



Queensland University of Technology
Brisbane Australia

This may be the author's version of a work that was submitted/accepted for publication in the following source:

Bon, Joshua J, Bretherton, Adam, Buchhorn, Katie, Cramb, Susanna, Drovandi, Christopher, Hassan, Conor, Jenner, Adrianne L, Mayfield, Helen J., McGree, James M, Mengersen, Kerrie, Price, Aiden, Salomone, Robert, Santos-Fernandez, Edgar, Vercelloni, Julie, & Wang, Xiaoyu (2023)

Being Bayesian in the 2020s: opportunities and challenges in the practice of modern applied Bayesian statistics.

Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 381(2247), Article number: 20220156.

This file was downloaded from: <https://eprints.qut.edu.au/239007/>

© 2023 The Authors

This work is covered by copyright. Unless the document is being made available under a Creative Commons Licence, you must assume that re-use is limited to personal use and that permission from the copyright owner must be obtained for all other uses. If the document is available under a Creative Commons License (or other specified license) then refer to the Licence for details of permitted re-use. It is a condition of access that users recognise and abide by the legal requirements associated with these rights. If you believe that this work infringes copyright please provide details by email to qut.copyright@qut.edu.au

License: Creative Commons: Attribution 4.0

Notice: *Please note that this document may not be the Version of Record (i.e. published version) of the work. Author manuscript versions (as Submitted for peer review or as Accepted for publication after peer review) can be identified by an absence of publisher branding and/or typeset appearance. If there is any doubt, please refer to the published source.*

<https://doi.org/10.1098/rsta.2022.0156>

Discussion



Cite this article: Bon JJ *et al.* 2023 Being Bayesian in the 2020s: opportunities and challenges in the practice of modern applied Bayesian statistics. *Phil. Trans. R. Soc. A* **381**: 20220156.
<https://doi.org/10.1098/rsta.2022.0156>

Received: 22 August 2022

Accepted: 6 January 2023

One contribution of 16 to a theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

Subject Areas:
statistics

Keywords:
intelligent data collection, federated analysis, new data sources, implicit models, model transfer, Bayesian software products

Author for correspondence:

Robert Salomone


e-mail: robert.salomone@qut.edu.au

Being Bayesian in the 2020s: opportunities and challenges in the practice of modern applied Bayesian statistics

Joshua J. Bon^{1,2}, Adam Bretherton^{1,2}, Katie Buchhorn^{1,2}, Susanna Cramb^{1,4}, Christopher Drovandi^{1,2}, Conor Hassan^{1,2}, Adrienne L. Jenner^{1,2}, Helen J. Mayfield^{1,5}, James M. McGree^{1,2}, Kerrie Mengersen^{1,2}, Aiden Price^{1,2}, Robert Salomone^{1,3}, Edgar Santos-Fernandez^{1,2}, Julie Vercelloni^{1,2} and Xiaoyu Wang^{1,2}

¹Centre for Data Science, ²School of Mathematical Sciences, ³School of Computer Science, and ⁴School of Public Health and Social Work, Queensland University of Technology, Brisbane, Queensland, Australia

⁵School of Public Health, The University of Queensland, Saint Lucia, Queensland, Australia

 CD, 0000-0001-9222-8763; CH, 0000-0002-6200-2795; KM, 0000-0001-8625-9168

Building on a strong foundation of philosophy, theory, methods and computation over the past three decades, Bayesian approaches are now an integral part of the toolkit for most statisticians and data scientists. Whether they are dedicated Bayesians or opportunistic users, applied professionals can now reap many of the benefits afforded by the Bayesian paradigm. In this paper, we touch on six modern opportunities and challenges in applied Bayesian statistics: intelligent data collection, new data sources, federated analysis, inference for implicit models, model transfer and purposeful software products.

This article is part of the theme issue 'Bayesian inference: challenges, perspectives, and prospects'.

© 2023 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, provided the original author and source are credited.

1. Introduction

Bayesian data analysis is now an established part of the lexicon in contemporary applied statistics and machine learning. There is now a wealth of practical know-how to complement the continued development and increasing access to Bayesian models, algorithms and software. There is also a weighty body of published case studies that testify to the successful implementation and associated benefits of the Bayesian paradigm in practice. However, as with all fields of knowledge, the task is unfinished: each success begets further opportunities and challenges, which in turn drive new directions for innovation in research and practice. In this paper, we identify six such directions that, among many others, are driving the evolution of applied Bayesian modelling in this decade. For each of these, we provide a brief overview of the issue and a case study that outlines our experience in practice.

The first direction focuses on intelligent data collection: instead of collecting and analysing all possible data, or alternatively relying on traditional static experimental or survey designs, can we devise efficient, cost-effective approaches to collecting those data that will be most informative for the inferential purpose? In §2, authors Buchhorn and McGree focus on the opportunity to address this issue through Bayesian optimal experimental design. While there is an emerging literature on this approach in the context of clinical trials, they extend this attention to sampling designs for complex ecosystems. Furthermore, they address the challenge of exact implementation of the derived design in practice by introducing sampling windows in the optimal design. The new methodology and computational solution are illustrated in a case study of monitoring coral reefs.

Following from consideration of data collection, the second direction considered in this paper focuses on opportunities and challenges afforded through the emergence of new data sources. In §3, authors Price, Santos-Fernández and Vercelloni focus on two such sources: quantitative information elicited from subjects in virtual reality (VR) settings, and data provided by citizen scientists. Bayesian approaches to modelling and analysing these data can help to increase trust in these data and facilitate their inclusion in mainstream analyses. Some methods for achieving this are set in the context of two case studies based in the Antarctic and the Australian Great Barrier Reef.

The challenges of data collection are considered from a different direction in §4. Here, authors Hassan and Salomone reflect on the exponential rise in interest in federated analysis and learning. A canonical application of these approaches is the analysis of sensitive data from multiple data sources held by different data custodians, while leaving the data *in situ* and maintaining data privacy. The case study in this section focuses on federated learning with spatially dependent latent variables.

In §§5 and 6, we swing attention away from data to the models themselves. First, authors Drovandi, Jenner, Salomone and Wang consider the challenge of modelling increasingly complex systems via implicit models, i.e. models with intractable likelihoods that can nevertheless be simulated, and the opportunity afforded by likelihood-free algorithms such as sequential Monte Carlo-based approximate Bayesian computation (SMC-ABC). These approaches are applied to a substantive case study of calibrating a complex agent-based model (ABM) of tumour growth. In §6, another direction for modelling is discussed by authors Bon, Bretherton and Drovandi. This focuses on the challenge of transferring models developed in one context (dataset, location etc.) to another context. Fully Bayesian approaches to this challenge are still emerging and promise great opportunities in both research and practice.

The final direction we explore is in the translation of Bayesian practice to software products. We acknowledge the plethora of Bayesian packages embedded in software such as *R*, *Matlab* and *Python*, as well as stand-alone Bayesian products such as *BUGS*, *INLA* and *Stan*. These have revolutionized the practice of Bayesian data analysis and have placed this capability in the hands of applied researchers and practitioners. In §7, we focus on substantive software products created to support purposeful decision-making that are underpinned by Bayesian models. Author Mayfield describes a COVID-19 vaccine risk-benefit calculator (CoRiCAL) driven by a Bayesian network model; Vercelloni describes a platform for global monitoring of coral reefs (ReefCloud)

based on a Bayesian hierarchical model; and Cramb describes an interactive visualization of small area cancer incidence and survival across Australia (the Australian Cancer Atlas) based on a Bayesian spatial model.

2. Intelligent data collection

(a) Overview

The ability to determine and characterize underlying mechanisms in complex systems is paramount to pioneering research and scientific advancement in the modern era. Over the last decade, the rise of data generation from sensor and internet enabled devices has catalysed the advancement of data collection technologies and analysis methods used to extract meaningful information from complex systems. However, the sheer size of these complex systems (e.g. natural ecosystems like the Great Barrier Reef and river networks) and the expense of data collection means that data cannot be collected throughout the whole system. Further, practical constraints like connectivity, accessibility and data storage issues reduce our ability to sample frequently through time. This has led to innovation in statistical methods for data collection, promoting an emerging era of ‘intelligent data collection’ where data are collected for a particular purpose such as understanding mechanisms for change, monitoring biodiversity and identifying threats or vulnerabilities to long-term sustainability. Bayesian optimal experimental design is one such area of recent innovation.

Bayesian design offers a framework for optimizing the collection of data specified by a design \mathbf{d} for a particular experimental goal, which may be to increase precision of parameter estimates, maximize prediction accuracy and/or distinguish between competing models. More specifically, Bayesian design is concerned with maximizing an expected utility, $U(\mathbf{d}) = \mathbb{E}u(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y})$ through the choice of design \mathbf{d} within a design space \mathcal{D} , while accounting for uncertainty about, for example, the parameter $\boldsymbol{\theta} \in \Theta$ and all conceivable datasets we might observe $\mathbf{y} \in \mathcal{Y}$. A Bayesian optimal design \mathbf{d}^* can therefore be expressed as

$$\begin{aligned} \mathbf{d}^* &= \arg \max_{\mathbf{d} \in \mathcal{D}} \mathbb{E}u(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y}) \\ &= \arg \max_{\mathbf{d} \in \mathcal{D}} \int_{\mathcal{Y}} \int_{\Theta} u(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}, \mathbf{y}; \mathbf{d}) \, \mathrm{d}\boldsymbol{\theta} \, \mathrm{d}\mathbf{y}, \end{aligned}$$

where $p(\boldsymbol{\theta}, \mathbf{y}; \mathbf{d})$ defines the joint distribution of $\boldsymbol{\theta}$ and \mathbf{y} given a design \mathbf{d} .

Unfortunately, determining \mathbf{d}^* can be challenging. Firstly, the utility function $u(\mathbf{d}, \boldsymbol{\theta}, \mathbf{y})$ typically involves computing some form of expected value with respect to the posterior distribution, which itself is typically analytically intractable. Further, $U(\mathbf{d})$ is itself an expectation taken with respect to the prior-predictive distribution, which is also typically intractable. This means numerical or approximate methods are needed, which may impose substantial compute time and/or require a stochastic approximation. For example, Monte Carlo integration has been proposed as an approach to form an approximation to the expected utility as follows:

$$\mathbf{d}^* \approx \arg \max_{\mathbf{d} \in \mathcal{D}} \frac{1}{M} \sum_{m=1}^M u(\mathbf{d}, \boldsymbol{\theta}^{(m)}, \mathbf{y}^{(m)}), \quad (2.1)$$

where $\boldsymbol{\theta}^{(m)} \sim p(\boldsymbol{\theta}; \mathbf{d})$ and $\mathbf{y}^{(m)} \sim p(\mathbf{y}|\boldsymbol{\theta}^{(m)}; \mathbf{d})$, for some large value of M . Thus, computations involving M different individual posterior distributions are required just to approximate the expected utility of a design.

Secondly, \mathbf{d} may be high-dimensional, meaning that a potentially large optimization problem needs to be solved for a computationally expensive and noisy objective function. Accordingly, the majority of research in Bayesian design has focused on developing new methods to address one of these two challenges. Below we provide a brief summary of some of the relevant literature from the last 25 years.

Since the conception of statistical decision theory [1,2] upon which the decision-theoretic framework of Bayesian design is based [3], there have been numerous strategies presented in the literature to address the above challenges. Curve-fitting methods were proposed by Müller & Parmigiani [4] to the approximate expected utility in equation (2.1). Here, the fitted curve is optimized as a surrogate for the true expected utility to determine the choice of (approximately optimal) design. An alternative simulation-based method proposed by Müller [5] formed the following augmented joint distribution on \mathbf{d} , $\boldsymbol{\theta}$ and \mathbf{y} :

$$h_j(\mathbf{d}, \boldsymbol{\theta}_{1:j}, \mathbf{y}_{1:j}) \propto \prod_{j=1}^J u(\mathbf{d}, \boldsymbol{\theta}_j, \mathbf{y}_j) p(\mathbf{y}_j, \boldsymbol{\theta}_j; \mathbf{d}),$$

where it can be shown that the marginal distribution of \mathbf{d} is proportional to $U(\mathbf{d})$. Markov chain Monte Carlo (MCMC) methods were then used to sample from this distribution, and subsequently to approximate the marginal mode of \mathbf{d} . Extensions of this approach were given in [6,7] which include adopting a sequential Monte Carlo (SMC) algorithm to more efficiently sample from the augmented distribution as J increases. However, such approaches are limited to low-dimensional design problems (i.e. 3–4 design points) and simple models due to difficulties in sampling efficiently in high dimensions.

Recently, there has been a shift from sampling-based methods to rapid, approximate posterior inference methods. Combined with a Monte Carlo approximation as given in equation (2.1), this has enabled expected utility functions to be efficiently approximated for realistic design problems. This includes those based on complex models (such as nonlinear models) and models for data that exhibit complex dependence structures (such as those with different sources of variability including spatially and between groups). Such approximate inference methods include the Laplace approximation [8] and variational Bayes [9], which have been combined with new optimization algorithms (e.g. the approximate coordinate exchange algorithm; ACE [10]) to solve the most complex and high-dimensional design problems to date.

The most prominent application of Bayesian design methods appears in the clinical trial literature [11]. Recently, this has been exacerbated by the outbreak of COVID-19 where it has been desirable to conduct clinical trial assessments as quickly as possible, with Bayesian (adaptive) designs shown to yield more resource efficient and ethical clinical trials [12,13]. More recently, Bayesian design methods have been proposed as a basis to efficiently monitor large environmental systems like the Great Barrier Reef [14,15]. In the following case study, we show how such methods can be used to form sampling designs to monitor a coral reef system, and extend these methods to provide flexible designs that address major practical constraints when sampling real-world ecosystems.

(b) Case study: sampling windows for coral reef monitoring

Coral reefs are biodiversity hot spots for marine species under threat from anthropogenic impacts related to climate change, water pollution and over-exploitation, among other factors [16]. Coral cover is a commonly used indicator to infer the health of coral reef environments [17], where data collection relies on a series of images taken underwater, along a transect (a line across a habitat). Monitoring of coral reef environments is expensive in terms of monetary, human and technological costs, particularly for remote locations. Informative data are critical to support conservation decisions, but with limited resources to invest in monitoring programs, the need is to optimize in-field activities that will result in the intelligent collection of data. Following [18], we consider monitoring submerged shoals which are coral reefs that exist at depths of around 18–40 m below sea level. Data collection at such depths requires unmanned vehicles to be deployed along a design, i.e. a series of transects which specify where images should be collected. However, spatially precise sampling is known to be difficult in deeper reefs due to unpredictable weather and water currents. Therefore, our aim is to provide Bayesian designs that offer flexibility in

where transects will be placed while taking into consideration the complex nature of the systems we are monitoring such as the spatial dependence of natural processes.

In order to define a design, we specify the placement of each transect $k = 1, \dots, q$ on the shoal by its midpoint given in Easting and Northing coordinates, i.e. E_k and N_k , and the angle of the transect, α_k , in degrees. Each transect line is expressed as a design point $\mathbf{d}_k = (E_k, N_k, \alpha_k)$. The exact sampling locations (equally spaced along the fixed-length transect) are specified as s_i . For each transect, we introduce a radius parameter $r_k > 0$ for the purpose of allowing the sampled image locations to disperse by $\delta_1, \delta_2 \stackrel{\text{iid}}{\sim} \text{Unif}(-r_k, r_k)$, i.e. sampling at $\mathbf{s} + \boldsymbol{\delta}$. For image i , a number, n_i , of randomly selected points on the image are classified as either hard coral or not. Accordingly, the number of points within an image that contain hard coral, y_i , is modelled as

$$y_i | \boldsymbol{\beta}, Z_i \sim \text{Binomial}(n_i, \text{logit}^{-1}(\boldsymbol{\beta}^\top \mathbf{x}_i + Z_i))$$

$$\text{and } \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma(\boldsymbol{\gamma})),$$

for regression parameters $\boldsymbol{\beta} = (\beta_0, \dots, \beta_n)^\top$, covariance kernel parameters $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_m)^\top$ for spatially correlated random effect \mathbf{Z} , and covariates \mathbf{x}_i . The priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are based on consideration of historical data (depth and depth squared) collected on the shoal. See [18] for further details.

As a basis for improved monitoring, we consider the amount learned from the data regarding parameters of the above model as our goal of data collection. For this, we specify our utility function as the Kullback–Liebler divergence of the posterior from the prior distribution, where larger values suggest the data is more informative with respect to model parameters.

For a computationally efficient approximation of the utility for a given design, we employ a Laplace approximation of the posterior distribution, i.e. an approximation of the form

$$\mathcal{N}(\boldsymbol{\theta}^*, \text{H}(\boldsymbol{\theta}^*)^{-1}),$$

where $\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta} \in \Theta} \log p(\mathbf{y}, \boldsymbol{\theta}; \mathbf{d})$ and $\text{H}(\boldsymbol{\theta}^*)$ is the Hessian matrix evaluated at $\boldsymbol{\theta}^*$. Here $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$, and marginalization of \mathbf{Z} is performed approximately using Monte Carlo integration. To obtain an optimal design to for monitoring of the shoal, we propose a two-step approach

- (i) Firstly, a global search for the Bayesian optimal design $\mathbf{d}^* = (\mathbf{d}_1^*, \dots, \mathbf{d}_q^*)$, where $q = 3$ (the total number of transects) is conducted. We consider a discretized design space, and find designs via a discrete version of ACE; and
- (ii) Secondly, we form design efficiency windows (illustrating robustness to imprecise sampling) across r_k for each transect $k = 1, \dots, q$. To do so, we specify a zero-mean Gaussian process (GP) prior for the approximate expected utility across $\mathbf{r} \in \mathbb{R}^q$ by $\tilde{U}(\mathbf{r}; \mathbf{d}^*)$, i.e. $\tilde{U}(\mathbf{r}; \mathbf{d}^*) \sim \mathcal{GP}(\mathbf{0}, \text{K}(\cdot) + \zeta_0 \text{I})$, for some kernel matrix $\text{K}(\cdot)$, and $\zeta_0 > 0$. The windows are then obtained as follows:
 - (a) Centre the radius on \mathbf{d}^* , i.e. the Bayesian design from (i), and specify a maximum value for r_k for $k = 1, \dots, q$;
 - (b) Randomly sample $\delta_{i,1}, \delta_{i,2} \stackrel{\text{iid}}{\sim} \text{Unif}(-r_k, r_k)$, where k is the transect from which image i is obtained, and evaluate the approximate expected utility of the design at locations $s_i + \boldsymbol{\delta}_i$;
 - (c) Fit a GP defined on $\mathbf{r} \in \mathbb{R}^q$ to the approximate expected utilities;
 - (d) Emulate the expected utility surface across values of \mathbf{r} using the posterior predictive mean of the GP, denoted $\tilde{U}(\mathbf{r})$;
 - (e) Normalize the predicted expected utility values by that of the original Bayesian design as follows:

$$\text{eff}(\mathbf{r}) = \frac{\tilde{U}(\mathbf{r}; \mathbf{d}^*)}{\tilde{U}(\mathbf{0}; \mathbf{d}^*)}, \quad (2.2)$$

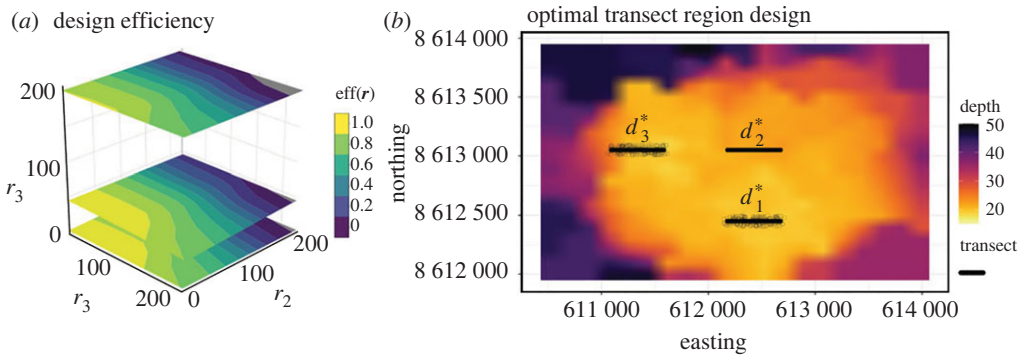


Figure 1. The Bayesian design across the Barracouta East coral shoal, $\mathbf{d}^* = (d_1^*, d_2^*, d_3^*)$, are illustrated as black transect lines (b). Sampling windows are formed around these transects allowing for flexibility in sampling locations while retaining 0.99 of the optimal utility. Design efficiency contours across $\mathbf{r} \in \mathbb{R}^q$ are shown (a). (Online version in colour.)

and use the above to obtain design efficiency contours (plotted in figure 1a). For some design efficiency contour value $c > 0$, the corresponding sampling window is the region in space defined by radii $\mathbf{r}(c)$ that satisfy $\text{eff}(\mathbf{r}(c)) = c$.

Based on the approach, the Bayesian design, \mathbf{d}^* , shown in figure 1, situates the transects in shallower areas of the reef, but at different depths in the shallow areas, presumably to provide information about the depth effects, β . Avoiding the deeper regions of the shoal makes sense physiologically, as the corals monitored here are photosynthetic organisms, and therefore rely on light to survive. This design thus avoids the collection of data in areas where there is little chance of observing coral. The design efficiency contours are also shown in figure 1. If we consider a design efficiency of 0.99 in equation (2.2), then possible radius values are (50, 0, 44) for the three transects, with the flexibility (sampling windows) this provides shown around each transect. As can be seen, transect d_2^* is more sensitive than the other two, suggesting more effort should be placed in sampling this transect precisely. In practical terms, sampling from shallow areas of the reef, d_1^* and d_3^* , can be undertaken when the conditions are more unpredictable (e.g. strong currents), and samples from d_2^* can be obtained when field conditions are more preferable.

In conclusion, Bayesian optimal design addresses a fundamental problem in science: the intelligent collection of data resulting in greater information efficiency, reduced sampling cost and improved estimation. Such benefits have been observed in clinical trials [19–22] and environmental monitoring [15,23], and we have shown how they can be used to offer flexible yet efficient sampling in a real-world context. One limitation of the approach is the potential reliance of designs on a number of assumptions, e.g. an assumed model for the data, so we would encourage future research in areas that reduce this reliance and thus provide more robust designs for data collection.

3. New data sources

As part of the digital revolution, data from new types of technologies (e.g. VR technology, satellite imagery and *in situ* sensor data) are becoming available, providing opportunities to gain insights into challenging applied research areas such as environmental conservation. In this section, we describe new sources of data arising from subject elicitation using VR and citizen science (CS), as well as illustrating how Bayesian modelling can be applied to such data for the purposes of informing management decisions in Antarctica and the Australian Great Barrier Reef.

(a) Elicitation using virtual reality

Recent advancements in digital technologies have led to the large-scale collection of more advanced data such as sensor data, satellite imagery, and a host of varied resolution imagery and video including those taken using 360° cameras. It is possible, using these advances in technology, to enable location-specific data to remote researchers for analysis. The emergence of VR technology, for example, acts as a way to connect the public and the scientific community, creating innovative pathways for environmental conservation research by immersing subjects in an otherwise inaccessible vivid virtual scene for elicitation purposes [24,25]. The opinions and knowledge extracted from this process is itself new data, which can be used for educational purposes [26] or incorporated into statistical models.

Increases in the volume of more complex types of data has led to the development of more effective and efficient analysis methodology. Recently, Bayesian models have seen use as a method to evaluate subject elicitation in the areas of coral reef conservation [27], jaguar and koala habitat suitability assessments [28,29], and the aesthetic value of sites in the Antarctic Peninsula.

(i) Case study: quantifying aesthetics of tourist landing sites in the Antarctic Peninsula

In the Antarctic Peninsula, the effects of climate change and the associated increase of ice-free areas are threatening the fragile terrestrial biodiversity [30]. As well as high ecological importance, these ecosystems also have a unique aesthetic value which has been formally recognized in Article 3 of the Protocol on Environmental Protection to the Antarctic Treaty [31]. There is value in protecting beautiful landscapes, as tourism in Antarctica is based largely on the natural beauty of the environment. This case study quantifies aesthetic values in the Antarctic Peninsula by recording elicitation from subjects immersed in a VR environment using a state-of-the-art web-based framework *R2VR* [32].

Subject elicitation in this case study is drawn from 16 photos, obtained via 360° photography at tourist landing sites in the Antarctic Peninsula. Consultation produced landscape characteristics of interest, e.g. the presence of certain animals and the weather. These characteristics and images were then used to construct an interview, to be held while the subject was immersed in the VR environment, with responses recorded on the Likert scale, from strongly disagree to strongly agree. From this elicitation process, responses to each question are recorded for each scene presented to the participant, as well as their opinion of the aesthetic value of the scene itself. Additionally, general participant characteristics such as gender identity and age are also recorded.

A Bayesian hierarchical model is used for modelling the response of whether or not a subject i determines scene j ($j = 1, \dots, o$) as aesthetically pleasing (y_{ij}) as a function of responses to statements such as 'there are animals in this image' and 'this image is monotonous' ($x_{ik}, k = 1, \dots, m$), subject characteristics such as age and gender ($x_{ih}, h = m, \dots, m + n$), and subject-reported confidence in their response to each interview statement ($s_{ij}, j = 1, \dots, m$), where zero represents low confidence and one represents high confidence. The model is

$$\begin{aligned}
 y_{ij} | \alpha_j, \beta_{0s_{ij}}, \beta_1, & \stackrel{\text{ind}}{\sim} \text{Bernoulli} \left(\text{logit}^{-1}(\alpha_j + (\beta_{0s_{ij}}^\top, \beta_1^\top)^\top x_i) \right) \\
 \alpha | \tau_\alpha & \sim \mathcal{N}(\mathbf{0}, \tau_\alpha^{-1} \mathbf{I}_o), \\
 \beta_{0l} | \mu, \tau_l & \stackrel{\text{ind}}{\sim} \mathcal{N} \left(\mu, \text{diag}(\tau_l^{-1}) \right), \quad l = 0, 1, \\
 \beta_1 & \sim \mathcal{N}(\mathbf{0}, 10^2 \mathbf{I}_n), \\
 \tau_{lk}, \tau_\alpha & \stackrel{\text{iid}}{\sim} \text{Gamma}(10^{-2}, 10^{-2}), \quad k = 1, \dots, m, \quad l = 0, 1 \\
 \text{and} \quad \mu & \sim \mathcal{N}(\mathbf{0}, 10^2 \mathbf{I}_m).
 \end{aligned}$$

The development of conservation plans should, in accordance with the Protocol on Environmental Protection to the Antarctic Treaty, include recommendations based on aesthetic

value. This case study is among the first to propose the incorporation of aesthetic value into conservation plans by leveraging subject-reported uncertainty. Understanding aesthetic attributes in Antarctica can be applied to other regions, especially through the implementation of similar surveys and models. The landscape of VR data assets continues to expand as more researchers are made aware of the value added to methods of subject inquiry by including multi-modal features such as text, sounds and haptic feedback. Modern Bayesian modelling approaches allow insights to be drawn from these such novel approaches to data collection.

(b) Citizen science

CS represents one of the most popular emerging data sources in scientific research. CS involves engaging members of the general population in one or more parts of the scientific process. Its applications can be found across almost all disciplines of science, especially in ecology and conservation where scientists are harnessing its power to help solve critical challenges such as climate change and the decline in species abundance. Examples of citizen scientists' contributions include reporting sightings of species, measuring environmental variables and identifying species on images. Hundreds of CS projects can be found in popular online platforms including Zooniverse [33], eButterfly [34], eBird [35] and iNaturalist [36]. A fundamental issue often discussed surrounding CS is the quality of the data produced, which is generally error-prone and biased. For example, bias can arise in CS datasets due to (i) the unstructured nature of the data, (ii) collecting data opportunistically, with more observations from frequently visited locations [37] or at irregular frequencies across time [38], and (iii) as a result of differing abilities of the participants to perform tasks such as detecting or identifying species [39,40]. However, recent advances in statistics, machine learning and data science are helping realize its full potential and increase trustworthiness [40–42].

Frequently, CS data are elicited via image classification. For example, asking the participants whether images contain a target class or species. In this section, we illustrate two modelling approaches for these types of data.

In the first approach, we consider a binary response variable y_{ij} representing whether the category has been correctly identified by the participant ($i = 1, \dots, m$) in the image ($j = 1, \dots, n$). The probability of obtaining a correct answer can be modelled using an item response model such as the three-parameter logistic model (3PL) [40,43],

$$y_{ij}|Z_i, B_j, \eta_j, \alpha_j \sim \text{Bernoulli}(\eta_j + (1 - \eta_j)\text{logit}^{-1}(\alpha_j(Z_i - B_j))), \quad (3.1)$$

where each $\eta_j \in (0, 1)$ is a pseudo-guessing parameter accounting for a participants' chance of answering correctly by guessing, Z_i is the latent ability of the i th participant, $\alpha_j > 0$ is the slope parameter and B_j is the latent difficulty of the j th image. Sometimes, the correct answer for certain images is unknown. In this case, we estimate the latent labels for the images via the estimates of Z_i , by using the latter as weights in popular methods such as majority or consensus voting. Code to fit these models and exemplar datasets can be found in [44].

The second approach is for the case where we are interested in the proportion of species in elicitation points in images. Here, we compute a statistic $\hat{y}_{ij} \in [0, 1]$ giving the apparent proportion of species in a number of elicitation points in image j classified by the participant i . The true latent proportion Y_j can be estimated based on each participant's overall performance measures se_i and sp_i (which denote the sensitivity and specificity scores of participant i , respectively). A Beta prior is placed on the true proportion, yielding the model

$$\begin{aligned} \hat{y}_{ij} &= Y_j se_i + (1 - Y_j)(1 - sp_i), \\ Y_j &\sim \text{Beta}(\alpha_j, \beta_j), \end{aligned} \quad (3.2)$$

where α_j and β_j are the shape and the scale parameters in the beta distribution, respectively. The above model can be parametrized via a specified prior mean μ_j for each Y_j , and a common precision parameter ϕ , via $\alpha_j = \mu_j\phi$ and $\beta_j = -\mu_j\phi + \phi$, which in turn implies that

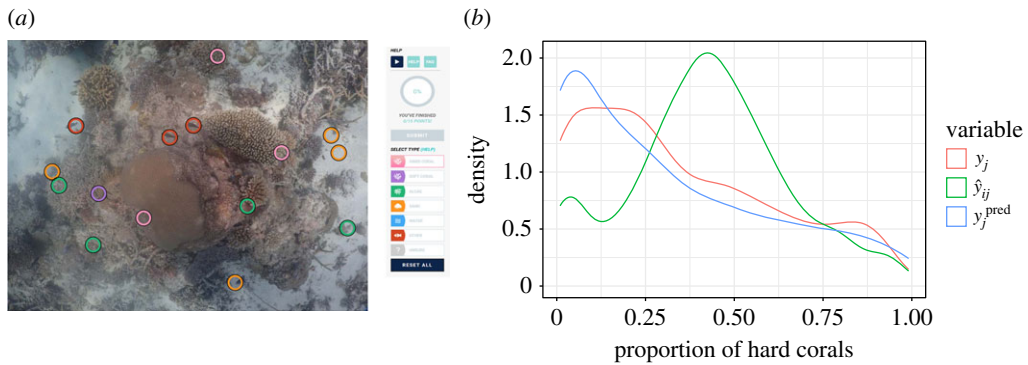


Figure 2. (a) Elicited points with benthic categories in an underwater image from Great Barrier Reef, Australia. (b) True latent proportion (in red) and the apparent proportion of hard corals (in green). The predicted proportion is represented in blue. (Online version in colour.)

$\text{Var}[Y_j] = \mu_j(1 - \mu_j)/(1 + \phi)$. Covariates can also be incorporated by defining a beta regression with $\text{logit}(\mu_j) = \xi^T x_j + U_j + \varepsilon_j$, where ε_j are error terms, and U_j are spatially dependent random effects. Both approaches account for spatial variation (captured in B_j or U_j for the first and second approach, respectively) using different spatial structures (e.g. conditional autoregressive (CAR) priors, covariance matrices, or Gaussian random fields). See more details in [40,42,45].

The following case study illustrates the estimation of the latent proportion of hard corals across the Great Barrier Reef in Australia, obtained from underwater images classified by citizen scientists. Figure 2 shows 15 spatially balanced random points in one of the images used in the study. The apparent proportion of hard coral in the image was obtained using the number of points selected by participants containing this category out of 15. Using equation (3.2), the (biased) estimates obtained from the citizen scientists can be corrected producing a similar density to the latent unobserved proportions.

The integration of CS data with current monitoring efforts from Australian federal agencies and non-governmental organizations is a breakthrough to increase the amount of information about changes along the Great Barrier Reef, learn about climate change impacts and adapt management actions consequently. This model introduced here is the root of a digital platform that estimates the health of the Great Barrier Reef using all available information. This study contributes to increasing the trust of CS and produce reliable data for environmental conservation while engaging and arising awareness about coral reefs.

4. Federated analyses and inference methods

(a) Overview

In many areas of study (health, business and environmental science, for example), the collective dataset of interest one wishes to use in modelling is often under the control of different data custodians, i.e. parties responsible for ensuring data is only used or released in instances deemed appropriate to governance requirements. Such requirements often stipulate that the data itself, and information pertaining to it can only be shared in a manner deemed sufficiently private.

Federated learning is the process of fitting a model in the setting where data resides with multiple custodians. Approaches typically place privacy as having the utmost importance, but computational efficiency is important from a practical perspective. There are two broad data settings that occur, each requiring their own style of algorithm. The first is the horizontal setting. Here, multiple data custodians have the same set of variables for different entities. By contrast, the vertical federated learning setting has data custodians who possess different variables for the

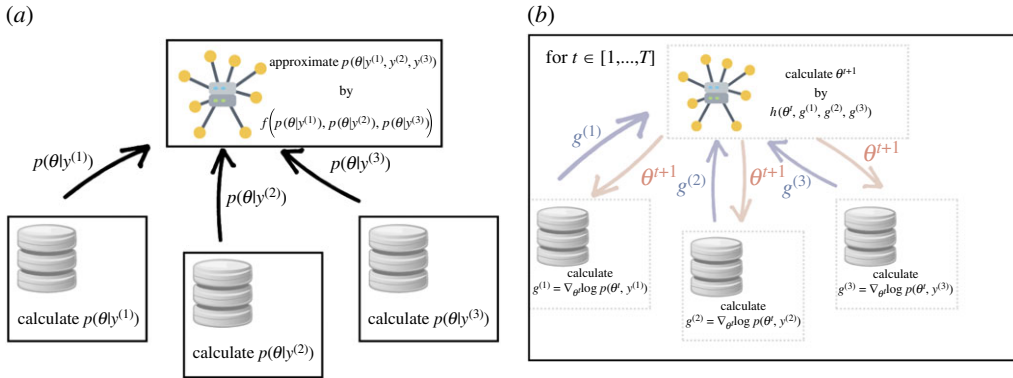


Figure 3. Federated approaches lie on a continuum between *post hoc* posterior amalgamation approaches as used in certain distributed MCMC approaches (a) and collaborative multi-round approaches (b). (Online version in colour.)

same entities. An example of a horizontal setting is where different countries possess the data for those primarily residing within. By contrast, an example of vertical federated learning would be where two companies possess their respective sales data for the same collection of customers. The term ‘federated learning’ originated in the deep learning literature with the introduction of the *FedAvg* algorithm [46]. *FedAvg* involves updating parameter values of a global model to be the weighted average of parameter values obtained by updating the same model locally (possibly many times) at each iteration. This work led to many related optimization algorithms, e.g. *FedProx* [47], and *FedNova* [48] which account for heterogeneous (non i.i.d.) data sources, and the Bayesian nonparametric approach for learning neural networks of [49], where local model parameters are matched to a global model via the posterior distribution of a Beta-Bernoulli process [50]. To date, practical federated analyses appear restricted to the frequentist setting. Examples include the prediction of breast cancer using distributed logistic regression [51] and modelling of the survival of oral cavity cancer through a distributed proportional Cox hazards model [52]. Both these approaches conduct parameter estimation via a Newton–Raphson algorithm [53] and result in equivalent maximum-likelihood estimates to those obtained in a standard, non-federated setting. Algorithms for the maximum-likelihood estimation of log-linear and logistic regression models in vertical federated learning settings [54–58] use ideas such as secure multiparty computation [59], and formulating the parameter estimation task as a dual problem [60]. Several overarching software infrastructures such as *VANTAGE6* [61] ensure the correct and secure use of the data of each custodian within the specified algorithm, given acceptable (model- and application-specific) rules for information exchange.

Despite the potentially enabling capabilities of federated methods, to our knowledge, Bayesian federated learning methods have yet to impact real-world applications. In the Bayesian inference setting, the ‘learning’ task becomes one of performing posterior inference, e.g. via MCMC or variational inference techniques. Note that Bayesian federated learning approaches may involve multiple communication rounds, though this is only sometimes the case. For example, many distributed MCMC approaches (e.g. [62–64]), combine individually-fit model posteriors, requiring only a single communication step from each local node. A recent intermediate approach [65] is to construct a surrogate likelihood of the complete dataset via an arbitrarily specified number of communication steps. After constructing the surrogate likelihood, an MCMC algorithm is run on a single device. As the number of communication steps increases, the approximation error introduced by the surrogate likelihood decreases. Figure 3 illustrates the difference between *post hoc* posterior amalgamation strategies and collaborative multi-round approaches.

In certain cases, carrying out federated Bayesian inference is (at least in principle) relatively straightforward. For example, a naive MCMC algorithm would be trivial to construct for a simple

model class, such as any generalized linear model (which assumes the data are independent), provided that one is not concerned with the number of communication steps. To see this, note that the (log-)posterior density function decomposes as

$$\log p(\boldsymbol{\theta}|\mathbf{y}) = \log p(\boldsymbol{\theta}) + \sum_{k=1}^n \log p(y_k|\boldsymbol{\theta}) + \text{const.} \quad (4.1)$$

Hence, for the horizontal setting, all that is required is the nodes sharing the sum of their respective log-likelihood terms with the server. However, this approach would require a minimum of two communication steps per iteration of the Markov chain. Recent MCMC methods, similar in style to the *FedAvg* algorithm (which use Langevin dynamics to update the Markov chain), require only a single communication step per iteration [66,67]. Such approaches exploit gradient information which decomposes as a sum similarly to (4.1), though eschew the usual Metropolis–Hastings correction and are hence asymptotically inexact. In some instances, a formally justified notion of privacy may be required, as opposed to simply an intuitive one given by aggregation of terms. Differential privacy (DP) (e.g. [68]) provides such guarantees, and there are variants of MCMC that ensure this, such as DP-MCMC [69], which accomplishes privacy guarantees at the cost of a slight perturbation of stationary distribution of the chain. It is worth noting that all of the above examples mentioned are specific to the horizontal setting, with the vertical setting proving especially challenging as one does not have a beneficial decomposition like that of (4.1).

As the above alludes to, the development and use of Bayesian federated learning algorithms are complex for several reasons. A method is only suitable for a prescribed application if it satisfies a combination of requirements, such as being able to work with the desired model, computational and communication costs, privacy and accuracy. For each application, the choice of model and federated method will depend on where the priorities lie, e.g. accuracy, efficiency or privacy. In some cases, there may be no feasible algorithm (an example is given in the upcoming case study). Thus, inference approaches that improve upon some (or even all) of these aspects are important and warrant future research.

The ultimate goal of federated Bayesian analysis is to circumvent the need for data merging [70,71] in scenarios where merging is considered infeasible. However, for Bayesian federated learning to reach this point, these approaches must offer custodians and interested parties an accurate inference for complex models while maintaining a level of privacy acceptable to those data custodians. Thus, the methodological development that enables federated inference for more advanced Bayesian models efficiently and/or with additional privacy guarantees is likely to emerge as a critical area of interest in the coming years.

(b) Case study: federated learning with spatially dependent latent variables

The greatest hindrance to employing federated learning in real-world applications is the lack of possible model types that current algorithms address. Commonly, applied statistical modelling involves incorporating hierarchical structures, and latent variables [72]. To our knowledge, there are no federated Bayesian analysis algorithms *at all* for such models. To briefly illustrate the unique challenges and the need for developments that account for the nuances of different models, the case study considers spatially dependent latent variables based on neighbourhood structures. For simplicity, the focus is on the Intrinsic Conditional AutoRegressive (ICAR) prior [73], although variations such as the Besag–York–Mollie [74] and Leroux [75] models are similar in what follows (the latter is used for example, in the Australian Cancer Atlas described in §7). The ICAR prior posits a vector of spatially dependent latent variables, denoted here as \mathbf{Z} . Each element of \mathbf{Z} corresponds to a latent area-level effect of a ‘site’, which is influenced by neighbouring sites. Writing $i \sim j$ to denote that sites i and j are considered neighbours, and assuming the graph arising from the neighbourhood structure is fully connected, the ICAR prior with precision

hyperparameter τ has log-density

$$\log p(z; \tau) = \frac{n}{2} \log \tau - \frac{\tau}{2} \sum_{i \sim j} (z_i - z_j)^2 + \text{const.}$$

The above may be problematic if the data custodians insist that the latent variables corresponding to their areas must be kept private to themselves. To see why, consider the case that there are two (2) data custodians, with the sets \mathcal{C}_1 and \mathcal{C}_2 containing the indices of data possessed by the first and second custodian, respectively. Then,

$$\sum_{i \sim j} (z_i - z_j)^2 = \sum_{i,j \in \mathcal{C}_1: i \sim j} (z_i - z_j)^2 + \sum_{i,j \in \mathcal{C}_2: i \sim j} (z_i - z_j)^2 + \sum_{i \in \mathcal{C}_1, j \in \mathcal{C}_2: i \sim j} (z_i - z_j)^2, \quad (4.2)$$

where terms in blue are those relevant to the sites under the first custodian, and those in red to the second. When computing the log-posterior density (as required, for example, in Markov chain Monte Carlo algorithms), the first two terms on the right-hand side above can be aggregated and sent to the central server. However, the final term cannot as each individual summand requires the individual latent variables to be processed. This is because the latter term considers interactions across custodian boundaries.

Consequently, solutions such as (i) employing judicious reparameterization of the latent variables (possibly compatible with the one that is often also required to enforce identifiability), (ii) changing the model to add additional auxiliary variables or (iii) otherwise approximating the troublesome term, are required. An additional challenge is that even if inferences on individual latent variables are only available to their respective custodians, they may nevertheless ‘leak’ information across custodian boundaries to neighbouring sites due to the underlying dependency structure.

While the above certainly highlights particular challenges, the first two terms of (4.2) split nicely across custodians and hint that latent variables need not always be problematic. For more straightforward cases such as the latter, specialized accurate and efficient inference approaches that allow individual custodians to avoid ever sharing their latent variables (either directly or indirectly) or data are the subject of forthcoming work by the authors of this section, who have a longer-term goal of tackling more challenging cases such as ICAR and its relatives in different settings.

5. Bayesian inference for implicit models

(a) Overview

The appetite for developing more realistic data-driven models of complex systems is continually rising. Development of such complex models can improve our understanding of the underlying mechanisms driving real phenomena and produce more accurate predictions. However, the calibration of such models remains challenging, as the associated likelihood function with complex models is often too computationally cumbersome to permit timely statistical inferences. Despite this, it is often the case that simulating the model is orders of magnitude faster than evaluating the model’s likelihood function. Such models with intractable likelihoods that nevertheless remain simulable are often referred to as implicit models. Such models are now prevalent across many areas of science (see e.g. various application chapters in [76]).

Currently, the most popular statistical approach amongst practitioners for performing Bayesian inference for implicit models is approximate Bayesian computation (ABC), popularized by Beaumont *et al.* [77]. A related method called generalized likelihood uncertainty estimation [78,79] predates ABC, and [80] explore its connections to ABC. The ABC approach approximates the true posterior as

$$p_\epsilon(\boldsymbol{\theta} | \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int p(\mathbf{x} | \boldsymbol{\theta}) \mathbb{I}(\|\mathbf{x} - \mathbf{y}\| \leq \epsilon) \, \mathrm{d}\mathbf{x}. \quad (5.1)$$

Here, \mathbf{x} denotes simulated data that has the same structure as \mathbf{y} , $\|\cdot\|$ is some norm (i.e. $\|\mathbf{x} - \mathbf{y}\|$ measures the closeness of the simulated data to the observed data), and ϵ stipulates what is considered ‘close’. Intuitively, values of θ more likely to produce simulated data \mathbf{x} close enough to \mathbf{y} have increased (approximate) posterior density. Rather than compare \mathbf{y} and \mathbf{x} directly, it can be more efficient to compare \mathbf{y} and \mathbf{x} in a lower dimensional space via a summarization function that aims to retain as much information from the full dataset as possible. For the posterior in (5.1) to equate to the exact posterior, we require that the observed and simulated datasets are matched perfectly (i.e. as $\epsilon \rightarrow 0$) in terms of some sufficient summarization. However, in the majority of practical applications, a low-dimensional sufficient statistic does not exist and it is computationally infeasible to take $\epsilon \rightarrow 0$, so we must accept some level of approximation.

Given the wide applicability of the approach, i.e. that only the ability to simulate the model is required to conduct inference, there has been an explosion of research in the past 10–15 years advancing ABC and related methods that lie within the more general class of so-called likelihood-free inference methods. A substantial portion of methodologically focused ABC research considers aspects including the effective choice of $\|\cdot\|$ (e.g. [81,82]), efficient sampling algorithms to explore the approximate posterior in (5.1) (e.g. [83]) and ABC’s theoretical properties (e.g. [84]). Many of the developments of ABC and some related methods (e.g. Bayesian synthetic likelihood [85,86]) prior to 2018 are discussed in [76], the first-ever monograph on ABC.

The following case study considers a popular class of sampling algorithms for ABC based on SMC. SMC-based ABC algorithms improve efficiency compared to sampling naively from the prior by gradually reducing the ABC tolerance ϵ where the output produced at iteration t is used to improve the proposal distribution of θ at iteration $t + 1$. The output of the algorithm is N samples, or ‘particles’, from the ABC posterior in (5.1) with a final ϵ that is either pre-specified or determined adaptively by the algorithm. Each particle has attached to it a ‘distance’, which is the value of $\|\mathbf{x} - \mathbf{y}\|$ for \mathbf{x} simulated from the model based on the particle’s parameter value. Here, we use the adaptive SMC-ABC algorithm in [87], itself a minor modification of the replenishment algorithm of [88]. The algorithm is summarized below.

- (i) Draw N samples from the prior, and for each sample, simulate the model and compute the corresponding distance. Initialize ϵ as the largest distance among the set of particles.
- (ii) Set the next ϵ as the α -quantile of the set of distances. Retain the N_α particles with distance less than or equal to ϵ .
- (iii) Resample the retained particle set $N - N_\alpha$ times so that there are N particles.
- (iv) Run MCMC on each of the resampled $N - N_\alpha$ particles with stationary distribution (5.1) with the current ϵ . This step helps to remove duplicate particles created from the previous resampling step. The number of MCMC iterations can be adaptively set based on the MCMC acceptance rate.
- (v) Repeat steps (ii)–(iv) until a desired ϵ is reached or the MCMC acceptance rate in step (iv) is too small (i.e. the number of MCMC steps becomes too large for the computational budget).

A key computational inefficiency of ABC and closely related methods such as BSL is that many of the model simulations yield MCMC proposals that are rejected. To obtain a suitable quality of approximation, it is not uncommon to require continuing the algorithm past the point where ϵ is small enough to have average acceptance probabilities of 10^{-2} or less. To overcome this issue, there has been significant attention devoted to machine learning based approaches to likelihood-free inference, especially in the past 5 years. These methods use model simulations (from different parameter values) as training data for building a conditional density estimator of the likelihood (e.g. [89]), likelihood ratio (e.g. [90]) or posterior density (e.g. [91]). Following this estimation, standard methods from the Bayesian inference toolkit can be used. Many machine learning approaches to likelihood-free inference can be implemented sequentially, so that samples from the approximate posterior in the previous (or all previous) iterations can comprise an increasingly

informed training set that yields a more accurate conditional density estimator in regions of non-negligible posterior probability. For the case study below, we compare the SMC-ABC approach with the sequential neural likelihood (SNL) method of [89], which is outlined below.

- (i) Set the initial proposal distribution of parameter values as the prior, i.e. $q(\theta) = p(\theta)$.
- (ii) Generate a training dataset by drawing M parameter/simulated data pairs according to $q(\theta)p(x|\theta)$. Fit a conditional normalizing flow (a flexible type of regression-density estimator) to the training data to estimate the conditional density of $\mathbf{X}|\theta$.
- (iii) Run MCMC to obtain approximate posterior samples, using the learned conditional density of $\mathbf{X}|\theta$ evaluated at the observed data \mathbf{y} as the approximation to the likelihood. Samples from this approximate posterior can also be used to update the proposal distribution $q(\theta)$.
- (iv) Repeat steps (ii) and (iii) for a desired number of rounds.

(b) Case study: calibrating agent-based models of tumour growth

In this case study, we apply likelihood-free methods SMC-ABC and SNL for calibrating a complex agent-based model (ABM) of tumour growth. We briefly compare the methods in terms of computational efficiency and their ability to fit simulated and real tumour growth data.

ABMs have been used in cancer modelling for some time now as they provide a spatial representation of the inherent cellular heterogeneity and stochasticity of tumours [92–94]. Largely, these models account for the individual cell-based behaviours of proliferation, movement and death and aim to predict the impact of stochasticity on spatial tumour growth over time. Previous works have considered this in the context of angiogenesis [95], immune involvement [96,97] and also treatment [98]. In some cases, data have been used to calibrate or validate aspects of the models [99,100], although due to the computational cost and their intractable likelihood it is not always easy to infer parameters in an ABM using data.

For this case study, we use a previously published ABM called a Voronoi cell-based model (VCBM) [101,102]. In this model, cancer cells and healthy tissue cells are considered agents, whose centre is modelled by a point on a two-dimensional lattice, and whose boundary is defined by a Voronoi tessellation. To mimic tumour growth and spatial tissue deformation, the model captures cell movement using force-balance equations derived from Hooke's Law. In this way, cell movement is captured off-lattice and is a function of the local cell-neighbourhood pressure, determined using a Delaunay Triangulation.

Tumour growth is captured by introducing a probability of an individual cancer cell proliferating P , which is a function of a cell's distance to the boundary of the tumour: $P = p_0(1 - (d/d_{\max}))$, where p_0 is the probability of proliferation, d is the cell's Euclidean distance to the tumour boundary (measured from the cell centre to the nearest healthy cell centre) and d_{\max} is the maximum radial distance a cell can be from the boundary and still proliferate. In this way, the model evolves stochastically over time with cells either proliferating or moving in a given timestep. The model also uses g_{age} to define the time taken for a cell to be able to proliferate and uses p_{psc} as the probability of cancer cell invasion. Hence, the model parameter θ to be estimated is $(p_0, p_{\text{psc}}, d_{\max}, g_{\text{age}})$.

To validate the VCBM, we use published *in vivo* tumour growth measurements for ovarian cancer [103]. In these experiments, tumour volume was recorded by measuring the tumour width and length as perpendicular axis using calipers and then calculating the approximate tumour volume. We simulate the VCBM in two dimensions and calculate the corresponding tumour volume measurements equivalently. We also consider one simulated dataset generated with parameter value $\theta = (0.2, 10^{-5}, 31, 114)$. The datasets are shown as solid black lines in figure 4. The prior distribution on θ is given by $p_0 \sim \text{Beta}(1, 1)$, $p_{\text{psc}} \sim \text{Beta}(1, 10^4)$, $d_{\max} \sim \text{LogNormal}(\log(30), 1)$ and $g_{\text{age}} \sim \text{LogNormal}(\log(160), 1)$, with parameters assumed independent *a priori* [104].

We run SMC-ABC until around 100 000 model simulations have been generated for each dataset. We use the SBI package [105] to implement SNL with five rounds of 10 000 model

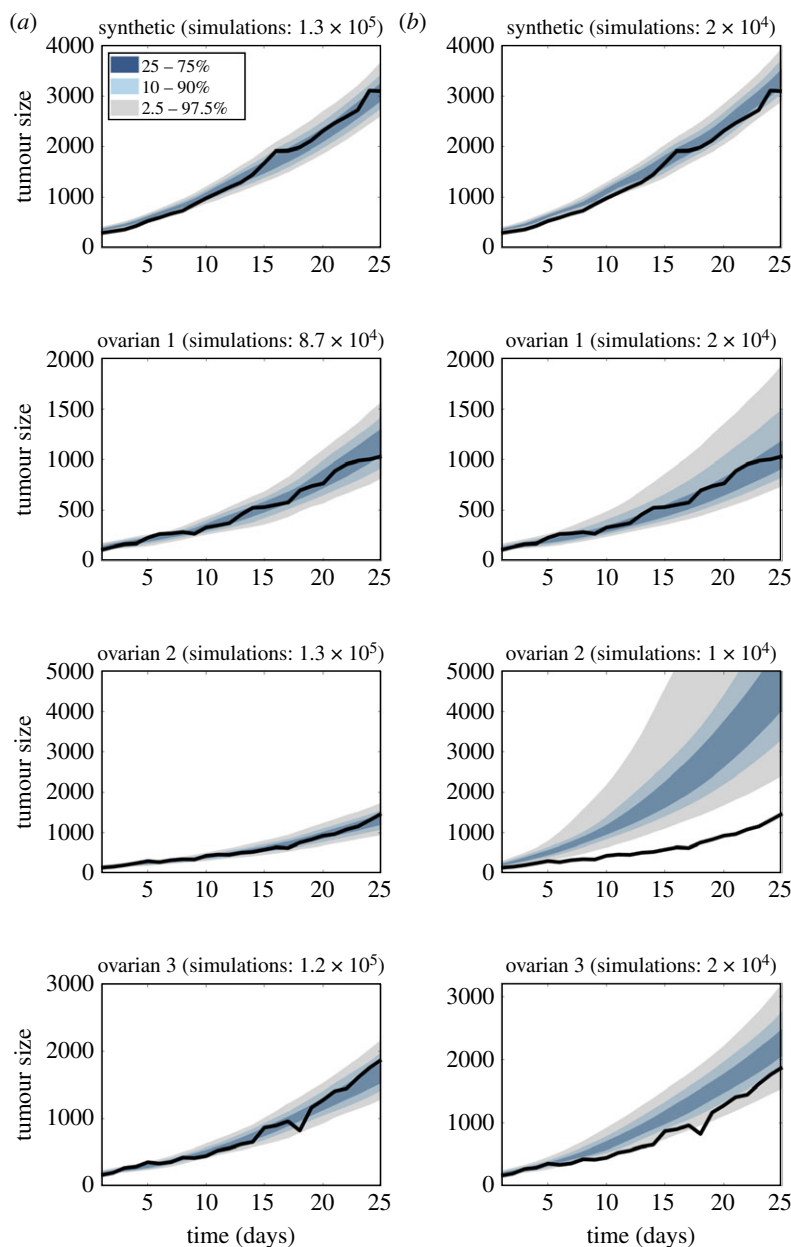


Figure 4. The posterior predictive distributions of (a) SMC-ABC and (b) SNL for the synthetic and ovarian cancer datasets. The black solid line is the tumour growth data. (Online version in colour.)

simulations for each dataset. To compare the performance of SMC-ABC and SNL, we compute the posterior predictive distribution for each dataset. For SNL, we choose the round that visually produces the most accurate posterior predictive distribution. We find that for SNL the performance can degrade with increasing rounds in three ovarian cancer datasets.

The results are shown in figure 4. It can be seen that SMC-ABC produces posterior predictive distributions that tightly enclose the time series of tumour volumes for three real-world ovarian cancer datasets. It is evident that SNL produces an accurate posterior predictive distribution for the synthetic dataset, with substantially fewer model simulations than that used for SMC-ABC.

This result is aligned with other synthetic examples in the literature (e.g. [106]). However, the SNL results for the real data are mixed, and for the three real datasets SMC-ABC produces more accurate posterior predictive distributions. Further, we do not necessarily see an improvement in SNL when increasing the number of rounds (i.e. number of model simulations). We suggest a reason for the potential poor performance of SNL is that the real data are more noisy than what the simulator is able to produce, and this may lead to a poorly estimated likelihood generated by SNL when evaluated at the observed data. By contrast, SMC-ABC produces results that are more robust to this misspecification, albeit at a higher computational cost in terms of model simulations. The potential poor performance of SNL under misspecification, and possible remedies to this problem, require further research (see [107] for the first approach to addressing this problem). In terms of computational cost, SMC-ABC takes approximately 3 h for each dataset. For SNL, it takes approximately 5 min to generate model simulations, approximately 20 min to train the conditional normalizing flow and approximately 3 h to generate approximate posterior samples (using the slice sampler as in [89]) for each round. The C++ and *Python* codes used in this study are available at [108].

6. Model transfer

Updating prior beliefs based on data is a core tenet of Bayesian inference. In the Bayesian context, model transfer extends Bayesian updating by incorporating information from a well-known source domain into a target domain. Consider the scenario where a target domain has insufficient data y_T to enable useful inference. Model transfer allows us to borrow information from a source domain with sufficient data y_S to improve inference. The transferability problem then is a question of when to transfer information, which information to transfer, and how to transfer this information. This problem appears across several domains, with some solutions exploiting the underlying properties of the source model, while others create informative priors with the source information. Below, we will discuss several different approaches to the model transfer problem. This broad topic is also known as transfer learning in the machine learning literature [109].

Naive updating, which uses all available source information, is a natural starting point to approach model transfer, though it can be detrimental. If the source and target distributions are dissimilar, negative transfer [110] may occur reducing the inference or predictive power from our posterior. Power priors [111] correct for the difference between source and target distributions by flattening the likelihood of the source distribution. This flattening is done by choosing a value $\phi \in [0, 1]$ and raising the source likelihood to the value of ϕ which gives

$$\pi(\theta|\phi, y_T, y_S) \propto f_T(y_T|\theta) f_S(y_S|\theta)^\phi \pi(\theta),$$

where $f_S(y_S|\theta)$ and $f_T(y_T|\theta)$ are the source and target likelihood functions, respectively. Naive updating would simply use the value $\phi = 1$. Finding an appropriate value for ϕ is challenging, intuitively we want to treat this as a latent variable and assign an appropriate prior. Unfortunately, even when both datasets are from the same distribution, the resulting posterior marginal of ϕ may exhibit only slightly less variance than the chosen prior. This phenomenon is analysed in [112] with illustrative examples. Other approaches attempt to determine an appropriate value of ϕ by optimization. Different information criteria, from the standard deviance information criterion to more complex penalized likelihood-type criterion, have been used [113] including the marginal likelihood [114] and the pseudo-marginal likelihood [115] which are evaluated using only the target data.

The transfer learning literature has a large number of methods for model transfer, evident by the recent review paper [109]. Many of these methods are specific to neural networks, but some can still be applied to broader classes of statistical models. An example of such a method is described in [116] which uses an ensemble of convolutional neural networks with a majority voting selection step that is easily generalized for use beyond neural networks. Another method, TrAdaBoost.R2 [117,118] adapts boosting [119] to the model transfer problem. This method

iteratively reweights each data point in the source and target domain to improve the predictive performance of the target model. There are also several methods specific to generalized linear models including; knockoff filters [120] to identify a subset of the source data to use, scaling the source likelihood function [121,122], and regularization [123,124] to adjust the weight of the source data. Finally, Transfer GPs [125–127] attempt to use information from the source kernel to improve model performance on the target domain. This is achieved by pooling the source and target datasets and producing a new joint kernel

$$\tilde{k}(x, x') = \begin{cases} \lambda k(x, x'), & \text{if } x \text{ and } x' \text{ are in different domains} \\ k(x, x'), & \text{otherwise.} \end{cases}$$

Above, $\lambda \in [0, 1]$, where $\lambda = 0$ indicates no information transfer and $\lambda = 1$ complete information transfer. For the interested reader, exemplar code is available via [128].

Current state-of-the-art Bayesian model transfer generalizes naive Bayesian updating but relies on fixed levels of transfer rather than incorporating uncertainty. It is still not clear how one should learn an optimal ϕ value in this paradigm but we expect future research will address this and use uncertainty more effectively. Moreover, given the interest in model-specific transfer learning, we believe that a Bayesian approach will be useful to develop general methods that are model agnostic.

7. Purposeful products

A key advantage of Bayesian methods is their ability to assist in decision making, and here three different case studies showcase innovative tools using Bayesian approaches.

(a) CoRiCAL: COVID-19 vaccine risk-benefit calculator

During the 1st year of the COVID-19 pandemic in 2020, border closures, lockdowns and other favourable conditions meant that Australia was spared from the high *per capita* case numbers and COVID-19-related deaths that were experienced in many other countries. When vaccines became available in February 2021 [129], the low number of COVID-19-related fatalities in Australia was coupled with uncertainty around highly publicized rare adverse-events for the vaccines: thrombosis and thrombocytopenia syndrome from AstraZeneca [130] and myocarditis from Pfizer [131]. This led to high levels of vaccination hesitancy in the general public [132]. Although emerging scientific evidence was increasingly available on the risks of the vaccines and their effectiveness against both becoming infected and becoming severely ill once infected [133,134], compiling and assessing this information from Scientific journals and Government reports is impossible for the majority of the population. Collating this evidence into an easily understood format that could be used by people to make an informed decision on COVID-19 vaccination in the Australian context became crucial.

Bayesian networks [135] are conditional probability models commonly represented as directed-acyclic graphs, with nodes and links representing variables of interest and the interactions between them. Conditional probabilities for the dependent child-nodes are stored in conditional probability tables (figure 5), which determine the probability of a node being in a given state for each possible combination of parent node states. Using Bayes theorem [135], the model calculates the probability of a given outcome for any defined scenario. Bayesian networks are widely used in a range of decision support settings including public health [136,137], environmental conservation [138] and natural resource management [139].

There are several characteristics that make Bayesian networks attractive for an evidence-based COVID-19 risk-benefit calculator. First, the conditional probability tables can be populated from different sources, such as data from government reports, results from scientific studies, or recommendations from experts and advisory committees. Second, the probabilistic output means

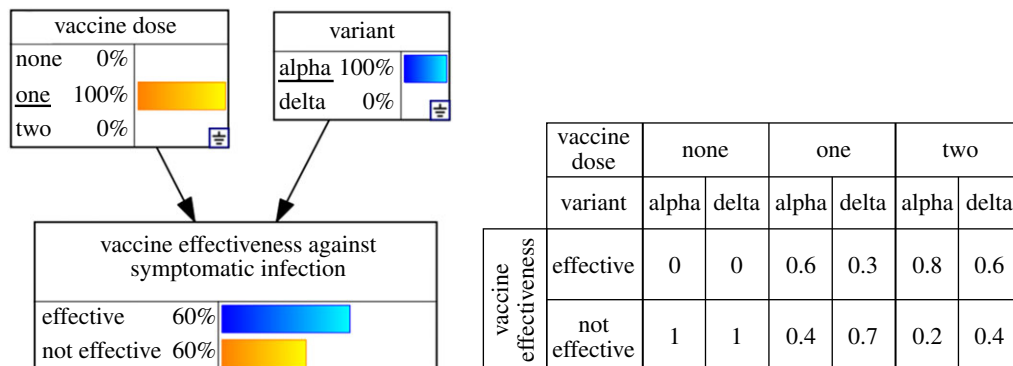


Figure 5. An example Bayesian network with a single, dependent child node (vaccine effectiveness) and two parent nodes (vaccine doses and variant). Conditional probability table for vaccine effectiveness is shown on the right. (Online version in colour.)

that the model can respond to user-defined scenarios such as ‘how likely is it that I will get sick’ rather than just ‘will I get sick’. Finally, Bayesian networks are highly interpretable models [140], as they allow exploration of the effect of different observed values (evidence) on the probability of certain outcomes.

The COVID-19 risk calculator (CoRiCAL—<https://corical.immunisationcoalition.org.au>) was developed to help the general public, as well as the doctors advising them, weigh-up the risks and benefits of receiving a COVID-19 vaccination. A Bayesian network model was constructed and parameterized based on the best available evidence from a range of sources that can be used to determine a person’s risk of developing symptomatic COVID-19, dying or other adverse effects from COVID-19, or suffering from adverse effects (including death) from the vaccine itself [141]. The model relied on Australian data to represent the context as accurately as possible, however in cases where local data was lacking, international data was used [142,143]. Full model information, along with model code is available via the link [144]. A web-based interface (figure 6) was developed to create a user-friendly tool that considers a person’s age and sex, the brand of the vaccine, how many vaccines they have had already, and the current levels of transmission within the community and displays their chances of an adverse event alongside common relatable risks. As the pandemic landscape changes, it remains crucial that the evidence for making informed choices on COVID-19 vaccination is made accessible. The model is updated in light of new variants, and as new vaccines become available and recommended (e.g. booster shots).

(b) ReefCloud: a tool to monitor coral reefs worldwide

Recent projections estimate that 99% of the world’s coral reefs will suffer from frequent marine heatwaves under 1.5°C of warming due to climate change [145]. Important ecological and socio-economic changes already occur in tropical oceans because of the decline of key corals that support thousands of species [146]. The latter impacts about one billion people whose income, food supply, coastal protection, and cultural practices depend on coral reef biodiversity. Robust estimation of changes in coral communities at large spatial scales is challenging because there are a lack of observations due to the remoteness of coral reefs and the absence of monitoring programs in sea countries. Also, the fine-scale variability of changes in coral cover result in disparate long-term coral trajectories at reef locations situated only few hundred metres apart [147]. For reef managers, these challenges (among others) contribute to slowdown the development of strategies that aim to reduce impacts of climate change on coral reefs.

A spatio-temporal Bayesian model is developed to estimate the coverage of total coral cover across spatial scales and predict coverage values at unsampled locations. The approach

if I get COVID-19, what are my chances of dying?

these results are for a 40–49-year-old female

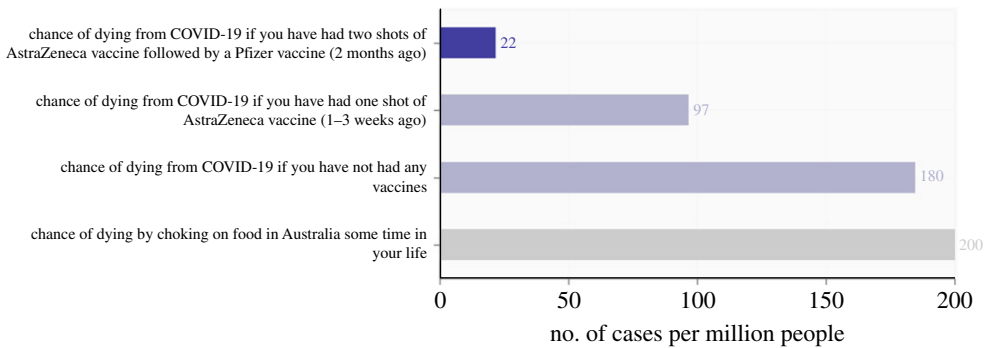


Figure 6. An example output from the CoRiCAL COVID-19 risk calculator tool. (Online version in colour.)

uses outputs from artificial intelligence algorithms trained to classify points on images [148]. For each Marine Ecoregion of the World (MEOW, [149]), a set of images $j = 1, 2, \dots, J$, each composed of $k = 1, 2, \dots, 50$ elicitation points is used across years of monitoring. Counts, y_{it} for observation i sampled at location s_i and time t , are modelled using a binomial distribution (with p the probability of positive and n_i the total number of positive cases) and controlled by additional components including the fixed effects of environmental disturbances (cyclones and mass coral bleaching events), sampling nested design (depth, transect, site, reef and monitoring program) modelled as independent and identically distributed Gaussian random effects, and spatio-temporal random effects.

The novelty in this model is the incorporation of a spatio-temporal random effects composed of a first-order autoregressive process in time and a Gaussian field that is approximated using a Gaussian Markov random field (GMRF), where the covariance is determined by a Matérn kernel. We employed the GMRF representation as a stochastic partial differential equation, using the method of [150], implemented in the R package *INLA* [151]. The spatial domain is based on the observed data locations and a buffer with adjacent MEOWs to allow information sharing between units. Spatial predictions are estimated at a grid level of 5×5 km resolution and posterior distributions used to reconstruct coral cover values at coarser spatial scales including MEOWs units and country level. Finally, estimations of coral cover are weighted by the proportion of coral reefs within a MEOW unit following the methodology developed as part of the global coral reef monitoring network [152]. We use the default *INLA* priors for different types of model parameters as discussed in [153]. The model is as follows:

$$y_{it} | \beta, \mathbf{Z}, V_i \sim \text{binomial} \left(n_i, \text{logit}^{-1} \left(\beta^\top x_i + r(s_i, t) + V_i \right) \right),$$

$$r(s_i, t) = \phi \cdot r(s_i, t - 1) + Z(s_i, t)$$

and
$$\mathbf{Z}(s, t) \stackrel{\text{ind}}{\sim} \mathcal{GP}(\mathbf{0}, \mathbf{K}), \quad t = 1, \dots, T.$$

The priors for the autoregressive parameter ϕ and independent Gaussian random effects V_i used are the *INLA* defaults. Research efforts focus on developing new technologies to assess the status of coral reefs in rapid and cost-effective ways through automatic image detection [148] and learn about impacts of multiple disturbances and management strategies [154,155]. ReefCloud is an open-access digital tool that support coral reef monitoring and decision-making by integration of data analyses and reporting (<https://reefcloud.ai/>). The online collection of worldwide data provides a unique opportunity to model these data together to (i) increase understanding on the impacts of environmental disturbances and (ii) reduce uncertainty when estimating coral

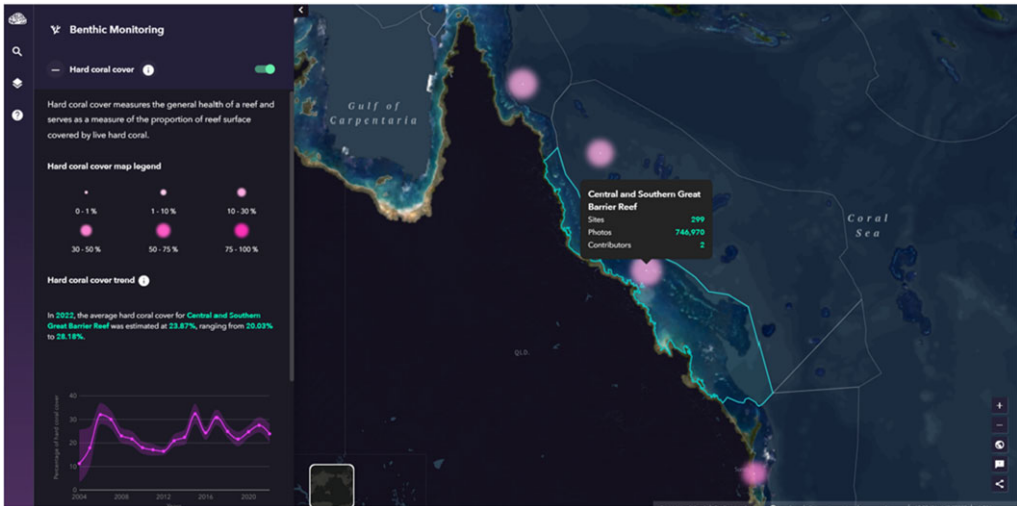


Figure 7. An example output from ReefCloud showing temporal trend in coral cover estimating from a Bayesian model for the central and southern parts of the Great Barrier Reef. (Online version in colour.)

trajectories at large spatial scales. The pilot product version is developed using the most extensive monitoring program in the world surveying the Great Barrier Reef, Australia. Machine learning outputs from one million of reef images are used to predict values in coral cover across 3000 coral reefs from 2004 onward. The ReefCloud online dashboard makes knowledge about reef changes accessible to everyone (figure 7). The project also educates the reef research community and managers on how Bayesian statistical modelling can help to increase our understanding of the impacts of climate change on coral reefs and supporting decision-making from local to global scales.

(c) Australian Cancer Atlas

Cancer is the leading cause of disease burden in Australia, which has comprehensive cancer incidence reporting for all cancers except common skin cancers [156]. Yet because Australia's population is heavily concentrated in specific coastal areas, cancer rates are commonly reported only for large regions. Difficulties when using sparse data for smaller areas include the reliability of estimates and the risk of identifying individuals. Yet, detailed small-area analyses have immense power to identify and understand inequities in cancer outcomes.

Using Bayesian hierarchical Poisson models incorporating Leroux priors [157] for spatial smoothing, robust and reliable cancer incidence and 5-year survival estimates were generated across Australian statistical areas level 2 (SA2; 2148 areas). These areas represent communities which interact together and while population sizes vary, the median is around 10 000 people [158]. Innovative visualizations helped rapidly identify areas which differed from the national average. Further details on the methods and visualizations are available in [159]. Example code for the Bayesian spatial models is available in §§7.3 and 9.8.2 of [160].

In September 2018, the Australian Cancer Atlas (atlas.cancer.org.au) was launched, providing the highest geographical resolution nationwide estimates available (figure 8). The website is designed to be interactive and engaging, featuring the ability to download all estimates, export pdfs of specific views, filter regions, rapidly compare different cancer types and rates for two areas, and more! There has been strong uptake and positive feedback, and in 2021 estimates were updated and cancer types expanded.

The Atlas has received prestigious spatial industry awards and is currently being replicated internationally. Australian Cancer Atlas 2.0 is underway, which will examine spatio-temporal

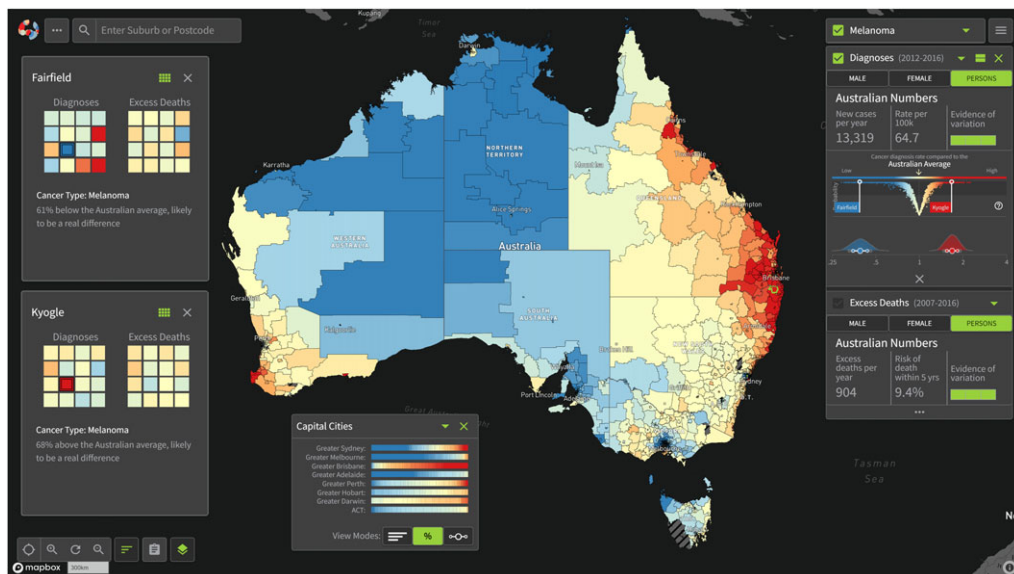


Figure 8. An example screenshot of the Australian Cancer Atlas showing melanoma incidence patterns with summary graphs. Red represents high incidence while blue is low in comparison to the national average (pale yellow). (Online version in colour.)

patterns, and further include cancer risk factors, some types of cancer treatment and selected cancer clinical/stage patterns. Underpinned by Bayesian methods, the Atlas will continue to provide the methods and visualizations necessary for accurate estimation, interpretation and making decisions.

8. Conclusion

This paper has focused on a small number of current opportunities and challenges in the application of the Bayesian paradigm. Of course, these are not the only issues, but collectively they point to the maturity of current Bayesian practice and the promise of a fully mature Bayesian future. As a final thought, we note that many advances in applied Bayesian statistics in recent years are deeply indebted to computational and methodological advances surrounding complex hierarchically structured models. Modern applied Bayesian statistics thus finds itself at the interface with not only its traditional neighbour mathematics, but also increasingly with the field of computer science. This partnership is one of considerable further promise in the years to come.

Data accessibility. This article has no additional data.

Authors' contributions. J.J.B.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; A.B.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; K.B.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; S.C.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; C.C.D.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; C.H.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; A.J.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; H.M.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; J.M.M.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; K.M.: conceptualization, formal analysis, investigation, methodology, project administration, resources, supervision, writing—original draft, writing—review and editing; A.P.: formal analysis, investigation, methodology, supervision, writing—original draft, writing—review and editing; R.S.: formal analysis,

investigation, methodology, software, writing—original draft, writing—review and editing; E.S.-F: formal analysis, investigation, methodology, writing—original draft, writing—review and editing; J.V.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing; X.W.: formal analysis, investigation, methodology, software, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. A.B. and J.J.B. acknowledge support from an Australian Research Council Discovery Project (DP200102101); K.B. was supported by a scholarship under the Australian Research Council Linkage Project (LP180101151), S.C. receives salary and research support from an NHMRC Investigator grant no. 2008313); C.D. acknowledges support from an Australian Research Council Future Fellowship (FT210100260); C.H. was supported by an Australian Research Council Linkage Project (LP200100468); J.M. was supported by an Australian Research Council Discovery (DP200101263) and Linkage Project (LP180101151).

Acknowledgements. We thank Dr Jasmine Lee for collecting and providing the 360° images in Antarctica.

References

1. Neyman J, Pearson ES. 1928 On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika* **20A**, 175–240.
2. Wald A. 1949 Statistical decision functions. *Ann. Math. Stat.* **20**, 165–205. (doi:10.1214/aoms/1177730030)
3. Lindley DV. 1972 *Bayesian statistics: a review*. Montpelier, VT: SIAM.
4. Müller P, Parmigiani G. 1995 Optimal design via curve fitting of Monte Carlo experiments. *J. Am. Stat. Assoc.* **90**, 1322–1330.
5. Müller P. 1999 Simulation-based optimal design. *Handb. Stat.* **6**, 459–474.
6. Müller P, Sansó B, De Iorio M. 2004 Optimal Bayesian design by inhomogeneous Markov chain simulation. *J. Am. Stat. Assoc.* **99**, 788–798.
7. Amzal B, Bois FY, Parent E, Robert CP. 2006 Bayesian-optimal design via interacting particle systems. *J. Am. Stat. Assoc.* **101**, 773–785. (doi:10.1198/016214505000001159)
8. Overstall AM, McGree JM, Drovandi CC. 2018 An approach for finding fully Bayesian optimal designs using normal-based approximations to loss functions. *Stat. Comput.* **28**, 343–358. (doi:10.1007/s11222-017-9734-x)
9. Foster A, Jankowiak M, Bingham E, Horsfall P, Teh YW, Rainforth T, Goodman N. 2019 Variational Bayesian optimal experimental design. Part of advances in neural information processing systems 32 (NeurIPS 2019).
10. Overstall AM, Woods DC. 2017 Bayesian design of experiments using approximate coordinate exchange. *Technometrics* **59**, 458–470. (doi:10.1080/00401706.2016.1251495)
11. Berry DA. 2006 Bayesian clinical trials. *Nat. Rev. Drug Discov.* **5**, 27–36. (doi:10.1038/nrd1927)
12. Connor JT, Elm JJ, Broglio KR, The ESETT and ADAPT-IT Investigators, 2013 Bayesian adaptive trials for comparative effectiveness research: an example in status epilepticus. *J. Clin. Epidemiol.* **66**, S130–S137. (doi:10.1016/j.jclinepi.2013.02.015)
13. Thorlund K, Haggstrom J, Park JJH, Mills EJ. 2018 Key design considerations for adaptive clinical trials: a primer for clinicians. *BMJ* **360**, k698. (doi:10.1136/bmj.k698)
14. Kang SY, McGree JM, Drovandi C, Mengersen K, Caley J. 2016 Bayesian adaptive design: improving the effectiveness of reef monitoring programs. *Ecol. Appl.* **26**, 2637–2648. (doi:10.1002/eap.1409)
15. Thilan P, Fisher R, Thompson H, Menendez P, Gilmour J, McGree JM. 2022 Adaptive monitoring of coral health at Scott Reef where data exhibit nonlinear and disturbed trends over time. *Ecol. Evol.* **12**, e9233.
16. Wagner D, Friedlander AM, Pyle RL, Brooks CM, Gjerde KM, Wilhelm TA. 2020 Coral reefs of the high seas: hidden biodiversity hotspots in need of protection. *Front. Mar. Sci.* **7**, 1–13.
17. AIMS, 2021 Annual summary report of coral reef condition 2020/21. See www.aims.gov.au/reef-monitoring/gbr-condition-summary-2020-2021.
18. Buchhorn K, Mengersen K, Santos-Fernandez E, Peterson EE, McGree JM. 2022 Bayesian design with sampling windows for complex spatial processes. Preprint (<https://arxiv.org/abs/2206.05369>).

19. Bassi A, Berkhof J, de Jong D, van de Ven PM. 2021 Bayesian adaptive decision-theoretic designs for multi-arm multi-stage clinical trials. *Stat. Methods Med. Res.* **30**, 717–730. (doi:10.1177/0962280220973697)
20. Giovagnoli A. 2021 The Bayesian design of adaptive clinical trials. *Int. J. Environ. Res. Public Health* **18**, 530. (doi:10.3390/ijerph18020530)
21. Kojima M. 2021 Early completion of phase I cancer clinical trials with Bayesian optimal interval design. *Stat. Med.* **40**, 3215–3226. (doi:10.1002/sim.8886)
22. McGree J *et al.* 2022 Controlled evaluation of angiotensin receptor blockers for COVID-19 respiratory disease (CLARITY): statistical analysis plan for a randomised controlled Bayesian adaptive sample size trial. *Trials* **23**, 1–18. (doi:10.1186/s13063-022-06167-2)
23. Leach CB, Williams PJ, Eisaguirre JM, Womble JN, Bower MR, Hooten MB. 2022 Recursive Bayesian computation facilitates adaptive optimal design in ecological studies. *Ecology* **103**, e03573. (doi:10.1002/ecy.3573)
24. Mazumdar S, Ceccaroni L, Piera J, Hölker F, Berre A, Arlinghaus R, Bowser A. 2018 Citizen science technologies and new opportunities for participation. In *Citizen science technologies and new opportunities for participation*. London, UK: UCL Press.
25. Queiroz ACM, Nascimento AM, Tori R, Silva Leme MID. 2019 Immersive virtual environments and learning assessments. In *Int. Conf. on Immersive Learning*, pp. 172–181. Berlin, Germany: Springer.
26. Fauville G, Queiroz ACM, Bailenson JN. 2020 Virtual reality as a promising tool to promote climate change awareness. *Technol. Health*, 91–108.
27. Vercelloni J *et al.* 2018 Using virtual reality to estimate aesthetic values of coral reefs. *R. Soc. Open Sci.* **5**, 172226. (doi:10.1098/rsos.172226)
28. Mengersen K *et al.* 2017 Modelling imperfect presence data obtained by citizen science. *Environmetrics* **28**, e2446. (doi:10.1002/env.2446)
29. Leigh C *et al.* 2019 Using virtual reality and thermal imagery to improve statistical modelling of vulnerable and protected species. *PLoS ONE* **14**, e0217809.
30. Lee JR, Raymond B, Bracegirdle TJ, Chadès I, Fuller RA, Shaw JD, Terauds A. 2017 Climate change drives expansion of Antarctic ice-free habitat. *Nature* **547**, 49–54. (doi:10.1038/nature22996)
31. Parties ATC. 1960 Protocol on environmental protection to the Antarctic treaty. Madrid, Spain.
32. Vercelloni J, Peppinck J, Santos-Fernandez E, McBain M, Heron G, Dodgen T, Peterson EE, Mengersen K. 2021 Connecting virtual reality and ecology: a new tool to run seamless immersive experiments in R. *PeerJ Comput. Sci.* **7**, e544. (doi:10.7717/peerj-cs.544)
33. Zooniverse, 2022 Zooniverse. See www.zooniverse.org (accessed 23 September 2022).
34. Prudic KL, McFarland KP, Oliver JC, Hutchinson RA, Long EC, Kerr JT, Larrivée M. 2017 eButterfly: leveraging massive online citizen science for butterfly conservation. *Insects* **8**, 53. (doi:10.3390/insects8020053)
35. Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Fink D, Kelling S. 2009 eBird: a citizen-based bird observation network in the biological sciences. *Biol. Conserv.* **142**, 2282–2292. (doi:10.1016/j.biocon.2009.05.006)
36. Nugent J. 2018 Inaturalist. *Sci. Scope* **41**, 12–13.
37. van Strien AJ *et al.* 2013 Occupancy modelling as a new approach to assess supranational trends using opportunistic data: a pilot study for the damselfly *Calopteryx splendens*. *Biodivers. Conserv.* **22**, 673–686. (doi:10.1007/s10531-013-0436-1)
38. Dwyer RG, Carpenter-Bundhoo L, Franklin CE, Campbell HA. 2016 Using citizen-collected wildlife sightings to predict traffic strike hot spots for threatened species: a case study on the southern cassowary. *J. Appl. Ecol.* **53**, 973–982. (doi:10.1111/1365-2664.12635)
39. Strebel N, Kéry M, Schaub M, Schmid H. 2014 Studying phenology by flexible modelling of seasonal detectability peaks. *Methods Ecol. Evol.* **5**, 483–490. (doi:10.1111/2041-210X.12175)
40. Santos-Fernández E, Mengersen K. 2021 Understanding the reliability of citizen science observational data using item response models. *Methods Ecol. Evol.* **12**, 1533–1548.
41. Freitag A, Meyer R, Whiteman L. 2016 Strategies employed by citizen science programs to increase the credibility of their data. *Citiz. Sci.: Theory Pract.* **1**, 2.

42. Santos-Fernandez E, Peterson EE, Vercelloni J, Rushworth E, Mengersen K. 2021 Correcting misclassification errors in crowdsourced ecological data: a Bayesian perspective. *J. R. Stat. Soc. C* **70**, 147–173.
43. Baker FB, Kim SH. 2004 *Item response theory: parameter estimation techniques*. New York, NY: CRC Press.
44. EdgarSantos-Fernandez, 2021 Hakuna. See <https://github.com/EdgarSantos-Fernandez/hakuna>.
45. Peterson EE *et al.* 2020 Monitoring through many eyes: integrating disparate datasets to improve monitoring of the great barrier reef. *Environ. Modell. Softw.* **124**, 104557. (doi:10.1016/j.envsoft.2019.104557)
46. McMahan HB, Moore E, Ramage D, Arcas BA. 2017 Communication-efficient learning of deep networks from decentralized data. In *Proc. of the 20th Int. Conf. on Artificial Intelligence and Statistics*, Ft. Lauderdale, FL. PMLR 54:1273–1282.
47. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. 2020 Federated optimization in heterogeneous networks. *Proc. Mach. Learn. Syst.* **2**, 429–450.
48. Wang J, Liu Q, Liang H, Joshi G, Poor HV. 2020 Tackling the objective inconsistency problem in heterogeneous federated optimization. *Adv. Neural Inf. Process. Syst.* **33**, 7611–7623.
49. Yurochkin M, Agarwal M, Ghosh S, Greenewald K, Hoang N, Khazaeni Y. 2019 Bayesian nonparametric federated learning of neural networks. In *Int. Conf. on Machine Learning*, pp. 7252–7261. Long Beach, CA: PMLR.
50. Thibaux R, Jordan MI. 2007 Hierarchical beta processes and the Indian buffet process. In *Conf. on Artificial Intelligence and Statistics*, pp. 564–571. San Juan, Puerto Rico: PMLR.
51. Deist TM *et al.* 2020 Distributed learning on 20000+ lung cancer patients—the personal health train. *Radiother. Oncol.* **144**, 189–200. (doi:10.1016/j.radonc.2019.11.019)
52. Geleijnse G, Chiang RCJ, Sieswerda M, Schuurman M, Lee K, van Soest J, Dekker A, Lee WC, Verbeek XA. 2020 Prognostic factors for survival in patients with oral cavity cancer: a comparison of the Netherlands and Taiwan using privacy-preserving federated analyses. *Sci. Rep.* **10**, 20526.
53. Cellamare M, van Gestel AJ, Alradhi H, Martin F, Moncada-Torres A. 2022 A federated generalized linear model for privacy-preserving analysis. *Algorithms* **15**, 243. (doi:10.3390/a15070243)
54. Fienberg SE, Fulp WJ, Slavkovic AB, Wrobel TA. 2006 ‘Secure’ log-linear and logistic regression analysis of distributed databases. In *Int. Conf. on Privacy in Statistical Databases*, pp. 277–290. Berlin, Germany: Springer.
55. Slavkovic AB, Nardi Y, Tibbits MM. 2007 ‘Secure’ logistic regression of horizontally and vertically partitioned distributed databases. In *7th IEEE Int. Conf. on Data Mining Workshops (ICDMW 2007)*, pp. 723–728. Omaha, NE: IEEE.
56. Shi H, Jiang C, Dai W, Jiang X, Tang Y, Ohno-Machado L, Wang S. 2016 Secure multi-party computation grid logistic regression (SMAC-GLORE). *BMC Med. Inform. Decis. Mak.* **16**, 175–187. (doi:10.1186/s12911-016-0316-1)
57. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. 2016 Vertical grid logistic regression (VERTIGO). *J. Am. Med. Inform. Assoc.* **23**, 570–579. (doi:10.1093/jamia/ocv146)
58. Kamphorst B, Rooijackers T, Veugen T, Cellamare M, Knoors D. 2022 Accurate training of the Cox proportional hazards model on vertically-partitioned data while preserving privacy. *BMC Med. Inform. Decis. Mak.* **22**, 1–18. (doi:10.1186/s12911-022-01771-3)
59. Cramer R, Damgård IB. 2015 *Secure multiparty computation*. Cambridge, UK: Cambridge University Press.
60. Minka TP. 2003 A comparison of numerical optimizers for logistic regression. *Unpublished Draft*, pp. 1–18.
61. Moncada-Torres A, Martin F, Sieswerda M, Van Soest J, Geleijnse G. 2020 VANTAGE6: an open source privacy preserving federated learning infrastructure for secure insight exchange. In *AMIA Annual Symp. Proc.*, vol. 2020, p. 870. Chicago, IL: American Medical Informatics Association.
62. Wang X, Dunson DB. 2014 Parallelizing MCMC via Weierstrass sampler. Preprint (<https://arxiv.org/abs/1312.4605>).

63. Neiswanger W, Wang C, Xing EP. 2014 Asymptotically exact, embarrassingly parallel MCMC. In *Proc. of the 13th Conf. on Uncertainty in Artificial Intelligence, UAI'14*, pp. 623–632. Arlington, Virginia, USA: AUAI Press.
64. Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI, McCulloch RE. 2016 Bayes and big data: the consensus Monte Carlo algorithm. *Int. J. Manage. Sci. Eng. Manage.* **11**, 78–88.
65. Jordan MI, Lee JD, Yang Y. 2019 Communication-efficient distributed statistical inference. *J. Am. Stat. Assoc.* **114**, 668–681. (doi:10.1080/01621459.2018.1429274)
66. Plassier V, Vono M, Durmus A, Moulines E. 2021 DG-LMC: a turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within Gibbs. In *Int. Conf. on Machine Learning*, pp. 8577–8587. Virtual: PMLR.
67. El Mekkaoui K, Mesquita D, Blomstedt P, Kaski S. 2021 Federated stochastic gradient Langevin dynamics. In *Uncertainty in Artificial Intelligence*, pp. 1703–1712. Virtual: PMLR.
68. De Cristofaro E. 2020 An overview of privacy in machine learning. Preprint (<https://arxiv.org/abs/2005.08679>).
69. Heikkilä M, Jälkö J, Dikmen O, Honkela A. 2019 Differentially private Markov chain Monte Carlo. *Adv. Neural Inf. Process. Syst.* **32**.
70. Bohensky MA, Jolley D, Sundararajan V, Evans S, Pilcher DV, Scott I, Brand CA. 2010 Data linkage: a powerful research tool with potential problems. *BMC Health Serv. Res.* **10**, 1–7. (doi:10.1186/1472-6963-10-346)
71. Harron K, Dibben C, Boyd J, Hjern A, Azimae M, Barreto ML, Goldstein H. 2017 Challenges in administrative data linkage for research. *Big Data Soc.* **4**, 1–12. (doi:10.1177/2053951717745678)
72. Gelman A, Hill J. 2006 *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.
73. Besag J. 1974 Spatial interaction and the statistical analysis of lattice systems. *J. R. Stat. Soc. B (Methodol.)* **36**, 192–225.
74. Besag J, York J, Mollié A. 1991 Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **43**, 1–20. (doi:10.1007/BF00116466)
75. Leroux BG, Lei X, Breslow N. 2000 Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pp. 179–191. Berlin: Springer.
76. Sisson SA, Fan Y, Beaumont M. 2018 *Handbook of approximate Bayesian computation*. London, UK: Chapman and Hall/CRC.
77. Beaumont MA, Zhang W, Balding DJ. 2002 Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035. (doi:10.1093/genetics/162.4.2025)
78. Beven K, Binley A. 1992 The future of distributed models: model calibration and uncertainty prediction. *Hydrol. Processes* **6**, 279–298. (doi:10.1002/hyp.3360060305)
79. Beven K, Binley A. 2014 Glue: 20 years on. *Hydrol. Processes* **28**, 5897–5918. (doi:10.1002/hyp.10082)
80. Nott DJ, Marshall L, Brown J. 2012 Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: what's the connection? *Water Resour. Res.* **48**. (doi:10.1029/2011WR011128)
81. Prangle D. 2018 Summary statistics. In *Handbook of approximate Bayesian computation*, pp. 125–152. London, UK: Chapman and Hall/CRC.
82. Drovandi C, Frazier DT. 2022 A comparison of likelihood-free methods with and without summary statistics. *Stat. Comput.* **32**, 1–23. (doi:10.1007/s11222-022-10092-4)
83. Sisson S, Fan Y 2018 *Handbook of approximate Bayesian computation*. ABC Samplers, 87–123. London, UK: Chapman and Hall/CRC (Chapter).
84. Frazier DT, Martin GM, Robert CP, Rousseau J. 2018 Asymptotic properties of approximate Bayesian computation. *Biometrika* **105**, 593–607. (doi:10.1093/biomet/asy027)
85. Price LF, Drovandi CC, Lee A, Nott DJ. 2018 Bayesian synthetic likelihood. *J. Comput. Graph. Stat.* **27**, 1–11. (doi:10.1080/10618600.2017.1302882)
86. Frazier D, Nott DJ, Drovandi C, Kohn R. 2022 Bayesian inference using synthetic likelihood: asymptotics and adjustments. *J. Am. Stat. Assoc.* 1–12.
87. Carr MJ, Simpson MJ, Drovandi C. 2021 Estimating parameters of a stochastic cell invasion model with fluorescent cell cycle labelling using approximate Bayesian computation. *J. R. Soc. Interface* **18**, 20210362. (doi:10.1098/rsif.2021.0362)

88. Drovandi CC, Pettitt AN. 2011 Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics* **67**, 225–233. (doi:10.1111/j.1541-0420.2010.01410.x)
89. Papamakarios G, Sterratt D, Murray I. 2019 Sequential neural likelihood: fast likelihood-free inference with autoregressive flows. In *The 22nd Int. Conf. on Artificial Intelligence and Statistics*, pp. 837–848. PMLR.
90. Thomas O, Dutta R, Corander J, Kaski S, Gutmann MU. 2022 Likelihood-free inference by ratio estimation. *Bayesian Anal.* **17**, 1–31. (doi:10.1214/20-BA1238)
91. Lueckmann JM, Goncalves PJ, Bassetto G, Öcal K, Nonnenmacher M, Macke JH. 2017 Flexible statistical inference for mechanistic models of neural dynamics. *Adv. Neural Inf. Process. Syst.* **30**, 1289–1299.
92. Wang Z, Butner JD, Kerketta R, Cristini V, Deisboeck TS. 2015 Simulating cancer growth with multiscale agent-based modeling. In *Seminars in cancer biology*, vol. 30, pp. 70–78. Elsevier.
93. Metzcar J, Wang Y, Heiland R, Macklin P. 2019 A review of cell-based computational modeling in cancer biology. *JCO Clin. Cancer Inform.* **2**, 1–13. (doi:10.1200/CCI.18.00069)
94. Macnamara CK. 2021 Biomechanical modelling of cancer: agent-based force-based models of solid tumours within the context of the tumour microenvironment. *Comput. Syst. Oncol.* **1**, e1018.
95. Cess CG, Finley SD. 2022 Multiscale modeling of tumor adaption and invasion following anti-angiogenic therapy. *Comput. Syst. Oncol.* **2**, e1032.
96. Norton KA, Gong C, Jamalian S, Popel AS. 2019 Multiscale agent-based and hybrid modeling of the tumor immune microenvironment. *Processes* **7**, 37. (doi:10.3390/pr7010037)
97. Cess CG, Finley SD. 2020 Multi-scale modeling of macrophage—T cell interactions within the tumor microenvironment. *PLoS Comput. Biol.* **16**, e1008519. (doi:10.1371/journal.pcbi.1008519)
98. Jenner AL *et al.* 2022 Agent-based computational modeling of glioblastoma predicts that stromal density is central to oncolytic virus efficacy. *iScience* **25**.104395 (doi:10.1016/j.isci.2022.104395)
99. Klowss JJ, Browning AP, Murphy RJ, Carr EJ, Plank MJ, Gunasingh G, Haass NK, Simpson MJ. 2022 A stochastic mathematical model of 4D tumour spheroids with real-time fluorescent cell cycle labelling. *J. R. Soc. Interface* **19**, 20210903. (doi:10.1098/rsif.2021.0903)
100. Gallaher JA *et al.* 2020 From cells to tissue: how cell scale heterogeneity impacts glioblastoma growth and treatment response. *PLoS Comput. Biol.* **16**, e1007672. (doi:10.1371/journal.pcbi.1007672)
101. Jenner A *et al.* 2022 Examining the efficacy of localised gemcitabine therapy for the treatment of pancreatic cancer using a hybrid agent-based model. *bioRxiv*.
102. Jenner AL, Frascoli F, Coster AC, Kim PS. 2020 Enhancing oncolytic virotherapy: observations from a Voronoi cell-based model. *J. Theor. Biol.* **485**, 110052. (doi:10.1016/j.jtbi.2019.110052)
103. Kim PH, Sohn JH, Choi JW, Jung Y, Kim SW, Haam S, Yun CO. 2011 Active targeting and safety profile of peg-modified adenovirus conjugated with herceptin. *Biomaterials* **32**, 2314–2326. (doi:10.1016/j.biomaterials.2010.10.031)
104. Wang X, Jenner AL, Salomone R, Drovandi C. 2022 Calibration of a Voronoi cell-based model for tumour growth using approximate Bayesian computation. *bioRxiv*.
105. Tejero-Cantero A, Boelts J, Deistler M, Lueckmann JM, Durkan C, Gonçalves PJ, Greenberg DS, Macke JH. 2020 SBI—a toolkit for simulation-based inference. Preprint (<https://arxiv.org/abs/2007.09114>).
106. Lueckmann JM, Boelts J, Greenberg D, Goncalves P, Macke J. 2021 Benchmarking simulation-based inference. In *Int. Conf. on Artificial Intelligence and Statistics*, pp. 343–351. PMLR.
107. Kelly RP, Nott DJ, Frazier DT, Warne DJ, Drovandi C. 2023 Misspecification-robust sequential neural likelihood. Preprint. (<https://arxiv.org/abs/2301.13368>)
108. Wang J. 2022 ABC and SNL. See <https://github.com/john-wang1015/ABCandSNL>.
109. Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, Xiong H, He Q. 2020 A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76. (doi:10.1109/JPROC.2020.3004555)
110. Agarwal N, Sondhi A, Chopra K, Singh G. 2021 Transfer learning: survey and classification. In *Smart Innovations In Communication and Computational Sciences*, pp. 145–155. Singapore: Springer.

111. Ye K, Han Z, Duan Y, Bai T. 2022 Normalized power prior Bayesian analysis. *J. Stat. Plann. Inference* **216**, 29–50. (doi:10.1016/j.jspi.2021.05.005)
112. Pawel S, Aust F, Held L, Wagenmakers EJ. 2022 Normalized power priors always discount historical data. Preprint (<https://arxiv.org/abs/2206.04379>).
113. Ibrahim JG, Chen MH, Gwon Y, Chen F. 2015 The power prior: theory and applications. *Stat. Med.* **34**, 3724–3749. (doi:10.1002/sim.6728)
114. Han Z, Ye K, Wang M. 2022 A study on the power parameter in power prior Bayesian analysis. *Am. Stat.* **77**, 1–8.
115. Bennett M, White S, Best N, Mander A. 2021 A novel equivalence probability weighted power prior for using historical control data in an adaptive clinical trial design: a comparison to standard methods. *Pharm. Stat.* **20**, 462–484. (doi:10.1002/pst.2088)
116. Chouhan V, Singh SK, Khamparia A, Gupta D, Tiwari P, Moreira C, Damasevicius R, de Albuquerque VHC. 2020 A novel transfer learning based approach for pneumonia detection in chest X-ray images. *Appl. Sci.* **10**, 559. (doi:10.3390/app10020559)
117. Gupta S, Bi J, Liu Y, Wildani A. 2022 Boosting for regression transfer via importance sampling.
118. Tang D, Yang X, Wang X. 2020 Improving the transferability of the crash prediction model using the TrAdaBoost.R2 algorithm. *Accid. Anal. Prev.* **141**, 105551. (doi:10.1016/j.aap.2020.105551)
119. Solomatine DP, Shrestha DL. 2004 AdaBoost. RT: a boosting algorithm for regression problems. In *2004 IEEE Int. Joint Conf. on Neural Networks (IEEE Cat. No. 04CH37541)*, vol. 2, pp. 1163–1168. IEEE.
120. Li S, Ren Z, Sabatti C, Sesia M. 2021 Transfer learning in genome-wide association studies with knockoffs. Preprint (<https://arxiv.org/abs/2108.08813>).
121. Maity S, Dutta D, Terhorst J, Sun Y, Banerjee M. 2021 A linear adjustment based approach to posterior drift in transfer learning. Preprint (<https://arxiv.org/abs/2111.10841>).
122. Reeve HW, Cannings TI, Samworth RJ. 2021 Adaptive transfer learning. *Ann. Stat.* **49**, 3618–3649. (doi:10.1214/21-AOS2102)
123. Guo S, Heinke R, Stöckel S, Rösch P, Bocklitz T, Popp J. 2017 Towards an improvement of model transferability for Raman spectroscopy in biological applications. *Vib. Spectrosc.* **91**, 111–118. (doi:10.1016/j.vibspec.2016.06.010)
124. Hector EC, Martin R. 2022 Turning the information-sharing dial: efficient inference from different data sources. Preprint (<https://arxiv.org/abs/2207.08886>).
125. Da B, Ong YS, Gupta A, Feng L, Liu H. 2019 Fast transfer Gaussian process regression with large-scale sources. *Knowl. Based Syst.* **165**, 208–218. (doi:10.1016/j.knsys.2018.11.029)
126. Wei P, Vo TV, Qu X, Ong YS, Ma Z. 2022 Transfer kernel learning for multi-source transfer Gaussian process regression. *IEEE Trans. Pattern Anal. Mach. Intell.*
127. Cao B, Pan SJ, Zhang Y, Yeung DY, Yang Q. 2010 Adaptive transfer learning. In *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 24, pp. 407–412.
128. Xiao-dong-Wang, 2021 Transfer-GP. See <https://github.com/Xiao-dong-Wang/Transfer-GP>.
129. Australian Government Department of Health and Aged Care, 2022 First COVID-19 vaccinations in Australia 2021. See www.health.gov.au/news/first-covid-19-vaccinations-in-australia.
130. Greinacher A, Thiele T, Warkentin TE, Weisser K, Kyrle PA, Eichinger S. 2021 Thrombotic thrombocytopenia after ChAdOx1 nCov-19 vaccination. *N. Engl. J. Med.* **384**, 2092–2101. (doi:10.1056/NEJMoa2104840)
131. Marshall M *et al.* 2021 Symptomatic acute myocarditis in 7 adolescents after Pfizer-BioNTech COVID-19 vaccination. *Pediatrics* **148**. (doi:10.1542/peds.2021-052478)
132. Leask J *et al.* 2021 Communicating with patients and the public about COVID-19 vaccine safety: recommendations from the collaboration on social science and immunisation. *Med. J. Aust.* **215**, 9–12. (doi:10.5694/mja2.51136)
133. Sheikh A, McMenamin J, Taylor B, Robertson C. 2021 SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness. *Lancet* **397**, 2461–2462. (doi:10.1016/S0140-6736(21)01358-1)

134. Zheng C, Shao W, Chen X, Zhang B, Wang G, Zhang W. 2022 Real-world effectiveness of COVID-19 vaccines: a literature review and meta-analysis. *Int. J. Infect. Dis.* **114**, 252–260. (doi:10.1016/j.ijid.2021.11.009)
135. Pearl J. 1988 *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Burlington, MA: Morgan Kaufmann.
136. Dickson BF *et al.* 2022 Bayesian network analysis of lymphatic filariasis serology from myanmar shows benefit of adding antibody testing to post-MDA surveillance. *Trop. Med. Infect. Dis.* **7**, 113. (doi:10.3390/tropicalmed7070113)
137. Wu Y, Foley D, Ramsay J, Woodberry O, Mascaro S, Nicholson AE, Snelling T. 2021 Bridging the gaps in test interpretation of SARS-CoV-2 through Bayesian network modelling. *Epidemiol. Infect.* **149**, 1–13. (doi:10.1017/S0950268821001357)
138. Camus EB *et al.* 2022 Using expert elicitation to identify effective combinations of management actions for koala conservation in different regional landscapes. *Wildl. Res.*
139. Xue J, Gui D, Lei J, Zeng F, Mao D, Zhang Z. 2017 Model development of a participatory Bayesian network for coupling ecosystem services into integrated water resources management. *J. Hydrol.* **554**, 50–65. (doi:10.1016/j.jhydrol.2017.08.045)
140. Uusitalo L. 2007 Advantages and challenges of Bayesian networks in environmental modelling. *Ecol. Modell.* **203**, 312–318. (doi:10.1016/j.ecolmodel.2006.11.033)
141. Mayfield HJ *et al.* 2022 Designing an evidence-based Bayesian network for estimating the risk versus benefits of AstraZeneca COVID-19 vaccine. *Vaccine* **40**, 3072–3084. (doi:10.1016/j.vaccine.2022.04.004)
142. Lau CL *et al.* 2021 Risk-benefit analysis of the AstraZeneca COVID-19 vaccine in Australia using a Bayesian network modelling framework. *Vaccine* **39**, 7429–7440. (doi:10.1016/j.vaccine.2021.10.079)
143. Sinclair JE, Mayfield HJ, Short KR, Brown SJ, Puranik R, Mengersen K, Litt JC, Lau CL. 2022 A Bayesian network analysis quantifying risks versus benefits of the Pfizer COVID-19 vaccine in Australia. *npj Vaccines* **7**, 1–11. (doi:10.1038/s41541-022-00517-6)
144. BayesFusion interactive model repository: CoRiCal AstraZeneca model. See <https://repo.bayesfusion.com/network/permalink?net=Small+BNs%2FCoRiCalAZ.xdsl>.
145. Dixon AM, Forster PM, Heron SF, Stoner AM, Beger M. 2022 Future loss of local-scale thermal refugia in coral reef ecosystems. *PLoS Clim.* **1**, e0000004. (doi:10.1371/journal.pclm.0000004)
146. Hughes TP *et al.* 2018 Spatial and temporal patterns of mass bleaching of corals in the anthropocene. *Science* **359**, 80–83. (doi:10.1126/science.aan8048)
147. Vercelloni J, Mengersen K, Ruggeri F, Caley MJ. 2017 Improved coral population estimation reveals trends at multiple scales on Australia's Great Barrier Reef. *Ecosystems* **20**, 1337–1350. (doi:10.1007/s10021-017-0115-2)
148. Gonzalez-Rivero M *et al.* 2020 Monitoring of coral reefs using artificial intelligence: a feasible and cost-effective approach. *Remote Sens.* **12**, 489.
149. Spalding MD *et al.* 2007 Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. *BioScience* **57**, 573–583. (doi:10.1641/B570707)
150. Lindgren F, Rue H. 2015 Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* **63**, 1–25. (doi:10.18637/jss.v063.i19)
151. Rue H, Martino S, Chopin N. 2009 Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B* **71**, 319–392. (doi:10.1111/j.1467-9868.2008.00700.x)
152. Souter D, Planes S, Wicquart J, Logan M, Obura D, Staub F, 2020 Status of coral reefs of the world: 2020. In *Global Coral Reef Monitoring Network; Int. Coral Reef Initiative*. Townsville, Australia: Australian Institute of Marine Science.
153. Moraga P. 2019 *Geospatial health data: modeling and visualization with R-INLA and Shiny*. New York, NY: CRC Press.
154. Vercelloni J, Liquet B, Kennedy EV, González-Rivero M, Caley MJ, Peterson EE, Puotinen M, Hoegh-Guldberg O, Mengersen K. 2020 Forecasting intensifying disturbance effects on coral reefs. *Glob. Change Biol.* **26**, 2785–2797. (doi:10.1111/gcb.15059)
155. Kennedy EV *et al.* 2020 Coral reef community changes in Karimunjawa National Park, Indonesia: assessing the efficacy of management in the face of local and global stressors. *J. Mar. Sci. Eng.* **8**, 760. (doi:10.3390/jmse8100760)

156. Australian Institute of Health Welfare. 2021 Cancer in Australia 2021. Report, AIHW.
157. Leroux BG, Lei X, Breslow N. 2000 *Estimation of disease rates in small areas: a new mixed model for spatial dependence*, pp. 135–178. New York, NY: Springer.
158. Australian Bureau of Statistics. 2011 Australian Statistical Geography Standard (ASGS): volume 1—main structure and greater capital city statistical areas, July 2011. Report, ABS.
159. Duncan EW, Cramb SM, Aitken JF, Mengersen KL, Baade PD. 2019 Development of the Australian Cancer Atlas: spatial modelling, visualisation, and reporting of estimates. *Int. J. Health Geogr.* **18**, 1–12. (doi:10.1186/s12942-019-0185-9)
160. Duncan EW, Cramb SM, Baade P, Mengersen KL, Saunders T, Aitken JF. 2020 Developing a Cancer Atlas using Bayesian methods: a practical guide for application and interpretation. Cancer Council Queensland and Queensland University of Technology. See <https://atlas.cancer.org.au/developing-a-cancer-atlas/>.