV Probit" framework estimates the correlation between the species in the guidebook using a multivariate - probit model. These two approaches weight the guidebook traits equally. The final framework, the "Latent Factor" approach, is the most flexible, using latent factors, $\boldsymbol{S}$, to combine and reweight the traits. All the approaches which use the guidebook are subject to a flexible sigmoidal transformation using a Normal CDF parameterised by θ and ψ

5.3.3 Model 1: Minimal

The simplest model (conceptually) is to populate the elements in $\boldsymbol{C}$ with parameters that allow the citizen scientists to confuse species (i.e. suggest the wrong label conditional on the species' identity). If we do not know *a priori* which species the citizen scientist is likely to confuse, we can use free parameters to populate $\boldsymbol{C}$. The elements of $\boldsymbol{C}$ can be drawn independently and the rows used to parameterise the categorical distribution under softmax transformation. To make the model identifiable, we must fix one parameter in each row (i.e. in each vector passed to the categorical distribution). Here, we fix the correct classification (the diagonal of $\boldsymbol{C}$) to one to maintain consistency with our later models that use a correlation structure (and thus also have ones on the diagonal). As we expect correct classification to be more likely than any given misclassification, we centre the off diagonal values around zero. The spread of misclassification values is determined by a standard half-Normal prior on σ. The larger σ is, the greater the spread in probabilities of misclassification. A smaller σ indicates most misclassifications are equally likely to each other, but much less likely than correct classifications. It is plausible that σ could vary by species so we examined two versions of the model: one with a global σ parameter (as below) and a more flexible one where it is indexed by species, σ_i.

$$
\begin{aligned}
U_n &\sim \text{Cat}(\boldsymbol{f}(\boldsymbol{C}_{T_n,.})) \\
P(U_n = \text{Species m}|T_n) = f_m(\boldsymbol{C}_{T_n,.}) &= \frac{e^{\boldsymbol{C}_{T_n m}}}{\sum_{i=1}^{M} e^{\boldsymbol{C}_{T_n i}}} \\
C_{i,j} &\sim N(0, \sigma^2) \\
\sigma &\sim \text{half-}N(0, 1) \\
C_{i,i} &= 1
\end{aligned}
$$

5.3.4 Estimating Species Similarity from ID Guides

Species identification by citizen scientists is most commonly done visually and we would expect similar-looking species to be more readily confused than highly distinctive-looking species. However, the question then arises: how do we know which species look similar to the citizen scientists? One source of information on this could be the identification guides provided to the citizen scientists by the scheme-organisers. These guides are often very different from professional taxonomic guides where subtle features and highly structured keys are relied upon. Citizen science guides are characterised by easy-to-recognise features and limited structure. Many guides are thus easily converted into simple $M \times H$ binary trait matrices, $\boldsymbol{G}$, with each row corresponding to a species, each column the level of a trait and a binary indicator in each cell indicating whether the indexed species has the indexed trait.

The distance between species in $\boldsymbol{G}$-space, $\boldsymbol{V}$, therefore, represents our prior expectation of which species are likely to be confused and can be used to inform our $\boldsymbol{C}$ matrix. There are multiple options for defining this distance depending

on the amount of flexibility given to estimating the correlation in $\boldsymbol{G}$-space and the weighting of the different dimensions (i.e. the different traits). The simplest of these methods, the "Deterministic Correlation" model described below, is the only method to use the empirical correlation of species in $\boldsymbol{G}$-space as a measure of distance, making the least flexible use of the guidebook data. The "MV Probit" method relaxes this approach by estimating the correlation of species in $\boldsymbol{G}$-space as a measure of distance by means of a multivariate probit. Both of these methods apply equal weighting to the trait dimensions in $\boldsymbol{G}$-space, but in reality it is unlikely that all traits are equally important to the citizen scientists for species determination. For example, some traits, like colour, may be easier to assess without specialist knowledge and be relied on more heavily. They may also be viewed in combination, so while head colour, thorax colour and abdomen colour are separate traits in the guidebook, many species have the same colour on multiple body parts and will be thought of as "the ginger bee" (like many of the carder bee species). Our final method, the "Latent Factor" model, accounts for this behaviour by allowing the traits in $\boldsymbol{G}$-space to be up- or down-weighted and recombined using latent factors (or "supertraits"). The correlation between species in latent factor space is then used to inform $\boldsymbol{C}$.

The correlation distance between species, regardless of how it is estimated, does not necessarily map directly onto the confusability distance in the probability space. While we would generally expect the confusability ranking to be maintained, the scalar distance in the two spaces will likely differ. We, therefore, introduce a convolution step, wherein we apply a flexible sigmoidal transformation to each vector of correlational distances. Specifically, we use the Normal CDF function which has two parameters, a slope (θ) and threshold (ψ). As the slope approaches zero, the sigmoid becomes a step function, with values smaller than the threshold transformed to zero and larger values become ones. The threshold determines where this step occurs. As the slope gets larger, the values around the threshold will become values between zero and one, with only more extreme values becoming zeroes or ones. We expect a small number of species to be confused with the true species, so place a prior having a mid-high threshold and small slope. Similarly to the σ parameter in the "Minimal" model, it is plausible that θ and ψ could vary by species so we examined two versions of the model: one with global values for those parameters parameter (as below) and a more flexible one where they are indexed by species (θ_{T_n}, and ψ_{T_n}).

$$
\begin{aligned}
V'_{T_n,m} &= \frac{1}{2}\left[1 + \text{erf}\left(\frac{V_{T_n,m} - \psi}{\theta\sqrt{2}}\right)\right] \\
\frac{\psi}{2} + 0.5 &\sim Beta(5, 2) \\
\theta &\sim \text{half-}N(0, 0.2^2)
\end{aligned}
$$

We also need to allow for confusion not associated with the guidebook. We achieve this by incorporating $V'_{i,j}$ as a prior for $C_{i,j}$. As $V'_{i,j}$ is bounded by zero and one, we use a Beta distribution with a mean-variance parameterisation. This parameterisation allows the variance, λ, to up or down-weight the contribution of $V'_{i,j}$ to $C_{i,j}$.

$$C_{T_n,m} = Beta\left(\lambda \cdot V'_{T_n,m}, \lambda \cdot \left(1 - V'_{T_n,m}\right)\right)$$

Although $C_{i,j}$ is bound by zero and one, the vector still needs to be normalised to generate a simplex of confusions. One option would be to use the softmax transformation (as in the Minimal Model), however, with constrained values (unlike in the Minimal Model) softmax tends to generate a large number of small probabilities. This contradicts our understanding of confusions which we expect to be sparse, with a few large probabilities (commonly confused species) and many zero probabilities (species which are never confused). Fortunately, as the values are now all positive, we can simply normalise by dividing each element of the vector by the sum of the vector. This is easy to calculate and is compatible with sparse probabilities.

$$C'_{T_n,m} = \frac{C_{T_n,m}}{\sum_{i=1}^{M} C_{T_n,i}}$$

These normalised values can now be used to parameterise the categorical distribution as in the Minimal Model:

$$U_n \sim \text{Cat}(\boldsymbol{f}(\boldsymbol{C}_{T_n,.}))$$

5.3.5 Model 2: Deterministic Correlation

There are several numerical methods for calculating the empirical correlation of a dataset. These point estimates for correlations, e.g. Pearson's, Kendall's and Spearman's coefficients assume no uncertainty in the correlations but are very computationally efficient, and thus useful as a baseline guidebook-based model. We use the Pearson's correlation coefficients for the species in the guidebook as data, $\boldsymbol{V}$.

$$V_{i,j} = \frac{cov(\boldsymbol{G}_{i,.}, \boldsymbol{G}_{j,.})}{\sigma_{\boldsymbol{G}_{i,.}} \sigma_{\boldsymbol{G}_{j,.}}}$$

5.3.6 Model 3: Multivariate Probit

The next most complex version of the model allows the uncertainty in the guide-based correlation, $\boldsymbol{V}$, to be estimated. Since we have binary data, a natural method for doing this is the multivariate probit, which assumes that binary variables are realisations of correlated, normally distributed latent processes [16; 17]. This assumption is both computationally expedient and often matches our biological knowledge, since many binary variables are functions of continuous underlying processes. For example, an individual is considered infected or not infected (binary) based on an underlying quantity of pathogens (continuous).

The binary data, $\boldsymbol{G}$ formed of H traits and M columns, are linked to the latent continuous variable, $\boldsymbol{z}$, by means of a thresholding function, $\mathbb{I}$, which returns a one if the latent variable is positive and a zero if it is negative. This thresholding process is equivalent to a probit link.

$$G_{h,m} = \mathbb{I}(z_{h,m} > 0)$$

The latent state, $\boldsymbol{z}$, is generated from an M length intercept-only linear predictor, $\boldsymbol{\beta}$ and a multivariate-Normal (MVN) error term, $\boldsymbol{\epsilon}$. The correlation between species is induced through the normalised covariance matrix parameter of the MVN.

$$\begin{aligned} z_{h,m} &= \beta_m + \epsilon_{h,m} \\ \boldsymbol{\epsilon_h} &\sim MVN(\mathbf{0}, \boldsymbol{V}) \\ V_{ii} &= 1 \end{aligned}$$

The intercept corresponds to the number of traits each species has (i.e. the number of traits indicated by a 1 in the guide). There is little biological interpretation for this value (as it is determined by encoding decisions), so we use a standard Normal prior which is minimally informative under a probit transformation.

$$\boldsymbol{\beta} \sim N(0, 1)$$

We place an LKJ prior on the correlation prior with $\eta = 1$. The LKJ prior corresponds to beta-distributed marginal correlations of Beta $\left(\frac{\eta+M-2}{2}, \frac{\eta+M-2}{2}\right)$. At $\eta = 1$ this is relatively uniform at small values of M but with slight peaking at 0 correlations for larger values of M. Lower marginal correlation between species as they increase is plausible although the degree of shrinkage should be monitored when re-applying. There are relatively few priors for correlation matrices and the LKJ distribution is computationally efficient.

$$\boldsymbol{V} \sim \text{LKJ}(1)$$

5.3.7 Model 4: Latent Factor Models

The approaches thus far all treat traits as equally important, however, this is unlikely to hold in reality. Field biologists (particularly ornithologists) have long referred to the "jizz" or "vibe" of an organism: the combination of shape, mode of movement, posture, colouration and myriad subtle traits which allow an organism to be identified from a quick glance. In these scenarios identification is not occurring on a trait-by-trait basis but some reading of the whole or of groups of traits together. While non-expert users may not achieve identification at a glance, it is likely they process the guidebooks and identification process in a similar way. Some species will be discounted immediately based on the dominant colour (a combination of thorax, abdomen, and tail colour), for instance, while other species will be more challenging to disentangle. The development

of citizen science identification guides is often an attempt to formalise this process by using measurable traits. Guides may be able to capture this with expert construction, but the need to function for novice citizen scientists as well as more experienced observers means that there will often need to be an imperfect match between the guide layout and its use by groups of different experience levels. As a result, to understand guide-based confusion we need a system by which we can up- and down-weight the contribution of different traits in the guidebook.

One approach is to consider the traits in $\boldsymbol{G}$ as functions of latent factors, $\boldsymbol{S}$. These latent factors can be thought of as "super-traits": continuous underlying processes that when combined in different proportions give rise to the traits that are measured and included in the guide. Crucially, the correlation between latent factors represent the similarity between species as seen by the citizen scientists, so correlations between species in $\boldsymbol{S}$ give us $\boldsymbol{V}$. The number of latent factors, R, can be estimated or may be pre-determined based on the size of data/previous experiments (or the latter used to inform a prior for estimation).

Estimating Latent Factors from Traits Most traits are binary so here we link traits, $\boldsymbol{G}$, and latent factors, $\boldsymbol{S}$, using logistic regression. The latent factors need to be flexible but estimable. For this reason, we assume a linear relationship between traits and their latent factors and provide standard Normal priors which are relatively uninformative under logit transformation. Where more information is known about the latent factors, more complex functional forms could be used. Similarly, where non-binary traits are present the approach can be generalised to accommodate more complex traits using other GLM formulations.

$$\begin{aligned} \boldsymbol{G}_{m,h} &\sim \text{Bern}(p_{m,h}) \\ \ln\left(\frac{p_{m,h}}{1-p_{m,h}}\right) &= \beta_{h,0} + \sum_{r=1}^{R} S_{m,r}\beta_{h,r} \\ S_{m,r} &\sim N(0,1) \end{aligned}$$

As the latent factors have exchangeable priors, we risk label-switching identifiability issues (i.e. the indexing of the latent factors may be inconsistent between MCMC chains). We need to impose some form of order on $\boldsymbol{S}$, however, as we are going to estimate the correlations between the rows of $\boldsymbol{S}$ (i.e. between the species in super-trait space) ordering the latent factors is undesirable. For this reason, we place a hierarchical prior on $\boldsymbol{\beta}$.

$$\begin{aligned} \beta_{m,r} &\sim N(0,\tau_r) \\ \tau_r &\sim \text{half-}N(0,0.5) \\ \tau_1 \leq \tau_2 \leq ... \leq \tau_{R-1} &\leq \tau_R \end{aligned}$$

Even with these restrictions, $\boldsymbol{S}$ and $\boldsymbol{\beta}$ are only identifiable when linked to another data source (in our case, the identity-label confusions).

Linking Latent Factors to Species Confusions through $\boldsymbol{V}$ To link $\boldsymbol{S}$ to $\boldsymbol{V}$ we simply calculate the correlations between the rows of$\boldsymbol{S}$ (i.e. between the species

in latent factor space). To do this, we must first z-score normalise the rows of $\boldsymbol{S}$ to achieve a mean of zero and variance of 1 to give us $\hat{\boldsymbol{S}}$. We can then calculate the Pearson correlation of $\boldsymbol{S}$ by post-multiplying $\hat{\boldsymbol{S}}$ by it's transpose to give us $\boldsymbol{V}$.

$$\hat{S}_{m,r} = \frac{S_{m,r} - \overline{S_{.,r}}}{\sigma_{S_{.,r}}}$$
$$\boldsymbol{V} = \frac{\hat{\boldsymbol{S}} \cdot \hat{\boldsymbol{S}}^T}{R-1}$$

5.3.8 Measuring Performance

The aim of these models is to reduce the work required by expert taxonomist reviewers to processing a small validation data set to which the model can be fit, while propagating and stating the uncertainty in corrected classifications. The best model, therefore, is the one that has highest out-of-sample predictive power from the smallest training data set. In this section, we define how we measure out-of-sample predictive performance using the correct classification rate and our experimental design for comparing performance across different sample sizes.

First, since the same performance metrics are measured across varying models we will use the following summary notation to represent all the models (including those that incorporate $\boldsymbol{V}$):

$$T_n \sim \text{Cat}(\boldsymbol{A}) \therefore \mathbb{P}(T_n = \text{Species } i) = A_i$$
$$U_n \sim \text{Cat}(\boldsymbol{f}(\boldsymbol{C}_{T_{n,\cdot}})) \therefore \mathbb{P}(U_n = \text{Species } j) = f_j(\boldsymbol{C}_{T_{n,\cdot}})$$

If we then want to predict T_n given U_n we apply Bayes rule:

$$\mathbb{P}(T_n = \text{Species } j | U_n = \text{Species } i) = \frac{\mathbb{P}(U_n = \text{Sp. } i | T_n = \text{Sp. } j)\mathbb{P}(T_n = \text{Sp. } j)}{\sum_{j=1}^{M} \mathbb{P}(U_n = \text{Sp. } i | T_n = \text{Sp. } j)\mathbb{P}(T_n = \text{Sp. } j)}$$

We will represent the $M \times M$ matrix that defines all the possible combinations of U and T using the symbol $\boldsymbol{\Psi}$. Each row of $\boldsymbol{\Psi}$ corresponds to a label and defines a simplex (probabilities summing to 1) which give the probability of the possible record-identities, T.

The simplest way to think about model performance is to measure how often the record-identity predicted by the model matches the true record-identity as validated by the model, i.e. the correct prediction rate, $\mathfrak{R}$. To estimate this value, we generate prediction values, $\hat{T}_n$, for each record in a holdout set of size N, and measure the proportion of records for which $\hat{T}_n = T_n$

$$\hat{T}_n \sim \text{Cat}(\mathbf{\Psi}_{U_n,\cdot})$$

$$\mathfrak{R} = \frac{\sum_{n=1}^{N} \begin{cases} 1 & \text{if } \hat{T}_n = T_n \\ 0, & \text{otherwise} \end{cases}}{N}$$

5.3.9 Comparing Performance Under Cross-Validation and Varying Data Richness

As outlined above, models need to be evaluated on both predictive performance under varying data richness (i.e., what proportion of the available data is used in training the model). We therefore use holdout cross-validation, where the data available are randomly assigned to two groups, d_0 (training) and d_1 (testing). The assignment is repeated J-times to estimate the average and spread of the performance metrics. Varying data richness (the sizes of d_0 and d_1) introduces two sources of uncertainty associated with the training and testing sets of data. As either the training or testing set shrinks, the number of data-point combinations will increase, leading to higher variability in the fitting process and prediction targets that will both be reflected in the performance metrics.

We are not interested in uncertainty due to testing set size therefore the simplest solution to this source of variation is to fix the size of d_1. Naturally, the size of available testing data is complementary to the amount of available training data. We therefore chose to test on 25% of the available data, as requiring more than 75% of the available data to be used in training would not represent a significant reduction in the work of the expert validators.

Varying the size of d_0 introduces two sources of variability. Firstly, there is the larger uncertainty in parameter estimates associated with smaller data sizes. Secondly, there is the larger number of combinations of data that may be used in the training data. The first is of vital importance to understanding model performance while the second is a nuisance that we should control for. Unfortunately, it is hard to predict what impact the latter will have, so we have to take a computationally-intensive approach.

We start by choosing a large value of J and running holdout cross-validation for all the model classes at the smallest d_0 of interest. The smallest size of d_0 will have the largest performance-metric variability due to training set effects (fortunately, they will also be the quickest models to run). We then repeatedly sub-sample from 1 to J of the cross-validation exercises and assess at what value, J', the centre and spread of the performance-metrics stabilises. The larger sizes of d_0 can then be run only J' times and the variability therein can be attributed solely to uncertainty in parameter estimation.

5.3.10 Case Study

We apply our modelling framework to real-world data from the "Blooms for Bees" citizen science program. In this program, citizen scientists were asked

to photograph and identify to species-level every visiting bumblebee to a single plant with at least one open flower in their garden or allotment. The scheme provided an unstructured identification guide to participants via a mobile phone app (through which they also submitted their records). The guide has 23 bumblebee species and 69 traits (including levels of traits). Falk *et al* then reviewed the photographs and corrected any misclassified species [10]. This generated 2314 records containing the original label from the citizen scientist and the corrected identity provided by the expert reviewer. The records are primarily concentrated around the West Midlands of England as the program was developed by Coventry University. This restricted geographic region, and focus on garden and allotment habitats allowed us to adopt a very simple biological process model using an intercept-only linear predictor for each species, α, corresponding to their relative abundance. The performance of the observation models is compared using the cross-validation protocol above.

5.3.11 Simulation Study

In some instances, we may not know exactly which guidebook the citizen scientists used. For example, participants may supplement the guidebook provided by the scheme with their own favourite guidebook. It is therefore important to understand how sensitive our models are to the exact guidebook used. We can explore this using simulations. Guidebooks for the same group of species may differ in a large number of ways - the rank order of species similarities, the frequency of the traits used, the determination of the correlations between species (i.e., how strongly correlated the species are in the trait-space defined by the guidebook). To assess the sensitivity of the models to these changes, we need to simulate under one guidebook scenario and then compare prediction performance when the model is fit with the correct guidebook vs a contrasting one. These kinds of transplant tests are computationally intensive, especially when testing under the cross-validation conditions described above.

To make these simulations computationally tractable, we need to prioritise how we simulate and change the guidebook. First, we choose one of our observation models to be the basis of the simulation. The "Multivariate Probit" model is the simplest model that allows us to generate a full guidebook. In this model (Figure 5.1), we can change the rank ordering of species similarities by re-organising the columns of the correlation matrix, the frequency of the traits using the mean β parameter, and the determinant of correlations by modifying the prior on $\boldsymbol{V}$. This brings us to the second prioritisation: which of these to change. The guidebook-space defined by these parameters is huge and not practical to explore fully. We choose to focus only on changing the rank order of species similarities as this is a commonly discussed decision by guidebook designers (e.g., when navigating a dichotomous key designers often try to ensure the final pair of species in each branch are easy to distinguish). Finally, as the "Multivariate Probit" model is stochastic, we need to repeat simulations to account for the inherent noise in the generative process. For this reason, we limit our comparison to two contrasting scenarios (i.e., two species rankings) to facilitate more repetitions.

In order to make our simulations realistic, we draw the parameters for each scenario from the posterior of the "Multivariate Probit" fit to the real data. To

change the species rankings in a consistent way, we modify the correlation parameter using hierarchical clustering with the complete linkage algorithm implemented using the "hclust" function in the R "stats" package [18; 19]. The precise nature of the reordering is not significant, but it is worth noting that by only reordering the correlation matrix we keep the same matrix determinant.

We now have full parameters for two contrasting scenarios (Figure 5.2). The "Real" scenario has the same rank species similarity as our real data while the "Restructured" scenario has a different rank species similarity

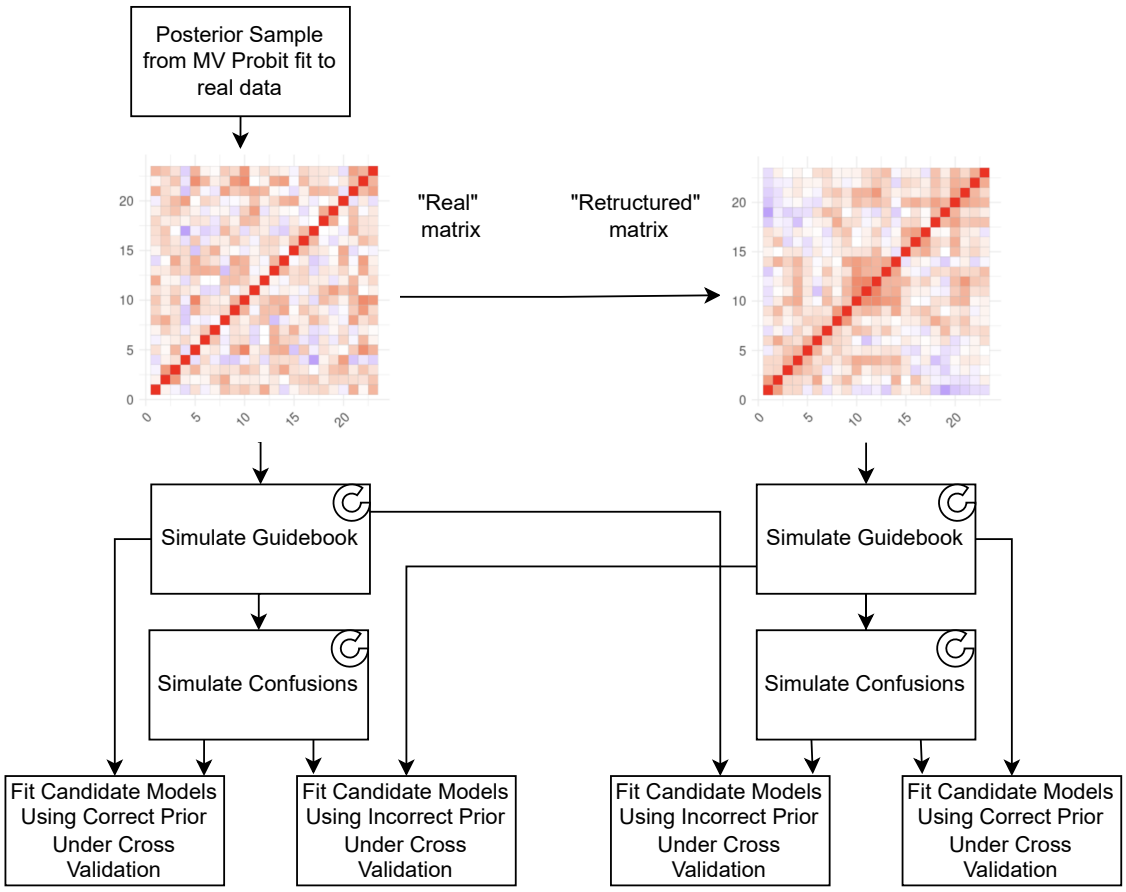

Figure 5.2: **Simulation study outline.** The simulation scenarios are based on species correlations estimated using the real data. This correlation matrix ("real", with a blue to red scale indicating negative to positive correlations) is used to generate one set of simulations directly, and restructured to generate a contrasting correlation matrix ("restructured") from which distinct but comparable simulations are created. Each of our candidate models is then fit (under cross validation) to the simulated data using either the correct or contrasting guidebook as a prior. This allows us to assess how sensitive to the guidebook the models are. Steps which are repeated are indicated with a partial concentric ellipse.

We take five samples from the posterior of the "Multivariate Probit" model fit to the real data. From these samples, we generate five guidebooks and five corresponding data sets (i.e., vectors of species labels and species identitities). To these data, we fit 7 candidate models (Minimal, Deterministic Correlation with correct and incorrect prior, Multivariate Probit with correct and incorrect prior and

Latent Factor with correct and incorrect prior) under the same cross-validation scheme described for the real data.

5.4 Results

5.4.1 Computational Resources

Analyses were run in R (v4.2.1) [18] with the CmdStanR (v0.5.3) [20] interface to Stan (v2.30.1) [21] on a 64-bit workstation with 32 AMD Ryzen Threadripper 3970X CPUs running a Ubuntu 20.04.5 LTS operating system.

5.4.2 Model Convergence

All models achieved convergence with $\hat{R} < 1.1$ [22], $< 2\%$ divergent transitions [23], and effective sample sizes of over 100 samples/chain (with the exception of a small number of lower level parameters in the "Latent Factor" model) [24]. Model runtimes varied by model type and data richness, with the more complex models and larger datasets taking longer to run (see Figure 5.3) with runtimes ranging from a minute to just over 3 hours.

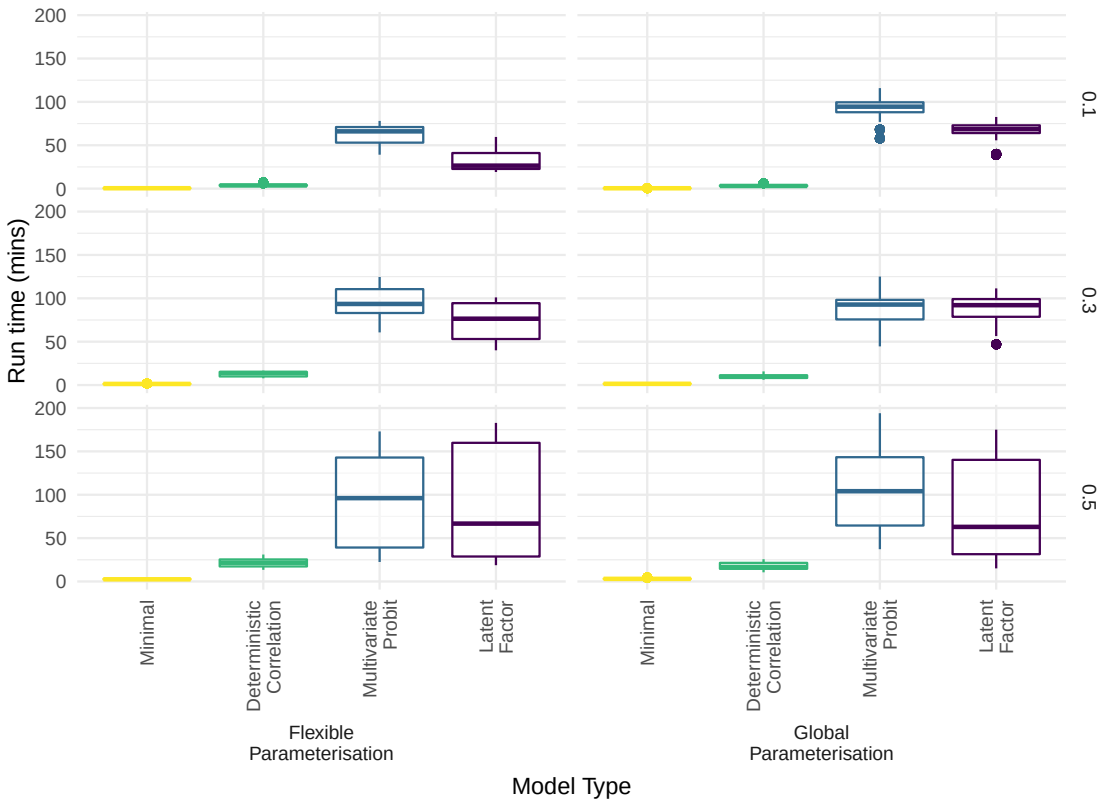

Figure 5.3: **Comparison of model run times**. Median and interquartile ranges for run times across cross-validation model fitting. Each row panel corresponds to different levels of data richness. As model complexity and data richness increases, models take longer. Run times vary from minutes to a few hours.

5.4.3 Case Study

When tested against the real data at the lowest level of data richness (10%), the best median performance of models which used the guidebook exceeded the "Trust User" model by around 10% and the "Minimal" models by 5%. The best performing parameterisation of the "Minimal" model predicted the proportion of correct classifications was 0.64 (90% Credible Interval (CI): 0.58-0.69), compared to the best performing parameterisations of the "Deterministic Correlation", 0.67 (0.62-0.71), the "Multivariate Probit", 0.69 (0.64-0.73), and "Latent Factor", 0.69 (0.64-0.73) models.

With higher data richness, the differences in model performance shrink and performance quickly plateaus. Indeed, between data richnesses of 30% and 50%, the best performing parameterisations of the models have the same median performance, with only small improvements in precision (if any). The "Minimal" model achieves a correct classification rate of 0.68 (0.64-0.71) at 30% with the 90% CI shrinking to 0.65-0.71 at 50%. These rates are very close to those achieved by the guide-based models across which there is almost no difference in performance. The "Deterministic Correlation" model achieves the same rate of 0.7 (0.67-0.73) at 30% and 50% data richness. The "Multivariate Probit" yields 0.7 (0.66-0.73) at 30% with the credible interval shrinking slightly to 0.66-0.72 at 50% data richness. The "Latent Factor" performs identically at both levels of data richness, with rates of 0.69 (0.66-0.72).

The flexibility of the parameterisations tested generally made little difference in performance except at low data richness. For each model, we tested two parameterisations and compared their 50% credible intervals (more sensitive to differences than the more conservative 90% CIs used for between-model comparisons). As shown in Figure 5.4, the less flexible parameterisation of the "Minimal" model performed best at 10% data richness (0.64 (0.62-0.66) vs 0.62 (0.6-0.64)), with no difference at higher levels of data richness. In contrast, the more flexible parameterisation of the "Deterministic Correlation" model performed slightly better at all levels of data richness. The "Multivariate Probit" and "Latent Factor" models seemed less affected by parameterisation (although the less flexible parameterisation of the "Latent Factor" model had much larger uncertainty in the tails at 10% data richness than the corresponding flexible parameterisation). Based on these results, the best model fits to the simulated data are the less flexible parameterisation of the "Minimal" model and the flexible parameterisation of the other models.

5.4.4 Simulation Study

The difference in performance between the "Minimal" model and guide-based models was much greater in the simulation study at low data richness. At 10% data richness, the "Minimal" model achieved 0.23 (0.17-0.36) for the "real" simulations and 0.31 (0.19-0.4) for the "restructured" simulations, rates approximately 0.2 lower than the guide-based models. There is a consistent but small improvement of performance when the "correct" guidebook prior is used for the "Deterministic Correlation" and "Multivariate Probit" models, while the "Latent Factor" model performs equally well with either prior, Figure 5.5. For both simulation scenarios, the "Deterministic Correlation" receives a bump as a consequence

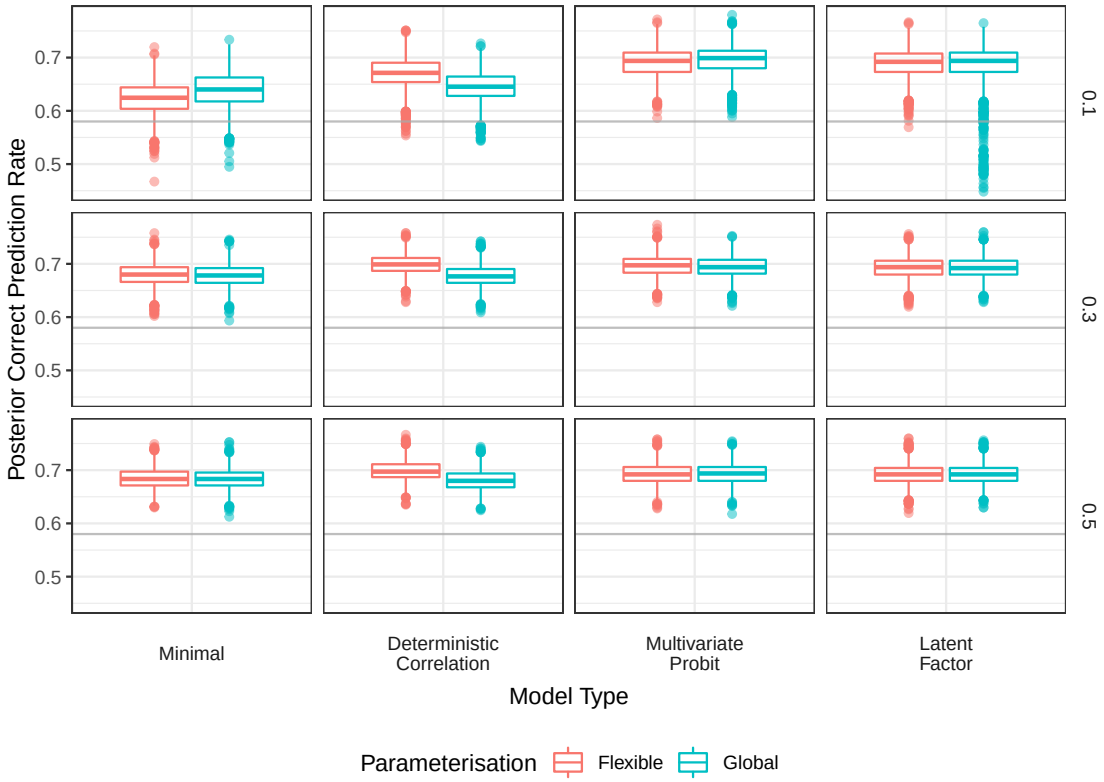

Figure 5.4: **Comparative performance of models fit to real data.** Aggregated posterior correct classification rate for models fit under cross - validation to the real data. The data richness for the cross - validation scheme is indicated by the horizontal panels (0.1=10% data richness, 0.3=30%, 0.5=50%) and model types by the vertical panels. There are two parameterisations for each model, the "flexible" one which allows the model to vary on a species-wise basis (the variance in the "Minimal" and sigmoidal transformation parameters for the others) vs "global" wherein these parameters are shared across species. The correct classification rate achieved by the citizen scientists is shown as a horizontal grey line.

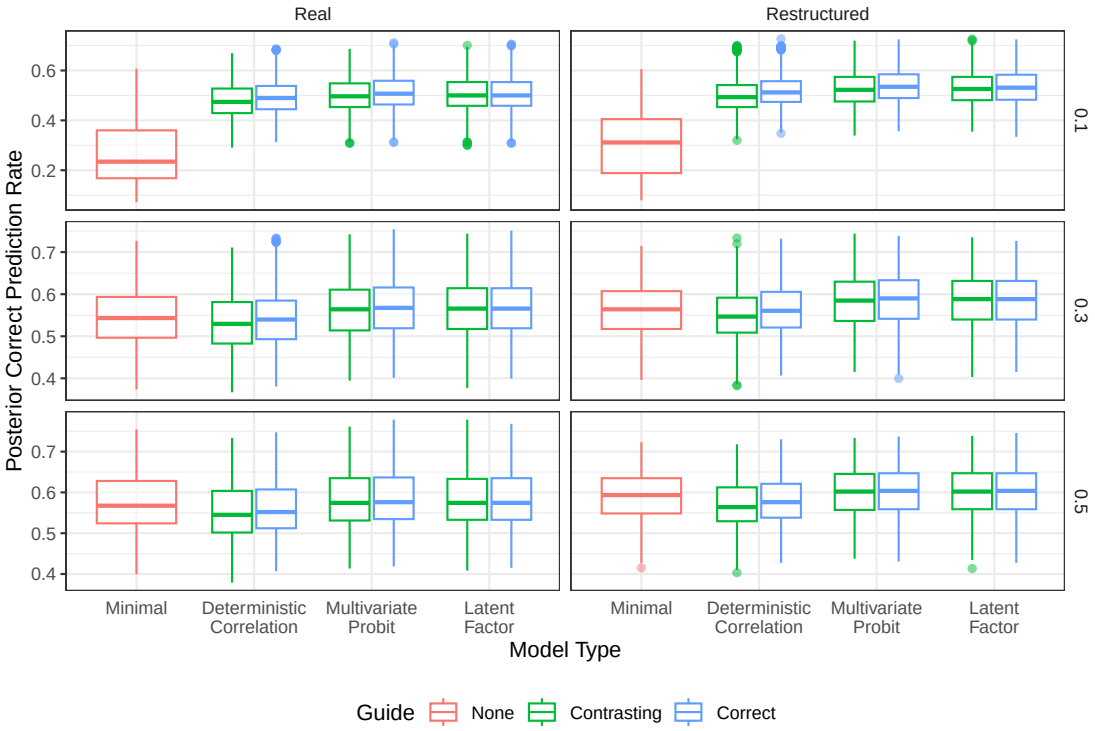

Figure 5.5: **Comparative performance of models fit to simulated data.** Data were simulated under two guidebook scenarios (indicated by the column titles), one drawn from the real data ("Real") and one from a clustered adaptation ("Restructured"). Each model was then fit to each scenario and given either the matching ("Correct") guidebook or the alternative ("Contrasting") guidebook. The "Minimal" model does not use the guidebook prior. The data richness for the cross validation scheme is indicated by the horizontal panels (0.1=10% data richness, 0.3=30%, 0.5=50%).

of using the correct prior in median correct prediction rate of 0.02 and the "Multivariate Probit" one of 0.01.

At medium data richness, the results are similar for both sets of simulations, so henceforth for clarity the results reported are for the "restructured" simulation. At 30% data richness, the "Minimal" model achieves a rate of 0.56 (0.52-0.61), a slightly better performance than the "Deterministic Correlation" model with the contrasting prior, 0.55 (0.51-0.59), and the same as it with the correct prior, 0.56 (0.52-0.61). The "Multivariate Probit" model performs better with both the incorrect prior, 0.58 (0.54-0.63) and with the correct one, 0.59 (0.54-0.63). The "Latent Factor" model equals the latter performance with both prior types.

At the highest data richness, 50%, the "Minimal" model performance, 0.59 (0.55-0.63), exceeds that of the "Deterministic Correlation" model: 0.56 (0.53-0.61) with the incorrect prior and 0.58 (0.54-0.62) with the correct one. The "Multivariate Probit" and "Latent Factor" models now perform identically with either prior, yielding a correct classification rate of 0.6 (0.56-0.65).

5.5 Discussion

We have demonstrated that there are predictable patterns to how citizen scientists confuse species and these patterns are informed by the guidebooks they use. The current approach to correcting citizen science data requires labour-intensive expert review of every single record. We have shown on real data that it is possible to reduce this work by 90 percentage points yet maintain a high rate of accurate classifications (70% with the guidebook and 65% without) with appropriate probabilistic uncertainty for each classification (compared to 58% without uncertainty when you trust the citizen scientists).

Misclassifications that are, in part, predictable, indicate that the identification guides could be improved to exploit the structure in these confusions. If the guidebooks do not improve misclassification prediction, this would indicate that there are similarities between species that are not being captured by the guidebook. If the guidebook does improve misclassification prediction (as we found here), it means the guidebooks may be the source of the confusion, for example, by making species seem too similar. In principle, the informativeness of the guidebook could vary between species, however, the global parameterisation (which links the guidebook to species more equally) generally performed similarly enough to the more flexible parameterisations of the same model. That all the models reach the same maximum correct classification rate (approximately 70%) indicates that there is perhaps no pattern in the rest of the misclassifications, with the species being confused at random. These unstructured confusions are unlikely to be affected by improvements to the guide but could potentially be mediated by increased training of participating citizen scientists if possible.

When comparing these models, we are most interested in which performs the best with the least amount of validation data. In both the real data case and simulation study, the "Multivariate Probit" and "Latent Factor" models comfortably achieve the highest correct classification rate. The other models do improve with increased data and eventually match the performance of the others. For the "Minimal" model, this is likely because it relies on uninformative priors. The

"Deterministic Correlation" model is fundamentally a point estimate version of the "Multivariate Probit" model (in terms of how they treat the guidebook). Essentially, this comes down to how estimable the correlation matrix is from the guidebook data. If the guidebook had a huge number of traits relative to the number of species, the two would give identical answers. In real applications, this is unlikely to ever happen (and with binary traits, correlations are even harder to estimate).

Our simulation study shows that using the same guidebook as a prior and to generate the data leads to a modest gain in performance for the guide-based models. The "correct" guidebook led to slightly improved performance of the "Deterministic Correlation" and "Multivariate Probit" models, and the "Latent Factor" model performed identically with the "correct" and "contrasting" priors. Nevertheless, at low data richness, these models all outperformed the "Minimal" models that used no guidebook. That the gains in model performance were only modest motivates a more thorough exploration of guidebook space. It's notable that the "Latent Factor" model (where the correlational structure is between supertraits, rather than straight guidebook traits), shows no difference at all between the two priors. Understanding the interplay between these features should form the basis of future work if we wish to use these models to directly inform future guidebook design.

There are also several exciting extensions to these models that could be developed. Currently, the models are self-contained, relying on no collection of covariates or other data sources, but it would be possible to incorporate additional information in both the biological and observation components of the model. In our studies here, we have assumed an intercept-only model for species incidence. This could be extended to include more sophisticated models that account for species-habitat associations [25] or species interactions [26]. In modelling these processes jointly, there would be a feedback loop between the observation and biological process, improving both simultaneously, and propagating uncertainty. It is worth noting that species misclassifications is only one of several issues that need to be addressed in citizen science data [3; 2].

The misclassification component of the model could also incorporate covariates. For example, in an open meadow it might be easy to follow a bee until an identification can be made confidently, while more difficult terrain might make this impossible. We could also incorporate information about the observers themselves, such as their level of experience or track record of making correct identifications [27]. The existing model structures could be easily adapted to include these features by placing a linear predictor on $\boldsymbol{C}$ with $\boldsymbol{V}$ acting as a prior for the intercept (i.e. baseline confuseability).

The widespread use of mobile applications for data collection and submission opens up the possibility of deploying these data in real time [28; 29; 30]. Several citizen science apps already offer suggestions of potential similar species (sometimes weighted probabilistically) [31; 32]. These models could be used to suggest such alternatives and to feedback common confusions to the scheme designers. In turn, this could facilitate experimentation on guidebook design, where different users are given different versions of the guidebook, or the guidebook is adapted and updated live.

5.6 Conclusions

Modelling observation processes is a challenging but essential step in modern ecological research. Frequently, we must learn these processes directly from the data but here we have shown that there are useful priors available in the form of guidebooks. The mutual benefit of combining explicitly modelling the observation process with input from citizen science scheme developers is currently underexplored, particularly when it comes to misclassification. Leveraging statistical models can help reduce the workload of taxonomic experts and thus unlock the scalability of citizen science data for ecological research. The development of these methods relies on citizen science scheme organisers adopting positive attitudes to data sharing, and methods developers engaging positively with that community to learn from them.

References

[1] R. Bonney, T. B. Phillips, H. L. Ballard, and J. W. Enck, "Can citizen science enhance public understanding of science?," *Public Understanding of Science*, vol. 25, no. 1, pp. 2–16, 2016.

[2] A. Johnston, E. Matechou, and E. B. Dennis, "Outstanding challenges and future directions for biodiversity monitoring using citizen science data," *Methods in Ecology and Evolution*, 2022.

[3] N. J. Isaac, A. J. van Strien, T. A. August, M. P. de Zeeuw, and D. B. Roy, "Statistics for citizen science: extracting signals of change from noisy ecological data," *Methods in Ecology and Evolution*, vol. 5, no. 10, pp. 1052–1060, 2014.

[4] B. Tang, J. S. Clark, and A. E. Gelfand, "Modeling spatially biased citizen science effort through the ebird database," *Environmental and Ecological Statistics*, vol. 28, no. 3, pp. 609–630, 2021.

[5] C. T. Callaghan, D. E. Bowler, S. A. Blowes, J. M. Chase, M. B. Lyons, and H. M. Pereira, "Quantifying effort needed to estimate species diversity from citizen science data," *Ecosphere*, vol. 13, no. 4, p. e3966, 2022.

[6] D. A. Miller, J. D. Nichols, B. T. McClintock, E. H. C. Grant, L. L. Bailey, and L. A. Weir, "Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification," *Ecology*, vol. 92, no. 7, pp. 1422–1428, 2011.

[7] A. I. Spiers, J. A. Royle, C. L. Torrens, and M. B. Joseph, "Estimating species misclassification with occupancy dynamics and encounter rates: a semi-supervised, individual-level approach," *Methods in Ecology and Evolution*, 2022.

[8] R. Altwegg and J. D. Nichols, "Occupancy models for citizen-science data," *Methods in Ecology and Evolution*, vol. 10, no. 1, pp. 8–21, 2019.

[9] R. Rempel, J. Jackson, S. Van Wilgenburg, and J. Rodgers, "A multiple detection state occupancy model using autonomous recordings facilitates correction of false positive and false negative observation errors," *Avian Conservation and Ecology*, vol. 14, no. 2, 2019.

[10] S. Falk, G. Foster, R. Comont, J. Conroy, H. Bostock, A. Salisbury, D. Kilbey, J. Bennett, and B. Smith, "Evaluating the ability of citizen scientists to identify bumblebee (bombus) species," *PloS One*, vol. 14, no. 6, p. e0218614, 2019.

[11] M. S. Engel, L. M. Ceríaco, G. M. Daniel, P. M. Dellapé, I. Löbl, M. Marinov, R. E. Reis, M. T. Young, A. Dubois, I. Agarwal, *et al.*, "The taxonomic impediment: a shortage of taxonomists, not the lack of technical approaches," *Zoological Journal of the Linnean Society*, vol. 193, no. 2, pp. 381–387, 2021.

[12] S. Kelling, D. Fink, F. A. La Sorte, A. Johnston, N. E. Bruns, and W. M. Hochachka, "Taking a 'big data'approach to data quality in a citizen science project," *Ambio*, vol. 44, no. 4, pp. 601–611, 2015.

[13] W. J. Wright, K. M. Irvine, E. S. Almberg, and A. R. Litt, "Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity," *Methods in Ecology and Evolution*, vol. 11, no. 1, pp. 71–81, 2020.

[14] J. C. Tweddle, L. D. Robinson, M. Pocock, and H. E. Roy, *Guide to citizen science: developing, implementing and evaluating citizen science to study biodiversity and the environment in the UK*. NERC/Centre for Ecology & Hydrology, 2012.

[15] R. F. Comont and K. Ashbrook, "Evaluating promotional approaches for citizen science biological recording: bumblebees as a group versus harmonia axyridis as a flagship for ladybirds," *BioControl*, vol. 62, no. 3, pp. 309–318, 2017.

[16] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[17] S. Chib and E. Greenberg, "Analysis of multivariate probit models," *Biometrika*, vol. 85, no. 2, pp. 347–361, 1998.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[19] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?," *Journal of Classification*, vol. 31, no. 3, pp. 274–295, 2014.

[20] J. Gabry and R. Češnovar, *cmdstanr: R Interface to 'CmdStan'*, 2022. https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org.

[21] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of Statistical Software*, vol. 76, no. 1, 2017.

[22] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner, "Rank-normalization, folding, and localization: An improved r for assessing convergence of mcmc. arxiv," *arXiv preprint arXiv:1903.08008*, 2019.

[23] M. Betancourt, "Diagnosing suboptimal cotangent disintegrations in hamiltonian monte carlo," *arXiv preprint arXiv:1604.00695*, 2016.

[24] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, "Bayesian data analysis," 2014.

[25] J. Matthiopoulos, J. R. Fieberg, and G. Aarts, "Species-habitat associations: Spatial data, predictive models, and ecological insights," 2020.

[26] O. Ovaskainen, G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego, "How to make more out of community data? a conceptual framework and its implementation as models and software," *Ecology Letters*, vol. 20, no. 5, pp. 561–576, 2017.

[27] D. E. Bowler, N. Bhandari, L. Repke, C. Beuthner, C. T. Callaghan, D. Eichenberg, K. Henle, R. Klenke, A. Richter, F. Jansen, *et al.*, "Decision-making of citizen scientists when recording species observations," *Scientific Reports*, vol. 12, no. 1, pp. 1–11, 2022.

[28] G. Newman, A. Wiggins, A. Crall, E. Graham, S. Newman, and K. Crowston, "The future of citizen science: emerging technologies and shifting paradigms," *Frontiers in Ecology and the Environment*, vol. 10, no. 6, pp. 298–304, 2012.

[29] T. August, M. Harvey, P. Lightfoot, D. Kilbey, T. Papadopoulos, and P. Jepson, "Emerging technologies for biological recording," *Biological Journal of the Linnean Society*, vol. 115, no. 3, pp. 731–749, 2015.

[30] T. A. August, O. L. Pescott, A. Joly, and P. Bonnet, "Ai naturalists might hold the key to unlocking biodiversity data in social media imagery," *Patterns*, vol. 1, no. 7, p. 100116, 2020.

[31] A. Affouard, J.-C. Lombardo, H. Goëau, P. Bonnet, and A. Joly, "Pl@ntNet," Apr. 2019.

[32] B. Qarabaqi and M. Riedewald, "Merlin: Exploratory analysis with imprecise queries," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 342–355, 2015.

The L.I.E.S. Framework

The observation processes in the following case study include:

Latency: the true species identity for each record is hidden and only the label attached to the record by the citizen scientist is observable. How similar species appear to each other from the perspective of citizen scientists is also unobservable but the case study aims to use the observable guidebook to estimate this.

Identifiability: (*mathematical*) the data in this case study are factors and modelled using the categorical distribution which is parameterised by a simplex. Simplexes are normalised and thus invariant to scaling factors making them non-identifiable without fixing a value or using a strong prior.

Identifiability: (*practical*) the model is currently predictive so practical identifiability issues could manifest as poor prediction or classification. In the future, these models could be used to make inference about *why* species are confused and parameter uncertainty will become more important.

Scaling: the guidebooks are used to estimate how close species are (and thus how likely they are to be confused), however, there is not a natural scale for this distance.

The application of the framework to the case study will be discussed in more detail in the conclusion.

Chapter 6

Conclusion

6.1 Overview

The aim of this thesis was to explore how methodological siloes in the study of observation processes could be broken down. Across ecology and epidemiology, data sets with complex observation processes are increasingly relied upon in research. The complexity of these problems, particularly the requirement to understand both sophisticated statistical methodology and nuances of data collection, tend to generate bespoke solutions that cannot be easily transferred between disciplines. As a result, the gains in efficiency and innovation that could be generated from unifying research efforts are being lost, and instead research is duplicated and commonalities that would encourage innovation are hidden from view.

I approached this from two perspectives. I began looking at observation problems from the top-down by developing a conceptual framework that I believe can describe any observation process as a combination of problems of latency, identifiability, effort and scale. I then explored the problem from the bottom-up through a series of novel case studies from ecology and epidemiology from first principles. The outstanding question of the thesis is whether the two perspectives have led to compatible results. Does the framework adequately describe the observation problems addressed in the case studies, and were the methodological similarities between the case studies more visible, thanks to the overview provided by the framework?

6.2 Complementary Perspectives on Observation Processes

6.2.1 L.I.E.S. and the Multimordibity Case Study

In the first case study, "COVID-19 – exploring the implications of long-term condition type and extent of multimorbidity on years of life lost: a modelling study", I aimed to extract the plausible covariances between diseases (multimorbidity) using only the marginal counts for each disease. In this project, there were multiple levels of latency. First, diseases are diagnosed as present or absent in different patients but frequently they are the result of an underlying, continuous scale of severity [1; 2]. I therefore had to map the binary realisations (disease present or absent) of the disease onto a continuous sub-space to quantify patterns of multimorbidity. Second, because the study of correlations in urgent clinical studies is seen as secondary, the binary realisations are often aggregated into marginal counts. This represents some loss of information, so a perfect reconstruction of joint observations is not possible. Nevertheless, when taken together, the manifest marginal counts are only consistent with some combinations of latent binary realisations.

These compounded layers of latency introduced both mathematical and practical identifiability issues. The model used to translate from binary data to covariance between the continuous latent states of the diseases is the multivariate probit [3]. If we consider one disease in isolation, the latent state takes the form of a univariate normal distribution which is then thresholded at 0 (the probit link) to

transform the continuous states into the binary realisation. The normal distribution has two parameters: the mean and variance. The portion of the distribution less than zero corresponds to the frequency of the disease. Unfortunately, the proportion of the normal distribution smaller than zero is down to the ratio of the mean and standard deviation, so the two parameters are not mathematically identifiable. The solution is to fix the standard deviation to 1. In the multivariate context, this means normalising the covariance matrix to give a correlation matrix.

Even by taking the correlation approach, the first and second layers of latency - binarisation of the disease and the aggregation of the binary data into marginal counts - introduces risk of practical identifiability problems. At both stages, information is lost, making the correlations harder to estimate without large amounts of data. Multiple correlation structures of the latent states could give rise to the patients' disease profiles and multiple combinations of disease presences and absences are coherent with the marginal counts. This manifests as a relatively large uncertainty in the estimated correlation matrix.

I ensured the uncertainty caused by the latency issues was propagated to the down-stream models. I also exploited all evident dependencies in the data, to make sure that no more uncertainty than necessary was propagated through. The resulting inference was still useful from a practical point of view but this interplay between latency and potential identifiability problems is also interesting theoretically. In this instance, it was the degrees of latency that directly contributed to the identifiability issues. Do issues of latency (or, indeed, of effort or scale) always contribute to increased identifiability issues? It seems unlikely that there will always be mathematical identifiability problems solely due to latency, but it is possible that it will always lead to increased uncertainty and thus increase the chances of practical identifiability problems. As practical identifiability will only be an issue if the uncertainty is too high to make useful inference, these two issues can exist independently but it should be noted that problems linked to one can lead to problems in the other.

In this case study, I effectively had 100% effort coverage for the target population (individuals in Italy and Scotland who had died of COVID-19). The scaling was also easy to choose in the model because I was interested in the profile of individuals. Scaling issues do come into the problem slightly in the unmodelled part of this case study in which the findings are extrapolated. Reassuringly, however, at least one paper that used the estimates produced, explicitly modelled the differences between their target population and ours [4].

6.2.2 L.I.E.S. and the COVID-19 Diagnosis Case Study

In the second case study, "Combining rapid antigen testing and syndromic surveillance improves community-based COVID-19 detection in a low-income country", I aimed to integrate data from two sources that contained complementary information about COVID-19 infection status. The first data source, rapid antigen tests, represent a binary summary of the quantity of antigen produced by the patient's body in response to COVID-19. A patient having a stronger immune response will have higher antigen production and thus be more likely to test positive. The second data source, symptomatic information, represents a multivari-

ate binary categorisation of several continuous processes linked to the immune system, but not necessarily in response to COVID-19. The latent variable of interest, infection with COVID-19, informs both data sources in similar but different ways.

As discussed with respect to the previous case study, binary data are less informative than continuous data and the same mathematical identifiability issues around estimating the means and variances apply. The model was designed to be purely predictive, so the only practical identifiability issue that can manifest is poor prediction/classification. Since these data were from a planned experiment (albeit not for the use-case to which I applied), effort is as homogeneous across the target population as could be achieved. Increasing the intensity of effort would likely improve the quality of prediction but naturally effort intensity is a function of resources which are extremely limited in this setting.

Determining the appropriate scaling for a predictive model is of practical importance. While I was explicitly not interested in interpreting parameters biologically, I did need to ensure that the evaluation of predictive power is at the appropriate scale. In this case, the structure for testing predictive power emerges as a natural property of the data. The data come in sequentially at two weekly intervals, meaning that the model needs to predict ahead two weeks at a time on existing data.

6.2.3 L.I.E.S. and the Species Misclassification Case Study

In the third and final case study, "Do identification guides hold the key to species misclassification by citizen scientists?", the aim was to identify whether citizen scientists mistake species' names in a systematic way and whether this could be linked to the guidebooks used. Much of the literature into species misclassification focuses on concepts of false positives and uses methods from that literature. Instead, I framed the label assigned by the citizen scientist as a manifest variable that carries information about the latent variable of interest: the true identities. The mechanism for this latency (i.e., why citizen scientists get species wrong) is not directly known, but I hypothesised that the structure of the guidebooks they use may act as a useful prior for these confusions.

In this case study, the data were categorical rather than binary and so were modelled using a single-trial multinomial (i.e., the categorical distribution) parameterised by a simplex. Generating simplexes often involves a normalisation step (to ensure the values sum to one), using the softmax function or similar. Often, normalisations are invariant to scaling factors (i.e., normalising a vector will give the same result as normalising the vector multiplied by a constant) which can cause mathematical identifiability problems. The issue can be resolved with either a strong prior or by setting one value to a constant - here we set correct classifications equal to one pre-normalisation.

Some of the candidate models for this case study used latent variables that represent concepts we may wish to interpret, for example, the contribution of a given trait or trait-cluster to the confusion probability of two species. While I do not interpret these parameters in the case study (I focus on prediction), I do suggest that future research could revolve around interpreting these parameters to improve guidebook design. The practical identifiability of these parameters is

not currently clear. It may be that modelling multiple guidebooks or using studies with higher effort intensity (i.e., more records) simultaneously would improve our estimation of these low level parameters. Unusually for a citizen science study, I did not worry about recording effort. This is because I explicitly chose not to incorporate any habitat parameters, although I did structure the models so they could accommodate an effort model.

Scaling manifests in this model in the observation process itself. To quantify how confuseable two species are I use correlation distance between them in the guidebook. However, I acknowledge that the correlational distance may not have a linear mapping onto the confusion distance for the citizen scientist. To address this, I include a convolutional term which effectively allows the scale to be selected as a parameter in the model.

6.2.4 L.I.E.S: a Successful Framework

The observation problems arising from the case studies can be easily described solely using the L.I.E.S. framework. Indeed, the framework provides a natural structure for summarising these issues for each case study. This is positive news for the robustness and utility of the framework although, naturally, these are only three case studies that were developed alongside the framework. There are three steps to continue development of the framework:

1. Publish the framework so that other researchers can use and test it;

2. Conduct an ongoing literature review to categorise new methods papers under the framework;

3. Identify generalisable classes of statistical solution to the different combinations of problems.

The first of these steps is perhaps the most rigorous. The framework could be conceptually sound but if it is not used by other researchers then it has failed in its main task. It is also the best way to identify conceptual or implementation flaws in the framework as it stands. The choice of journal is crucial as the framework bridges ecology, epidemiology and statistics, and requires engagement from all three disciplines. Too broad a journal and the examples are unlikely to resonate with the bulk of the audience, too narrow a journal and the benefit of cutting across themes will be lost. There are relatively few journals that occupy this space but one example is *Trends in Ecology and Evolution*[5]. If the framework is taken up with enthusiasm, there may be a case for more targeted versions of the framework aimed at sub-fields of ecology and epidemiology (e.g. which draws all the examples from fisheries science, ecotoxicology, or vector-borne diseases), or indeed at broader versions, aimed at incorporating other fields (e.g. economics, or medicine).

A systematic and ongoing review of the literature (similar to [6] reviews of COVID-19 prediction models) could benefit the framework in two ways. Firstly, it would extend the bottom-up approach taken in this thesis. If new observation process models can be readily and meaningfully explained using L.I.E.S. then the

framework will be shown to be comprehensive and robust. Secondly, a classification of existing and novel methods using the framework would aid in achieving the the third step.

An exciting outcome for this framework would be to achieve a mapping of the four canonical observation problems it defines onto generalised statistical solutions. To an extent, this is realised in the field of latent variable modelling, where different model classes are well-defined [7; 8; 9] and for identifiability issues where tools to find problems are under active development [10; 11]. However, statistical solutions to problems of effort and scale are yet to be unified. The next question would be whether the combined problems can be solved with the combined solutions, or whether additional methods would be needed to deal with each combination specifically. Being able to map each problem onto a coherent class of statistical methods would perhaps indicate some deeper truth to the framework but it is not essential for the framework to be successful. Indeed, it may represent a pyrrhic victory if the notional solutions are too complex or depart too far from existing methods, they may hinder observation process modelling more than they help.

6.3 Other Lessons Learned

6.3.1 Effective Statistics Relies On Non-Statisticians

Many applied statisticians enjoy statistics as it gives them the chance "to play in everyone's backyard" (to quote John Tukey [12]). This attitude can reflect the absolute best and worst of what it means to be an applied statistician. At our best, we help inform work in other disciplines, working with the experts in those areas to translate their knowledge and questions into meaningful inferential frameworks. At our worst, we treat non-statisticians as generators of data that allow us to build fancy models and generate publications but no insights. To stretch a metaphor, playing in someone's backyard is a lot more fun when you have been invited and your host gets to play too.

The COVID-19 pandemic has highlighted the dangers of statisticians moving into fields for which they are ill-equipped. [13; 14; 15; 16] each highlight potentially damaging attempts by ecological statisticians to contribute to pandemic modelling. The authors of the studies criticised were probably trying to add help in a time of global emergency, but by adding uninformed but convincing "solutions" into the mix, they risked drowning out more relevant voices. In the COVID-19 work in this thesis [17; 18], I was engaged in both instances by epidemiological experts and developed our models in constant dialogue with them. My models, without their input, would likely have been, at best, unhelpful, or, at worst, dangerous.

The same applies when analysing citizen science data. Many citizen science schemes, particularly long-term ones such as the Bee, Wasp and Ant Recording Society [19], have committed, long-term experts and organisers. These individuals understand the observation processes and what drives their citizen scientists better than could ever be achieved from looking at the data alone [20]. The motivations of citizen scientists are heterogeneous and in constant flux, and only those in regular contact with them tend to know why and how they change [21; 22].

Even among professional scientists, those with field experience tend to better understand where and why data collection plans are not followed (e.g., [23]), or why certain instruments give peculiar readings [24].

While I believe that this lesson is pertinent across applied statistics, it is non-negotiable for the study of observation processes. Much progress in this field will come from translating insights from non-statisticians into statistical solutions (even if not done through the L.I.E.S. framework). The frequent mutual unease between statisticians and field scientists is a major block to progress, and anything that can be done to bridge these groups is likely to have large, positive consequences.

6.3.2 Priors Are Always Informative So They Might As Well Be Informative In The Right Way
(And Reparameterisations Are Our Friends)

While the previous lesson on the value of non-statisticians was a core value to me from the start of my PhD (and reinforced during it), my understanding of the value and potential of prior modelling was hard-won. The techniques outlined in this lesson and the next (and the "Bayesian Workflow" in [10] more generally) sound like a huge amount of work at the start of model development that can be hard to motivate oneself to do. However, debugging and understanding complex models without these techniques is extremely hard. In my future work, I fully intend to embrace the "Bayesian Workflow" and wish it had featured more prominently in my thesis.

Prior specification in Bayesian models has a long and controversial history (nicely summarised in [25]). There is a general notion that we should use "non-informative" priors in our models. The precise nature of "non-informative" is poorly defined but it is often implemented as a uniform prior or diffuse normal. However, such priors are rarely actually "non-informative". An example that surprised me early in my PhD was the impact of these priors in a logistic regression. The logit link function transforms more extreme values than ±3 to 0 and 1. When a prior stretches to $\pm\infty$ (as with various "non-informative" default priors), it places the majority of the prior mass outside of the ±3 range. Thus, the prior is highly informative, suggesting extreme probabilities. I then found that as my models became more complex (for example, in the multivariate setting), the relationship between the prior distribution and the propagated prior density (the way that prior affects downstream values in the model) became even harder to understand.

I believe there are two parts to the solution to this problem that helped me, particularly in the third case study (on species misclassification). First, I needed to understand how information was pushed-forward through my model. Second, I needed a better idea of what information I wanted to push forward (as no information is not an option). I had previously thought about priors as being likelihood independent but that is not the case [26], the prior meaning is changed depending on the structure and dimensions of the likelihood. Fortunately, the likelihood structure and dimensions of a problem contain little information that bias our inference in a meaningful way. I simply used this likelihood shell to conduct prior push-forward checks, where I simulated from the prior and evaluate

how these simulations manifest when push-forward. This is a form of simulation-based calibration [11]. As there is no data being brought into the likelihood, I was only observing the impact of the priors. Where the push-forward distribution did not reflect our prior, I could refine the prior and repeat the exercise.

Now that I understood how information was projected through the model, I had to decide what level of information we want to project. In theory, the prior-push-forward approach can be used to develop the super-diffuse priors many researchers are so keen on. However, I strongly believe that the situations in which a super-diffuse prior does not actively contradict our understanding of a model are few and far between. I instead needed to understand my understanding! A thought experiment I found useful was the "bet the farm" approach. Were there any values for the parameter of interest I would willingly rule out with confidence (i.e., risk losing the farm far)? The answer, invariably, was yes. Even in situations where we do not have as much knowledge, there are generally some physical limits or practical boundaries that we can call upon when setting a prior. When estimating the size of an animal, we know they are multi-cellular, we know they have to fit within a continent or ocean. Even these very broad assumptions are often an improvement over the diffuse prior approach - all without risking the farm.

Sometimes I did have strong prior information but it did not correspond to a parameter in my model. For example, in the species misclassification case study, I often did not know have information on individual confusion pairs (outwith the guidebook, that is) so had to set exchangeable priors, but I knew something about the overall patterns of confusion (i.e., they were unlikely to be uniform).. In this situation, my prior knowledge was perhaps best applied as a check on summary statistics for our prior-push-forward model. If the prior-push-forward suggests all species are equally confuseable all the time, I may have needed to modify the priors to change that behaviour. To do so, I still needed an understanding of which priors affect which bits of the model. Simulation experiments told me this, however, I also made my life easier by setting up my model generatively. A generative model is one which is meaningfully decomposed into interpretable parameters and sub-models, and from which data can be simulated. While the prior-push-forward effects can still be surprising, structuring our models in this way is extremely helpful for sense checking them.

Some default parameterisations of distributions are not conducive to generative thinking. For example, the Beta distribution has shape parameters which do not have a clear intrinsic meaning and can only be evaluated in combination. This can make setting priors for them extremely challenging (especially when trying to incorporate meaningful expert information). Fortunately, the distribution can be reparameterised in terms of the mean and variance (or even skew and kurtosis). These parameters are much more interpretable, making it easier to elicit expertise to set them (even if just through betting the farm) and to see how values are propagated through the model.

6.3.3 Posterior Sampling Is Difficult
(And Reparameterisations Are Our Friends)

The revolution in statistics that led to the widespread embrace of Bayesian methods was arguably not driven by the conceptual and philosophical benefits of the paradigm. These benefits did keep the Bayesian flame alive, but it was the development of Markov chain Monte Carlo (MCMC) algorithms and, in particular, their user-friendly implementations in the form of BUGS-type [27] probabilistic programming languages (PPLs) that has led to widespread uptake of Bayesian methods [28]. The relief for a statistician like me is huge! I can access this modelling paradigm with its myriad benefits without needing to calculate the posterior using analytic solutions or hand-coding my own MCMC algorithms. This is particularly useful more generally given the skills needed to do high-powered algebra or to code efficient algorithms and those needed to develop interesting and useful models overlap far less than we might hope. The downside of this democratisation of Bayes, however, is that it's easy to forget what a complex task we are asking a few lines of code to achieve.

I, and most applied statisticians, think relatively little about the posterior geometries we are asking JAGS [29] or Stan [30] or any other PPL to navigate. This is largely appropriate; learning measure theory is not something many of us have the time or skill to do. However, I still rely on our MCMC algorithms having successfully navigated the high-dimensional, Escher-ian volumes I have created in order to give me important inferential guarantees. I therefore need to be convinced the samplers have done what I want based on summaries from the PPL. For this reason, I wanted to choose a PPL that gives informative diagnostic summaries for the samplers, as well as the more conventional convergence diagnostics for the parameters.

To me, Stan [30] stands head and shoulders above most other PPLs in the provision of sampling diagnostics. The reasons are manifold (pun intended). Firstly, the development team behind Stan believe in opinionated software (i.e. that things should "fail noisily" where possible). Secondly, the user community of Stan is extremely large meaning that the user interface (including error messages) is tested a lot and fed back for improvement. Thirdly, Stan implements a form of dynamic Hamiltonian Monte Carlo (HMC) [31]. When this class of MCMC fails it does so in an informative way leading to useful diagnostics (as well as including more general MCMC diagnostics such as effective sample size and split-$\hat{R}$).

There are three main HMC-specific diagnostics provided by Stan: max tree-depth exceeded, low energy Bayesian Fraction of Missing Information (eBFMI) and divergent transitions. The first of these relates to the efficiency of the sampler and, if all other metrics are okay, is not majorly concerning. The second and third of these metrics indicate the estimated posterior may be unreliable. Fortunately, both of them can be explored visually (using manual plotting or interactive tools like ShinyStan [32]) to identify the problematic region of the posterior. Often, there are a large number of options to tame the posterior geometry to improve sampling, kept up to date in the Stan manual [33] and on the Stan webstie [34].

One technique to improve model geometry that I used multiple times in this thesis is to reparameterise hierarchical model components [35]. Hierarchical parameters are often correlated with one another and relatively weakly informed

by the data. This can lead to a "funneling" in the posterior, i.e. in some portions of the parameter space, the region of mutually compatible values for the two parameters becomes extremely small. Funneling makes the parameters hard to sample, as when the sampler is adapted for the wider part of the funnel it is not be able to enter the narrow part of the funnel, and when the sampler is adapted for the narrow part of the funnel it is not be able to efficiently sample the wide end. Reparameterisation saved me here. I started with sampling parameters with Normal priors (although this technique works for other distributions too). Instead of drawing from a Normal with mean, μ, and variance, σ^2, I instead used standard normals (mean of zero and variance of 1) and then transformed by multiplying the results by σ and adding μ. In doing so, Stan can sample the full parameter space without having to directly sample the funnel geometry.

6.3.4 Beware of Conventions, Defaults and Asymptotic Properties

Every part of statistics is hard and we all want to do a good job. This can make it tempting to reach for off-the-shelf performance metrics and model components. Doing so allows us to defer to the wisdom of others (often backed up by complex but reassuring asymptotic properties) and reduce how much we have to think. Unfortunately, if we are too deferential, we can cause huge problems for ourselves if our problem is different to the one these tools were designed around. For example, generic concepts of model performance are hard to imagine in applied statistics. In general, the performance of the model needs to be contextualised with reference to the problem we are trying to solve. Similarly, asymptotic properties look extremely impressive (especially to those of us who are less mathematically minded) but the majority of the time applied statisticians are working in pre-asymptotic regimes [36].

Even when the asymptotic properties of a metric hold, it is often still debatable whether or not we actually want that metric. Many of the information criteria approximate leave-one-out cross-validation. Genuine leave-one-out cross-validation is extremely computationally intensive, so the appeal of an approximation is clear, *if leave-one-out is a relevant test of performance*. In most of my models, not all data points are born equal. Whenever there was structure in my data, my ability to predict some data points was higher than others. In the diagnostic case study, I knew that relationships between symptoms were changing over time. The prediction problem was one of forecasting (i.e. predicting the next few time points). Predicting within the existing time series of data was much easier as I knew what the relationship between symptoms was either side of the point we were predicting. Unfortunately, this did not reflect the prediction problem I had, so would have overestimated the power of my model. The same would apply to spatial models. It is much easier to predict the value of a cell in the middle of our study region than one on the edge of it. The same problems apply for almost any model with autocorrelated or hierarchical components. Unfortunately, this generally means that the best predictive measurements are computationally intensive, structured cross-validation schemes (although there is some work on automated [37] and approximated [38] cross validation for structured problems now).

Sometimes there were no good performance metrics. Multivariate methods for evaluating predictive performance are extremely challenging. In a multivariate categorical model, for a given input the model generates a prediction simplex. I generally wanted to evaluate a multivariate model on the full simplex, i.e. if the models first guess was not good, was the second guess and so on. Kullback-Leibler divergence and derived metrics like cross entropy should in principle have given me this behaviour, with lots of nice mathematical behaviours to back them up. They measure the distance between distributions, so we should be able to compare the predictive distribution with the true distribution and the divergence between the two tells us the relative performance. However, I never had the true distribution so I had to evaluate based on data which are realisations of the true distribution. This means that I needed to compare my predictive distribution with vectors of 0s and 1s. For the univariate case, these values converge with relatively little data. But as the dimensions increase, the convergence becomes extremely slow. While there are not necessarily good non-default metrics, the key here was not to be taken in by alleged multivariate properties here that do not manifest in practice.

These problems do not just apply to model evaluation, model components can also have surprising properties. The softmax (or multi-logit) transformation is a widely used transformation that takes a vector of reals and transforms it to a simplex, by taking the exponent of each element of the vector and dividing it by the sum of the exponent of each element of the vector. In one of the earlier iterations of the species misclassification models, I used softmax to take the vector of correlation distances between the species and return a simplex of confusion probabilities. However, softmax has no sparsity properties. This means that even in the situation where the input vector is a 1 (for correct classification) and series of 22x -1s (for each misclassification), the probability of correct identification is only 0.25. If the same was true for a 100 species system, the probability of correct identification becomes just 0.07. Such behaviour is extremely undesirable for most use-cases [39] but softmax is still used as a standard method to generate a simplex.

6.4 Conclusion

Observation process modelling is an exciting frontier in applied statistics. The potential gains to be made by unifying the excellent work, currently spread across numerous area of application, are vast. Unification allows us to reduce research waste, kick-start innovation, and continue to unlock the enormous potential of complex data sources from acoustic records to Zooniverse's citizen scientists. In this thesis, I have laid out my vision for achieving this (the L.I.E.S. framework) and tested it against three case studies. The framework has survived these trials and must now be pitted against a larger audience and used to start a conversation about how we can progress. The future of this field is exciting, daunting and waiting.

References

[1] R. Plomin, C. Haworth, and O. S. Davis, "Common disorders are quantitative traits," *Nature Reviews Genetics*, vol. 10, no. 12, pp. 872–878, 2009.

[2] R. Sukkar, E. Katz, Y. Zhang, D. Raunig, and B. T. Wyman, "Disease progression modeling using hidden markov models," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2845–2848, IEEE, 2012.

[3] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[4] T. Ferenci, "Different approaches to quantify years of life lost from covid-19," *European Journal of Epidemiology*, vol. 36, no. 6, pp. 589–597, 2021.

[5] A. Sugden, "Trends in ecology and evolution," *Trends in Ecology & Evolution*, vol. 1, no. 1, p. 2, 1986.

[6] L. Wynants, B. Van Calster, G. S. Collins, R. D. Riley, G. Heinze, E. Schuit, M. M. Bonten, D. L. Dahly, J. A. Damen, T. P. Debray, *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *bmj*, vol. 369, 2020.

[7] A. A. Beaujean, *Latent variable modeling using R: A step-by-step guide*. Routledge, 2014.

[8] W. H. Finch and B. F. French, *Latent variable modeling with R*. Routledge, 2015.

[9] J. Loehlin and A. Beaujean, *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis, Fifth Edition (5th ed.)*. Routledge, 2017.

[10] A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák, "Bayesian workflow," *arXiv preprint arXiv:2011.01808*, 2020.

[11] M. Modrák, A. H. Moon, S. Kim, P. Bürkner, N. Huurre, K. Faltejsková, A. Gelman, and A. Vehtari, "Simulation-based calibration checking for bayesian computation: The choice of test quantities shapes sensitivity," *arXiv preprint arXiv:2211.02383*, 2022.

[12] D. Leonhardt, "John tukey, 85, statistician; coined the word "software"," *New York Times*, vol. 28, 2000.

[13] C. J. Carlson, J. D. Chipperfield, B. M. Benito, R. J. Telford, and R. B. O'Hara, "Don't gamble the covid-19 response on ecological hypotheses," *Nature Ecology & Evolution*, vol. 4, no. 9, pp. 1155–1155, 2020.

[14] C. J. Carlson, J. D. Chipperfield, B. M. Benito, R. J. Telford, and R. B. O'Hara, "Species distribution models are inappropriate for covid-19," *Nature Ecology & Evolution*, vol. 4, no. 6, pp. 770–771, 2020.

[15] J. D. Chipperfield, B. M. Benito, R. O'Hara, R. J. Telford, and C. J. Carlson, "On the inadequacy of species distribution models for modelling the spread of sars-cov-2: response to araújo and naimi," 2020.

[16] A. Contina, S. W. Yanco, A. K. Pierce, M. DePrenger-Levin, M. B. Wunder, A. M. Neophytou, C. P. Lostroh, R. J. Telford, B. M. Benito, J. Chipperfield, R. B. O'Hara, and C. J. Carlson, "Comment on "a global-scale ecological niche model to predict sars-cov-2 coronavirus infection rate", author coro," *Ecological Modelling*, vol. 436, p. 109288, 2020.

[17] P. Hanlon, F. Chadwick, A. Shah, R. Wood, J. Minton, G. McCartney, C. Fischbacher, F. S. Mair, D. Husmeier, J. Matthiopoulos, *et al.*, "Covid-19–exploring the implications of long-term condition type and extent of multimorbidity on years of life lost: a modelling study," *Wellcome Open Research*, vol. 5, 2020.

[18] F. J. Chadwick, J. Clark, S. Chowdhury, T. Chowdhury, D. J. Pascall, Y. Haddou, J. Andrecka, M. Kundegorski, C. Wilkie, E. Brum, *et al.*, "Combining rapid antigen testing and syndromic surveillance improves community-based covid-19 detection in a low-income country," *Nature Communications*, vol. 13, no. 1, pp. 1–9, 2022.

[19] R. Edwards, *BWARS: Bees, Wasps and Ants Recording Society starter pack*. Biological Records Centre, 1996.

[20] A. Zizka, F. A. Carvalho, A. Calvente, M. R. Baez-Lizarazo, A. Cabral, J. F. R. Coelho, M. Colli-Silva, M. R. Fantinati, M. F. Fernandes, T. Ferreira-Araújo, *et al.*, "No one-size-fits-all solution to clean gbif," *PeerJ*, vol. 8, p. e9916, 2020.

[21] D. Rotman, J. Preece, J. Hammock, K. Procita, D. Hansen, C. Parr, D. Lewis, and D. Jacobs, "Dynamic changes in motivation in collaborative citizen-science projects," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pp. 217–226, 2012.

[22] S. E. West, R. M. Pateman, and A. Dyke, "Variations in the motivations of environmental citizen scientists," *Citizen Science: Theory and Practice*, 2021.

[23] I. Svanberg, "Encounters with fierce dogs and itchy bedbugs: why my first field work failed," *Journal of Ethnobiology and Ethnomedicine*, vol. 10, no. 1, pp. 1–8, 2014.

[24] E. A. Roznik and R. A. Alford, "Does waterproofing thermochron ibutton dataloggers influence temperature readings?," *Journal of Thermal Biology*, vol. 37, no. 4, pp. 260–264, 2012.

[25] A. Gelman and C. Hennig, "Beyond subjective and objective in statistics," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 180, no. 4, pp. 967–1033, 2017.

[26] A. Gelman, D. Simpson, and M. Betancourt, "The prior can often only be understood in the context of the likelihood," *Entropy*, vol. 19, no. 10, p. 555, 2017.

[27] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best, "The bugs project: Evolution, critique and future directions," *Statistics in Medicine*, vol. 28, no. 25, pp. 3049–3067, 2009.

[28] D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye, "Penalising model component complexity: A principled, practical approach to constructing priors," *Statistical Science*, vol. 32, no. 1, pp. 1–28, 2017.

[29] M. Plummer *et al.*, "Jags: A program for analysis of bayesian graphical models using gibbs sampling," in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, vol. 124, pp. 1–10, Vienna, Austria., 2003.

[30] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of Statistical Software*, vol. 76, no. 1, 2017.

[31] M. Betancourt, "A conceptual introduction to hamiltonian monte carlo," *arXiv preprint arXiv:1701.02434*, 2017.

[32] S. D. Team, "shinystan: Interactive visual and numerical diagnostics and posterior analysis for bayesian models," *R package version 2.4. 0*, 2017.

[33] S. D. Team, "Stan modeling language users guide and reference manual," *https://mc-stan.org*, 2022.

[34] S. D. Team, "Runtime warnings and convergence problems," *https://mc-stan.org/misc/warnings.html*, 2022.

[35] M. Betancourt and M. Girolami, "Hamiltonian monte carlo for hierarchical models," *Current Trends in Bayesian Methodology with Applications*, vol. 79, no. 30, pp. 2–4, 2015.

[36] A. Vehtari, A. Gelman, and J. Gabry, "Practical bayesian model evaluation using leave-one-out cross-validation and waic," *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[37] R. Valavi, J. Elith, J. Lahoz-Monfort, and G. B. Guillera-Arroita, "An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models," *Methods Ecol. Evol*, vol. 10, pp. 225–232, 2018.

[38] S. Ghosh, W. Stephenson, T. D. Nguyen, S. Deshpande, and T. Broderick, "Approximate cross-validation for structured models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8741–8752, 2020.

[39] A. Martins and R. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in *International Conference on Machine Learning*, pp. 1614–1623, PMLR, 2016.

Appendix A

Supplementary Materials for Chapter 3: Mathematical Description of Aggregated Comorbidities Model

We describe the mathematical procedure on which the inference of typical COVID - 19 patient comorbidity profiles from combined low - cardinality complete data (Scottish patients) and high - cardinality incomplete data (Italian records) is based.

A complete data vector $\mathbf{y}_i$ in our study is a K-dimensional binary vector of ones and zeros indicating the presence or absence of a medical condition in a given patient i. So $y_{ik} = 1$ indicates that medical condition k in patient i is present, and it is absent if $y_{ik} = 0$. Dependencies between comorbidities are modelled with a latent multivariate normal distribution

$$\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{A.1}$$

where $\boldsymbol{\Sigma}$ is a covariance matrix with all diagonal elements kept fixed at 1 (i.e. a correlation matrix). The conditional probability of an observation given the latent variable is

$$p(y_{ik}|x_{ik}, \boldsymbol{\xi}) = [\psi(x_{ik}; \boldsymbol{\xi})]^{y_{ik}}[1 - \psi(x_{ik}; \boldsymbol{\xi})]^{1-y_{ik}} \tag{A.2}$$

where ψ is a link function, like the probit or the logit, and $\boldsymbol{\xi}$ are its parameters. Conditional on the latent variables $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, the data $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ are assumed to be independent

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\xi}) = \prod_{i=1}^{N}\prod_{k=1}^{K} p(y_{ik}|x_{ik}, \boldsymbol{\xi}) \tag{A.3}$$

where N is the number of patients with complete records, and K is the number of comorbidities. As it turns out, in the data for the Scottish patients, a zero entry does not indicate the absence of a comorbidity, but an unknown disease status. These entries therefore have to be treated as missing values, leading to the following modification:

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\xi}) = \prod_{i=1}^{N}\prod_{k=1}^{K} \Big[p(y_{ik}|x_{ik}, \boldsymbol{\xi})\Big]^{y_{ik}} \tag{A.4}$$

The parameters $\boldsymbol{\xi}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can in principle be estimated by maximizing the likelihood

$$p(\mathbf{Y}|\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \int p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\xi})\mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma})d\mathbf{x} \tag{A.5}$$

where we define the matrix normal distribution as follows:

$$\mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N}\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{A.6}$$

Now, in addition to the complete patient data, we have incomplete data $\tilde{\mathbf{Y}} = (\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_{\tilde{N}})$, where the individual comorbidity indicators $\tilde{y}_{ik}$ are unknown, and we only have access to the marginal counts indicating the total number of occurrences of comorbidity k:

$$\tilde{y}_{.,k} = \sum_{i=1}^{\tilde{N}} \tilde{y}_{i,k} \tag{A.7}$$

In words: $\tilde{y}_{.,k}$ represents the total number of patients in the complementary incomplete data set (the "Italian" data) for which comorbidity k has been recorded. We also have the patient-wise comorbidity counts:

$$\phi(m) \;=\; \sum_{i=1}^{\tilde{N}} \delta\left(m, \sum_{k=1}^{K} \tilde{y}_{ik}\right) \tag{A.8}$$

where $\delta(.,.)$ is the Kronecker delta. In words: $\phi(m)$ denotes the number of patients in the complementary incomplete data set (the "Italian" data) for which m comorbidities have been recorded. Note that we deliberately use different symbols m and k in (A.7) and (A.8). While k in (A.7) is an identifier for a specific comorbidity, m in (A.8) is a count of comorbidities. In practice $\phi(m)$ may be subject to censoring, as we will discuss below. The total likelihood is given by

$$p(\mathbf{Y}, \tilde{\mathbf{Y}}|\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; \int p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\xi}) p(\tilde{\mathbf{Y}}|\tilde{\mathbf{X}}, \boldsymbol{\xi}) \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathcal{N}(\tilde{\mathbf{X}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{X} d\tilde{\mathbf{X}} \tag{A.9}$$

with

$$p(\tilde{\mathbf{Y}}|\tilde{\mathbf{X}}, \boldsymbol{\xi}) \;=\; \sum_{\mathbf{H}} p(\tilde{\mathbf{Y}}|\mathbf{H}) p(\mathbf{H}|\tilde{\mathbf{X}}, \boldsymbol{\xi}) \tag{A.10}$$

where $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_{\tilde{N}})$, $\mathbf{h}_i = (h_{i1}, \ldots, h_{iK})^{\mathsf{T}}$, and $h_{ik} \in \{0, 1\}$ is a latent indicator for the presence of comorbidity k in patient i. The structure of $\mathbf{H}$ mimics that of $\mathbf{Y}$ and augments the information that is missing in the incomplete "Italian" data $\tilde{\mathbf{Y}}$; so $p(\mathbf{H}|\tilde{\mathbf{X}}, \boldsymbol{\xi})$ is given by (A.3), with $\mathbf{Y}$ replaced by $\mathbf{H}$, and separate latent variables $\tilde{\mathbf{X}}$:

$$p(\mathbf{H}|\tilde{\mathbf{X}}, \boldsymbol{\xi}) \;=\; \prod_{i=1}^{\tilde{N}} \prod_{k=1}^{K} p(h_{ik}|\tilde{x}_{ik}, \boldsymbol{\xi}) \tag{A.11}$$

However, since this information is not actually available, $\mathbf{H}$ is a latent variable (our procedure follows the standard procedure of data augmentation in missing data problems). The first term in the sum on the right-hand side of (A.10) is given by

$$p(\tilde{\mathbf{Y}}|\mathbf{H}) \;=\; \prod_{k=1}^{K} \delta\left(\tilde{y}_{.,k}, \sum_{i=1}^{\tilde{N}} h_{i,k}\right) \times \prod_{m=1}^{K} \delta\left(\phi(m), \sum_{i=1}^{\tilde{N}} \delta\left(m, \sum_{k=1}^{K} h_{ik}\right)\right) \tag{A.12}$$

This is a filter that accepts those latent variable matrices $\mathbf{H}$ that are consistent with the marginal constraints given by the incomplete data $\tilde{\mathbf{Y}}$. The marginalization in (A.10) is computationally onerous, so we use a Bayesian sampling approach

$$\begin{aligned} & p(\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}, \tilde{\mathbf{X}}, \mathbf{H}|\mathbf{Y}, \tilde{\mathbf{Y}}) && \text{(A.13)} \\ \propto\; & p(\mathbf{Y}, \tilde{\mathbf{Y}}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}, \tilde{\mathbf{X}}, \mathbf{H}) && \text{(A.14)} \\ =\; & p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\xi}) \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\tilde{\mathbf{Y}}|\mathbf{H}) p(\mathbf{H}|\tilde{\mathbf{X}}, \boldsymbol{\xi}) \mathcal{N}(\tilde{\mathbf{X}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \pi(\boldsymbol{\mu}) \pi(\boldsymbol{\Sigma}) \pi(\boldsymbol{\xi}) \end{aligned}$$

with prior distributions $\pi(\boldsymbol{\mu}), \pi(\boldsymbol{\Sigma})$ and $\pi(\boldsymbol{\xi})$. We follow a Gibbs sampling strategy:

$$\begin{aligned} \mathbf{X} \;&\sim\; p(\mathbf{X}|\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{H}, \mathbf{Y}, \tilde{\mathbf{Y}}, \tilde{\mathbf{X}}) \;\propto\; p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\xi}) \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) && \text{(A.15)} \\ \tilde{\mathbf{X}} \;&\sim\; p(\tilde{\mathbf{X}}|\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{H}, \mathbf{Y}, \mathbf{X}, \tilde{\mathbf{Y}}) \;\propto\; p(\mathbf{H}|\tilde{\mathbf{X}}, \boldsymbol{\xi}) \mathcal{N}(\tilde{\mathbf{X}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) && \text{(A.16)} \\ \mathbf{H} \;&\sim\; p(\mathbf{H}|\boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}, \tilde{\mathbf{X}}, \mathbf{Y}, \tilde{\mathbf{Y}}) \;\propto\; p(\tilde{\mathbf{Y}}|\mathbf{H}) p(\mathbf{H}|\tilde{\mathbf{X}}, \boldsymbol{\xi}) && \text{(A.17)} \\ \boldsymbol{\mu}, \boldsymbol{\Sigma} \;&\sim\; p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{X}, \tilde{\mathbf{X}}, \boldsymbol{\xi}, \mathbf{H}, \mathbf{Y}, \tilde{\mathbf{Y}}) \;\propto\; \mathcal{N}(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mathcal{N}(\tilde{\mathbf{X}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \pi(\boldsymbol{\mu}) \pi(\boldsymbol{\Sigma}) && \text{(A.18)} \end{aligned}$$

In principle we could include $\boldsymbol{\xi}$ in this sampling scheme, but we decided to keep it fixed, set such that the link function in (A.2) reduces to a step function at zero. The parameters of the latent Gaussian distribution, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (A.18), can be directly sampled from a distribution that is available in closed form if the priors $\pi(\boldsymbol{\mu})$ and $\pi(\boldsymbol{\Sigma})$ are conjugate. The conditional distributions in (A.15–A.17) are not available in closed form, so we have to resort to a Metropolis-Hastings-within-Gibbs scheme. The computation of $p(\tilde{\mathbf{Y}}|\mathbf{H})$ in (A.17) via (A.12) is effectively a rejection sampling step, which in practice may have very low acceptance probability. To improve mixing and convergence of the Markov chain we have therefore introduced a slight "fudge" and replaced the Kronecker deltas in (A.12) by peaked Gaussians with very small variance.

Rather than choosing a conjugate prior for $\boldsymbol{\Sigma}$ (an inverse Wishart distribution), we choose a prior for the individual off-diagonal elements, shrinking them to zero with a rescaled beta distribution. This makes it easier to satisfy the constraint of keeping the diagonal elements fixed at 1. In addition, sampling from an inverse Wishart distribution is not available in JAGS, which we used for this project. The consequence is that sampling has now to be carried out with a Metropolis-Hastings-within-Gibbs scheme, slightly deteriorating the rate of convergence.
The output of interest is a collection of typical patient profiles $\{\mathbf{H}_i\}$. To this end, we sample from the joint posterior distribution (A.13) and marginalize over the other parameters (by simply discarding them).

A practical problem that we have ignored so far is censoring of the marginal comorbidity counts $\phi(m)$ in (A.8). To deal with this, we use the prior knowledge that these counts can be assumed to follow a Poisson distribution. A simple approach is to fit a Poisson distribution to the censored counts and use data augmentation to impute the missing values, following the straightforward procedure described in Selvin (1974). A disadvantage is that this would not take the uncertainty of the imputation into account.

We therefore integrate this data augmentation step into our Gibbs sampling routine. The set of $\boldsymbol{\phi(m)} = \{\phi(m)\}$ in (A.12) becomes a latent variable, on which the distribution in (A.12) now explicitly depends: $p(\tilde{\mathbf{Y}}|\mathbf{H}) \to p(\tilde{\mathbf{Y}}|\mathbf{H}, \boldsymbol{\phi(m)})$. The distribution of these counts follows a Poisson distribution:

$$\boldsymbol{\phi(m)} \;\sim\; \mathcal{P}(\boldsymbol{\phi(m)}|\lambda) \tag{A.19}$$

whose parameter λ is sampled from the posterior distribution of the available censored counts:

$$p(\lambda|\tilde{\boldsymbol{\phi}}(\boldsymbol{m})) \;\propto\; p(\tilde{\boldsymbol{\phi}}(\boldsymbol{m})|\lambda)\pi(\lambda) \;=\; \pi(\lambda) \sum_{\boldsymbol{\phi(m)}} \delta\Big(\tilde{\boldsymbol{\phi}}(\boldsymbol{m}), \boldsymbol{\phi(m)}\Big)\mathcal{P}(\boldsymbol{\phi(m)}|\lambda) \tag{A.20}$$

for which standard procedures are available; see [1]. The upshot is that (A.19) and (A.20) have to be included as additional sampling steps in our Gibbs sampling routine (A.15–A.18).

When feeding the sampled comorbidity profiles $\mathbf{H}$ into the survival analysis, we need to allow for the fact that these samples were drawn from a Markov

chain and that they are therefore not independent. Let $f(\mathbf{H})$ denote the life expectancy from SAIL for patients with comorbidity profiles $\mathbf{H}$, then the expected SAIL-equivalent life expectancy for the Italian Covid-19 patient population is

$$E \;=\; \int f(\mathbf{H}) p(\mathbf{H}|\mathbf{Y}, \tilde{\mathbf{Y}}) d\mathbf{H} \tag{A.21}$$

which in practice is approximated with a finite sample of comorbidity profiles from our MCMC sampler:

$$\hat{E} \;\approx\; \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{H}_m) \tag{A.22}$$

Here $\mathbf{H}_m$ is the mth sample from the Markov chain. The uncertainty quantification for our estimator $\hat{E}$ from a sample of M independent samples is given by

$$\sigma(\hat{E}) \;=\; \sqrt{\frac{\mathrm{var}(f)}{M}} \tag{A.23}$$

where

$$\mathrm{var}(f) \;=\; \frac{1}{M-1} \sum_{m=1}^{M} \left(f(\mathbf{H}_m) - \hat{E} \right)^2 \tag{A.24}$$

However, given that the M comorbidity profiles are dependent, this would systematically underestimate the uncertainty. The correct uncertainty quantification is given by

$$\sigma(\hat{E}) \;=\; \sqrt{\frac{\mathrm{var}(f)}{M_{\mathit{eff}}}} \tag{A.25}$$

where M_{eff} is the effective sample size, which represents the equivalent size of a sample of independent draws from the same distribution and is mathematically defined as

$$M_{\mathit{eff}} \;=\; \frac{M}{1 + 2\sum_{t=1}^{\infty} \rho_t} \tag{A.26}$$

The quantities ρ_t denote the t-step autocorrelations of the MCMC trajectory, which can be computed with standard MCMC analysis software packages, like the `coda` package from `CRAN` (which, in fact, directly computes M_{eff} as well). For the mathematical details, see chapter 11 in [2].

References

[1] S. Selvin, "Maximum likelihood estimation in the truncated or censored poisson distribution," *Journal of the American Statistical Association*, vol. 69, no. 345, pp. 234–237, 1974.

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. Duncan, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. Chapman and Hall/CRC, 2014.

Appendix B

Supplementary Materials for Chapter 4: Additional Results and Methodological Details

B.1 Supplementary Figures

The key result we would like to highlight in the paper is the power of the workflow we have developed. Individual parameter estimates will vary depending on the population to which the models are applied and we strongly advise against applying the parameter values found here (for example, the final symptoms chosen through model selection) outwith the population and time to which the models were fit.

B.1.1 Correlation Estimates

The relationship between symptoms and results can only be understood through the full correlation matrix. In Figures 1 and 2, we present the median correlations for the four symptom Syndromic-only and Syndromic-RAT Combined models. These results should not be used to prioritise future data collection because the most predictive symptoms are liable to change with time (e.g., the emergence of new COVID variants) and context (e.g., broader vaccination levels).

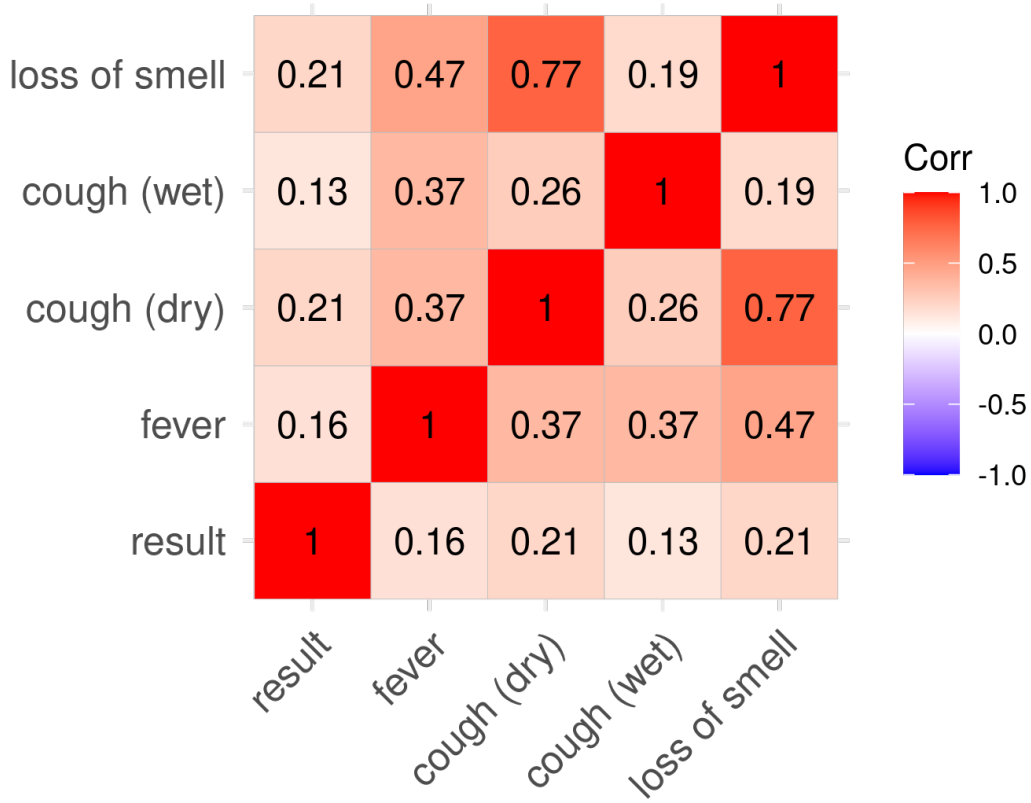

Figure B.1: Median correlation between PCR result and top symptoms for 4 symptom Syndromic-only Model

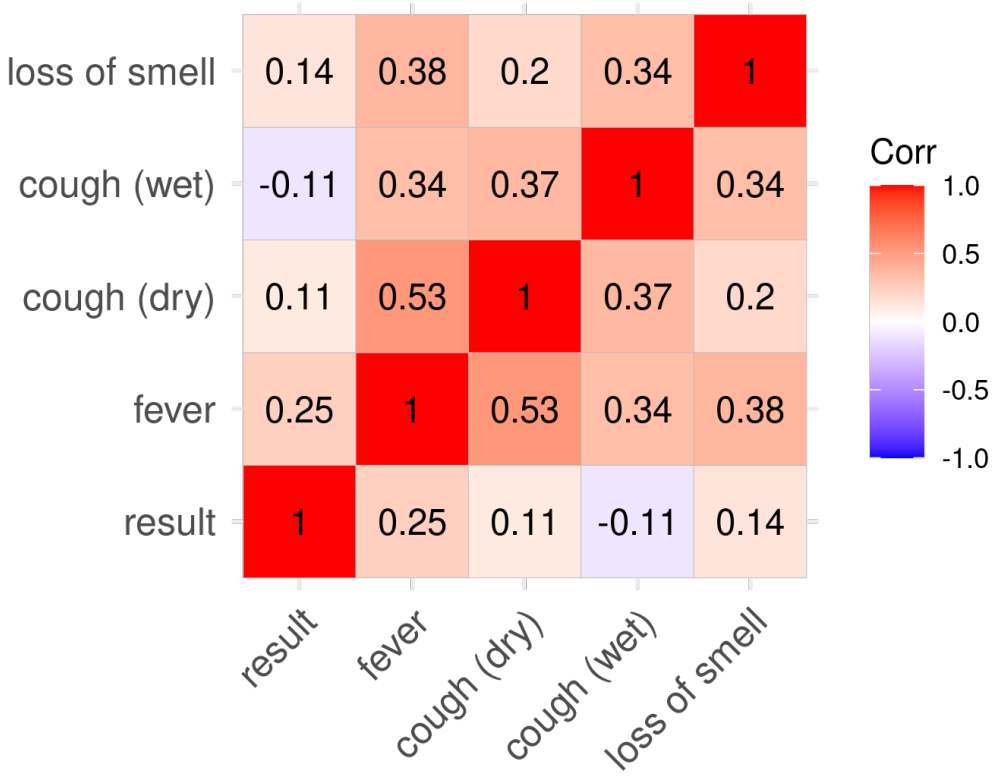

Figure B.2: Median correlation between PCR result and top symptoms for 4 symptom Syndromic-RAT Combined Model

B.2 Supplementary Tables

B.2.1 Translation of Error Rates into Raw Numbers Based on Case Positivity Rate

False positive and false negative rates can only be translated into numbers of people affected if the case positivity rate is known. To demonstrate how the numbers of misclassifications change for the same false positive and false negative rates, we have scaled these numbers with respect to low (5%), average (20%) and high (35%) CPRs in Bangladesh in Table B.1.

Table B.1: Translation of best model performance by scenario into number of patients per 1000 tested who were incorrectly diagnosed, broken down by case positivity rate (CPR). CPRs chosen to reflect low, average and high values in Bangladesh.

		Per 1000					
		5% CPR		20% CPR		35% CPR	
Model Class	Scenario	False Positives	False Negatives	False Positives	False Negatives	False Positives	False Negatives
RATonly	All	24	20	21	81	17	141
SyndOnly	1	400	21	337	85	274	149
SyndRAT	1	105	18	88	71	72	124
SyndOnly	2	703	10	592	40	481	69
SyndRAT	2	392	10	330	39	268	68
SyndOnly	3	189	34	159	137	129	240
SyndRAT	3	188	15	158	59	129	103

B.3 Supplementary Methods

Below we have extended the modelling description provided in the main text to include more technical detail. The code used to implement these tasks is available at https://github.com/fergusjchadwick/COVID19_SyndromicRAT_public.

B.3.1 Modelling

Structure

We examined the ability of the two imperfect identification methods, syndromic modelling and rapid antigen testing (RAT), to predict the patient's COVID-19 status when used separately and together. These combinations define three model classes (Main Text Figure 4).

RAT-only uses only the RAT result. It equates being RAT-positive with the patient being PCR-positive for COVID-19 (hereafter, PCR-positive), and being RAT-negative with PCR-negativity.

Syndromic-only uses only the syndromic data. For this model, we used a Bayesian multivariate probit model.[1] The multivariate probit structures the outcomes of the PCR test and symptoms presence/absence as a D-dimensional vector of binary outcomes ($\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{id})$, $y_{ij} \in \{0, 1\}$). These outcomes are determined by an indicator function which takes a D-dimensional vector of *continuous latent* variables ($\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iD})$, $z_{ij} \in \mathbb{R}$). These latent continuous variables then covary as realisations of a D-dimensional multivariate normal, with the mean of the error structure informed by a linear predictor (in our case formed of the covariates age and gender), $\sum_{j=1}^{J} x_{ij}\beta_{jd} + \epsilon_{id}$, and a covariance ($\Sigma$) between dimensions. The linear predictor allows us to condition the outcomes on risk factor variables (here, age and gender). The covariance structure allows us to account for the correlated nature of the symptoms with each other and the outcome. This multivariate approach (multiple response variables) is also a very efficient way of encoding complex relationships between symptoms. These relationships need to be accounted for because symptoms are not simply additive in their predictive power. For example, in the diagnosis of measles the "Three C's" are used: cough, coryza (irritation and inflammation of the mucous membrane in the nose leading to head cold, fever, sneezing) and conjunctivitis. These symptoms individually, and in pairwise combination could be indicative of a wide range of diseases, but when all three are present measles is a highly probable cause (obviously, this is a simplified example conditioning on patient age and vaccination status). In the alternative, univariate approach, symptoms would be encoded as covariates in the linear predictor for PCR-status, and the complex relationships would need to be reflected as high-order interaction terms. These interaction terms use a large number of parameters and can be hard to fit to data. Using a multivariate structure allows us to exploit more efficient posterior sampling algorithms, and in higher dimensional settings like this uses fewer parameters.

The covariance matrix formulation of the model described above is not identifiable, because the variance, $diag(\Sigma)$ and means of the latent variables, $\boldsymbol{z}_i$ trade off against each other.[1] For this reason, we use a correlation matrix, Ω, formulation

with the variance set to 1. A correlation based framework also makes communication with clinicians and other practitioners smoother as correlations are more familiar. We thus formulate the multivariate probit as:

$$\begin{aligned} y_{id} &= \mathbb{I}(z_{id} > 0) \\ \boldsymbol{z}_i &= \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \\ z_{id} &= \sum_{j=1}^{J} x_{ij}\beta_{jd} + \epsilon_{id} \\ \boldsymbol{\epsilon}_i &\sim N(\mathbf{0}, \boldsymbol{\Omega}) \\ \Omega_{ii} &= 1 \\ \beta &\sim N(0, 1) \\ \boldsymbol{\Omega} &\sim \text{LKJ}(1) \end{aligned} \tag{B.1}$$

Syndromic-RAT Combined combines the two data sources. We utilise the specificity of RAT by treating RAT-positive patients as PCR-positive patients. The RAT-negative patients are modelled using the sensitive syndromic approach using Syndromic-only to capture PCR-positive patients that are missed by the RAT. This approach leverages the potential different syndromic profiles of PCR-positve patients who are RAT-positive and -negative, allowing the model to adapt solely to the latter. Structurally, the model combines RAT-only and Syndromic-only, with RAT-positive patients being modelled using RAT-only, and RAT-negative patients with Syndromic-only.

By using a Bayesian formulation, we generate full posteriors for our parameter estimates, allowing natural quantification of uncertainty. Bayesian methods also facilitate the use of more informative priors. We used minimally informative priors here. For covariate coefficients (betas) we used standard normals which are relatively flat in the probit scale. For the correlation prior, we used the Lewandowski-Kurowicka-Joe (LKJ) distribution, a covariance matrix prior with unit variance (i.e. a prior for correlation matrices). The LKJ distribution has a single parameter, η, which controls the degree of marginal correlation shrinkage. We used minimal shrinkage, $\eta = 1$)[2]. More informative priors that incorporate spatio-temporal effects, for instance, would be natural extensions. The models were fitted to the data using Bayesian inference techniques based on Hamiltonian Monte Carlo in the Stan programming language[3]. The models all converged with zero divergent transitions and large effective sample sizes.

Model Selection

We conducted backwards model selection (starting with the most complex, biologically plausible model) to identify a subset of models with the highest predictive power under temporal cross-validation (Main Text Figure 5). For the cross-validation, we divided the data into 5 folds of equal sizes in time order (i.e. the first fold is formed of the chronologically first $\frac{N}{K}$ patients, where N is the number of patients and K is the number of folds, the second fold by the next $\frac{N}{K}$ etc.) To test the sensitivity of this cross-validation structure, we also did a strict temporal division (i.e. the first $\frac{T}{K}$ days where T is the number of days samples were taken on). The results did not change qualitatively between these approaches.

The coarse round of model selection (Main Text Figure 5) selected candidate symptoms based on whether they had a strong and consistent correlation with PCR as estimated according to Equation (B.1). The models were fit with both covariates throughout the coarse round and symptoms were compared in nested models. In the fine round of model selection, these candidate symptoms and the covariate combinations (age and gender, age, gender and no covariates) were permuted to more exhaustively explore the model space. Reducing the number of possible models using the two stages of model selection was necessary to reduce computational demand and reduce the risk of overfitting models to the test scenarios. The large number of symptoms corresponds to a high number of potential model configurations (>131 000 for 14 symptoms and two covariates) which might perform well on the test sets (even under the challenging conditions of temporal cross-validation) but lack transferability.

By using general predictive power to narrow down the number of candidate models and then testing those models, we are more likely to choose models that generalise well to new data. It was clear when fitting the models that there were "jumps" in performance (as defined below) between models containing five and four symptoms, so the models with one to four symptoms were used as the candidate models. Zero symptom models were not included in the analysis as they do not correspond to a feasible policy (with covariates they would require governments to ask individuals of a given gender and age as COVID-19 positive, and without covariates they would involve randomly assigning individuals as COVID-19 positive).

Predictive Performance

We scored the models' predictive power using binary cross-entropy (hereafter, cross-entropy). Cross-entropy measures the accuracy of models that generate probabilities of binary outcomes, rather than make binary classifications, similar in concept to a mean square error for normally-distributed data, but adapted for binary data.[4] A cross-entropy value close to zero corresponds to high levels of accuracy, with larger values indicating lower accuracy. More specifically, the metric allows us to compare a binary vector, $\boldsymbol{y} \in [0, 1]$, with a vector of probabilistic predictions ($p(\boldsymbol{y}) \in (0, 1)$) as follows:

$$\boldsymbol{H}_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \tag{B.2}$$

The resulting score is comparable across all methods for assigning predictions where the same test data are used, allowing us to compare predictions from Model Classes 1-3. $H_p(q) \in 0, \boldsymbol{R}_+$ with zero indicating perfect prediction (assigning probabilities of ones and zeroes to outcomes of ones and zeros exactly) and larger values indicating worse predictions.

Classification Performance

In applied settings, models must often be evaluated on their performance as classifiers rather than just as prediction engines (i.e. their ability to say a patient is COVID-19 positive or negative, not simply the probability the patient might be

COVID-19 positive or negative). To generate a classification, $\hat{Y}$, a probability threshold, $\hat{p}$, must be chosen over which patients are classified as COVID-19 positive:

$$\hat{Y} = \begin{cases} 1, & \text{if } p(y) \geq \hat{p} \\ 0 & \text{otherwise} \end{cases} \tag{B.3}$$

Receiver operating characteristics (ROCs) are a way to measure the performance of a set of classifications in terms of true and false positives and negatives (TP, FP, TN and FN) and the rates of each of these classification types (e.g. $TPR = \frac{TP}{TP+FN}$, and $FPR = \frac{FP}{FP+TN}$). The error rates are calculated with respect to a particular threshold, $\hat{p}$, or across the range of possible $\hat{p}$s to generate a ROC curve. In our epidemiological scenarios (outlined below) we use our ROC curve calculations to identify single thresholds which yield a required error rate.

We strongly emphasise that generic performance here is only used to show the flexibility of the model classes; the best model for a local situation can only be determined if the relative cost of false positives and false negatives is known. Here, we choose three representative scenarios. Each scenario has a requirement and error rate (defined in Main Text Table 2). We identify the threshold, $\hat{p}$, at which the requirement is most closely exceeded (i.e. if the requirement were, hypothetically, that an error rate should be a maximum 15%, the threshold that produces an error rate below 15% but as close to 15% as possible will be chosen).

In Scenario 1, we do not consider epidemiological context but simply minimise false negative and false positive rates equally. We do this by maximising the two correct classification rates both individually and in total, as measured by the harmonic mean. The harmonic mean is used widely in the classification literature as it is maximised by achieving large values in all its component parts, rather than the arithmetic mean which can be maximised by having one extremely large component at the expense of other components. In other words, the arithmetic mean could be large because it has a very high TPR but a small TNR, whereas the harmonic mean will maximise both TPR and TNR. While conceptually the harmonic mean is better suited than the arithmetic for this use case, both produce qualitatively the same results for these data.

Scenario 2 corresponds to the situation in Bangladesh at time of writing (Sep. 2021), with COVID-19 cases beginning to rapidly increase again. Under these circumstances, false negatives are extremely costly relative to false positives due to the exponential growth of the disease.

In Scenario 3, the pandemic is not declining but maintaining a steady rate of cases. In this situation, policy-makers may be keen to keep false positive diagnoses low to prevent lockdown fatigue and to keep the workforce active.

The requirements in Scenario 2 and 3 were developed in discussion with the Institute of Epidemiology, Disease Control and Research (IEDCR), Bangladesh, for illustrative purposes.

References

[1] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, vol. 88, no. 422,

pp. 669–679, 1993.

[2] D. Lewandowski, D. Kurowicka, and H. Joe, "Generating random correlation matrices based on vines and extended onion method," *Journal of Multivariate Analysis*, vol. 100, no. 9, pp. 1989–2001, 2009.

[3] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of Statistical Software*, vol. 76, no. 1, 2017.

[4] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 359–378, 2007.