*algorithms*

MDPI

*Article*

# Active Data Selection and Information Seeking

**Thomas Parr [1,2,*], Karl Friston [2,3,4] and Peter Zeidman [3]**

[1] Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford OX3 9DU, UK
[2] Stanhope AI, London EC2M 1NH, UK; k.friston@ucl.ac.uk
[3] Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3AR, UK; peter.zeidman@ucl.ac.uk
[4] VERSES AI Research Lab, Los Angeles, CA 90016, USA
* Correspondence: thomas.parr@ndcn.ox.ac.uk

**Abstract:** Bayesian inference typically focuses upon two issues. The first is estimating the parameters of some model from data, and the second is quantifying the evidence for alternative hypotheses—formulated as alternative models. This paper focuses upon a third issue. Our interest is in the selection of data—either through sampling subsets of data from a large dataset or through optimising experimental design—based upon the models we have of how those data are generated. Optimising data-selection ensures we can achieve good inference with fewer data, saving on computational and experimental costs. This paper aims to unpack the principles of active sampling of data by drawing from neurobiological research on animal exploration and from the theory of optimal experimental design. We offer an overview of the salient points from these fields and illustrate their application in simple toy examples, ranging from function approximation with basis sets to inference about processes that evolve over time. Finally, we consider how this approach to data selection could be applied to the design of (Bayes-adaptive) clinical trials.

**Keywords:** experimental design; active sampling; information gain; Bayesian inference



**Citation:** Parr, T.; Friston, K.; Zeidman, P. Active Data Selection and Information Seeking. *Algorithms* **2024**, *17*, 118. https://doi.org/10.3390/a17030118

Academic Editors: Kevin B Korb, Steven Mascaro and Erik P. Nyberg

Received: 15 February 2024
Revised: 8 March 2024
Accepted: 10 March 2024
Published: 12 March 2024



## 1. Introduction

When we look at the world around us, we are implicitly engaging in a form of active data sampling (also known as active sensing or active inference [1–5]). Despite the abundance of visual data available to us at any one moment, our visual systems are adapted to select (or foveate) only a small portion of these data. The advantage of this is that our brains can scale the processes that underwrite perceptual inference to large sensory datasets. All that is needed is an efficient means of sequentially selecting those data to best optimise our inferences about their causes [6,7].

This paper suggests that the same process of active sampling can be (and often is) used in situations in which there is a cost associated with collecting new data or analysing a very large dataset. Figure 1 illustrates the basic idea, akin to perception–action cycles in biology [8], of alternating between drawing inferences from the data we have available and selecting new data based on these inferences. This cycle is implicitly part of the scientific process, in which our pre-existing understanding is used to motivate experiments whose results update our understanding to motivate future experiments, and so on.

Within the schematic shown in Figure 1, we highlight the key quantities in this cyclical process. They include the causes of data (which may include model parameters or hidden states) ($\theta$) and the data ($y$). Bayesian inference involves taking our prior beliefs about causes $p(\theta)$, combining these with the likelihood ($p(y \mid \theta)$) of observed data, and arriving at a posterior belief ($p(\theta \mid y)$) as to the causes given the data. The priors and likelihoods together comprise what is known as a generative model, which can be represented in a number of ways, including as the Bayesian network shown in the centre of the graphic. Here, an arrow from one variable to another indicates a conditional probability distribution of the latter given the former.

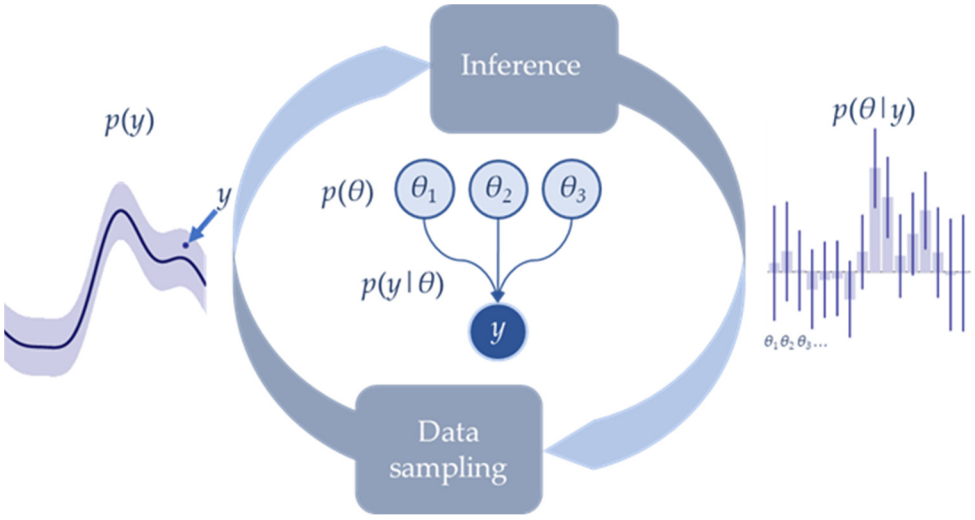


**Figure 1.** The active sampling cycle. This graphic illustrates the basic idea behind this paper. Analogous with action–perception cycles of the sort found in biological systems, it shows reciprocal interactions between the process of sampling data (in biology, through acting upon the world to solicit new sensations) and drawing inferences about these data (akin to perceptual inference).

Careful data selection is especially important when we consider the problems associated with very large datasets of the sort that are now ubiquitous in machine learning and artificial intelligence settings. Computations with such datasets can be costly in terms of the hardware and computing power required, and their energy consumption raises important sustainability questions [9–12]. Taking inspiration from the approach evinced by natural selection—that of sequentially selecting small amounts of sensory data—may help to alleviate the costs associated with big-data analysis.

To optimise data selection, we first need to identify an appropriate optimality criterion. In what follows, we base this upon the idea of expected information gain—a measure used in optimisation of experimental design [13,14], feature selection [15], and accounts of biological information-seeking and curiosity-driven behaviour [16,17]. The idea behind this measure is that we form beliefs about hidden states or parameters in our model of how data are generated. Expected information gain is the degree to which we anticipate changing our beliefs under a given experimental design or data-sampling strategy. The implication is that optimisation of beliefs—and of data selection—work in tandem, as in Figure 1, and that both depend upon our model of how data are generated.

In what follows, we consider the form this model, and therefore data-selection, might take in different settings, starting with abstract function-approximation and progressing to a more realistic example based upon the idea of an adaptive clinical trial. Before we move to these examples, we unpack the basic theory behind active data selection. We outline the basic principles behind Bayesian inference, the role of generative models, and the formulation of expected information gain. Our overall aim is to provide an intuitive overview of the principles that underwrite active data selection, and to illustrate this with some simple examples.

## 2. Bayesian Inference, Generative Models, and Expected Information Gain

Bayesian inference is the process of inverting a model of how data ($y$) are generated to obtain two things [18]. The first is the probability of observed data under our model—sometimes referred to as marginal likelihood. This is a ubiquitous measure of the evidence our data affords the hypothesis expressed in our model. Second is the probability of the random variables ($\theta$) in the model given our data—known as a posterior probability. These two things can be obtained by specifying the a priori plausible distributions of model variables and the likelihood of generating patterns of data given the values those variables might take. These prior and likelihood beliefs form our generative model. The

relationship between the generative model and the associated marginal likelihood and posterior probability is given by Bayes' theorem, which can be expressed as:

$$\begin{aligned} \underbrace{p(y|\theta)p(\theta)}_{\substack{\text{Generative}\\ \text{Model}}} &= \underbrace{p(\theta|y)}_{\text{Posterior}} p(y) \\ \underbrace{p(y)}_{\substack{\text{Marginal}\\ \text{Likelihood}}} &= \mathbb{E}_{p(\theta)}[p(y|\theta)] \end{aligned} \tag{1}$$

Equation (1) shows the generative model, its inversion, and the relationship between a model and its marginal likelihood. The second line shows that the marginal likelihood is obtained directly from the prior and likelihood beliefs simply by taking the expectation (or average) of the latter, assuming our $\theta$ variables are distributed according to the prior (also known as marginalisation). This leaves only one unknown in the first line—the posterior—which can then be obtained from rearrangement of the other three terms. In practice, $\theta$ may include many different parameters or hidden states, making exact computation impractical [19]. While we will touch upon the topic of approximate inference briefly in one of the examples, this is of limited importance for the primary message of this paper. However, the question of how we deal with complex models with multiple parameters is important to address.

Bayesian networks, and related graphical notations, offer a visualisation of a generative model. They tell us which variables depend upon which other variables. This is useful in that it suggests efficient message passing schemes that may be used to perform an inference (see [20,21] for overviews). Such representations have broad applicability, ranging from clinical reasoning [22] to environmental conservation [23]. For our purposes, message passing in graphical models is useful in that it helps us to find the quantities required for computing expected information gain.

The information we gain on making an observation can be formalised as the change in our beliefs as a consequence of that observation. A change in probabilities of this sort is quantified using a KL-Divergence (also known as relative entropy)—the average difference between the log probabilities before and after that observation (i.e., between the prior and posterior of Equation (1)). The value this KL-Divergence takes, averaged over the data we would predict using our marginal likelihood, is our expected information gain [14]:

$$\begin{aligned} I[\pi] &= \mathbb{E}_{p(y|\pi)}\underbrace{[D_{KL}[p(\theta|y,\pi)||p(\theta|\pi)]]}_{\text{Information gain}} \\ &= \underbrace{D_{KL}[p(y,\theta|\pi)||p(\theta|\pi)p(y|\pi)]}_{\text{Mutual information}} \\ &= \underbrace{\mathbb{E}_{p(y,\theta|\pi)}[\ln p(y|\theta,\pi)]}_{\text{(-) Conditional entropy}} - \underbrace{\mathbb{E}_{p(y,\theta|\pi)}[\ln p(y|\pi)]}_{\text{(-) Predictive entropy}} \\ &= \mathbb{E}_{p(\theta|\pi)}\Big[\mathbb{E}_{p(y|\theta,\pi)}[\ln p(y|\theta,\pi)]\Big] - \mathbb{E}_{\mathbb{E}_{p(\theta|\pi)}[p(y|\theta,\pi)]}\Big[\ln \mathbb{E}_{p(\theta|\pi)}[p(y|\theta,\pi)]\Big] \end{aligned} \tag{2}$$

In Equation (2), we have conditioned our model upon the variable $\pi$, which represents a choice we can make in selecting our data. The first line specifies expected information gain directly as the expectation of an information gain. The second line expresses information gain in terms of mutual information that quantifies the degree of conditional dependence between parameters and data. If the parameters were independent of data, the joint distribution would be equal to the product of their marginals, rendering the divergence zero. The third line provides a further interpretation in terms of the difference between two entropies. Essentially, the greater the predictive uncertainty the greater the potential

information we can gain, but only if the conditional entropy is sufficiently small that there is a precise mapping from parameters to the data they cause.

The final line of Equation (2) re-expresses the penultimate line by explicitly separating the joint distributions into their constituent singleton (i.e., marginal) and conditional distributions. This gives a sense of a recursive structure, in which the key constituents of information gain are themselves built up of expectations of expectations (of expectations as the model becomes more complex). This provides us with a method for quantifying the information gain about parameters in graphical models. Such models are typically expressed using a series of nodes (or edges, in factor graphs) representing variables linked together by the factors of their associated probability distributions. Typically, causative variables are placed towards the top of the graph and measurable data at the bottom. Pairs of nodes are linked if there is a conditional dependence between them. In other words, two nodes are depicted as linked if there is a (conditional probability) factor of the model in which the variable represented by one of the two nodes is in the conditioning set (the 'parent') and the variable represented by the other node is the support (the 'child') of that factor.

Using the notation *pa*( ) to represent parents of (i.e., the things something depends upon), and *ch*( ) to represent children of (i.e., the things that depend upon it), we can express the expected information gain in Equation (2)—about a parameter $\theta_i$—in terms of messages ($\mu$) that are passed along the edges of the factor graph:

$$\begin{aligned} I[\pi]^i &= \Lambda_i\left[\mu_i^{pa(i)}\Lambda_y\left[\mu_i^{ch(i)}\ln\mu_i^{ch(i)}\right]\right] - \Lambda_y\left[\mu_y^{pa(y)}\ln\mu_y^{pa(y)}\right] \\ \mu_i^{pa(i)} &= p(\theta_i\Big|\pi) = \Lambda_{pa(i)}\left[\mu_{par(i)}^{pa(pa(i))}p(\theta_i\Big|par(\theta_i),\pi)\right] \\ \mu_i^{ch(i)} &= p(y\Big|\theta_i,\pi) = \Lambda_{par(ch(i))\backslash i}\left[\Lambda_{ch(i)}\left[\mu_{ch(i)}^{ch(ch(i))}p(ch(\theta_i)\Big|pa(ch(\theta_i)),\pi)\right]\mu_{ch(i)}^{pa(ch(i))\backslash i}\right] \end{aligned} \tag{3}$$

The $\Lambda$ operator here is introduced to indicate either summation or integration—depending upon whether we are dealing with continuous or discrete variables—with respect to the variables indicated by the subscript. The recursive definitions of the messages ($\mu$) in Equation (3) are exactly the definitions that underwrite belief-propagation inference schemes, including the sum-product algorithm [24]. Superscripts and subscripts on the messages indicate the node or variable from which the message comes, and to which it is sent, respectively. In what follows, we consider the nature of the requisite messages.

## 3. A Worked Example

In Figure 2, we illustrate a simple worked example based on Equation (3) using a Bayesian network of normally distributed variables. The network includes three nodes. These are: the choice ($\pi$) we make about how to sample our data, the parameter ($\theta$) we seek information about, and our data ($y$). The Bayesian network is first transformed into a factor graph—with explicit representation of factors of the joint probability distribution as boxes—and then to a (normal) Forney factor graph [25] that omits the variable nodes (circles). The Forney factor graph formulation is perhaps the easiest to use when visualising the passing of messages along the edges of the graph. The messages can then be used to construct the two entropies we need to compute expected information gain.

The model in this toy example has the following factors:

$$\begin{aligned} p(y|\theta,\pi) &= \mathcal{N}(\theta\cdot\cos(\pi),1) \\ p(\theta) &= \mathcal{N}(0,1) \end{aligned} \tag{4}$$

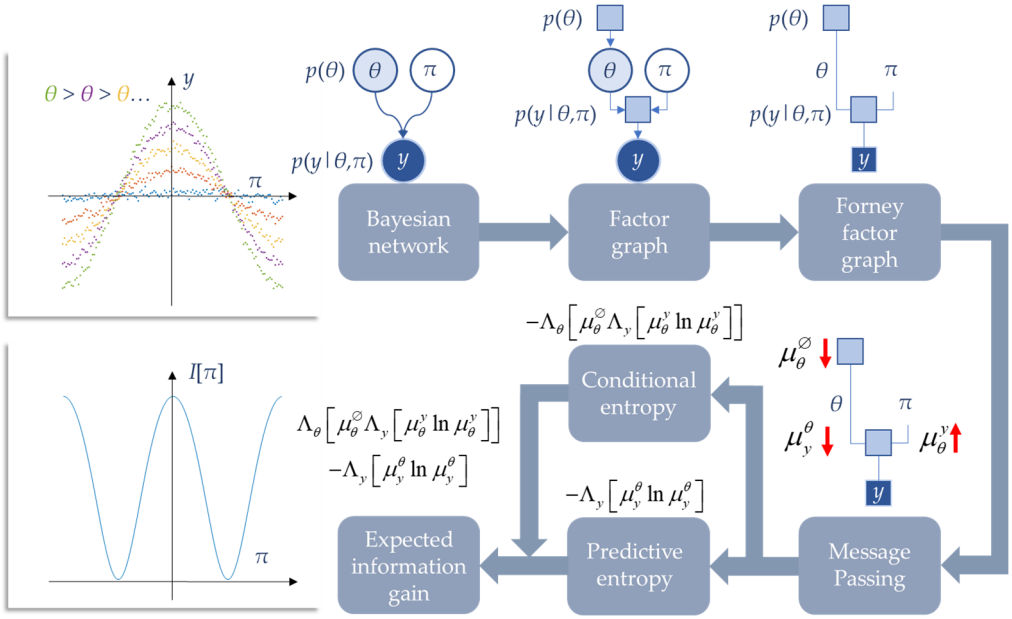


**Figure 2.** A worked example. This figure offers a worked example of how we compute expected information gain. It draws from the message passing perspective set out in Equations (3) and (4) which detail the analytic computations that lead us to expected information gain as a function of our choices, $\pi$, for this generative model. The colours in the upper left plot represent different values of $\theta$. The red arrows in the lower right factor graph indicate the directions of messages being passed.

In effect, this model amplifies or attenuates the amplitude of the predicted data depending upon a periodic function of our data-sampling policy, $\pi$. The upper-left panel of Figure 2 depicts the data sampled from this model for several possible values of $\theta$. Computing the relevant messages for the information gain, we have (omitting normalising constants for simplicity and using $\varnothing$ to denote the empty set)

$$\begin{aligned} \mu_\theta^\varnothing &= e^{-\frac{1}{2}\theta^2} \\ \mu_y^\theta &= \Lambda_\theta\left[\mu_\theta^\varnothing e^{-\frac{1}{2}(y-\theta\cdot\cos(\pi))^2}\right] = e^{-\frac{1}{2}\left(\frac{y^2}{(\cos^2(\pi)+1)}\right)} \\ \mu_\theta^y &= e^{-\frac{1}{2}(y-\theta\cdot\cos(\pi))^2} \end{aligned} \tag{5}$$

The directions of these messages are illustrated in the bottom-right panel of Figure 2. We next substitute these into the first line of Equation (3) to find an expression for the information gain about the parameter $\theta$. Once all terms that are constant with respect to $\pi$ are eliminated, we are left with:

$$I[\pi]^\theta = \frac{1}{2}\ln(\cos^2(\pi)+1) \tag{6}$$

Equation (6) is a special case of the third row of Table 1, which highlights analytical expressions for expected information gain for a few common model structures. This function is plotted in the lower-left panel of Figure 2. As we might intuit, the most informative places to sample data align with those in which differences in $\theta$ lead to large differences in the predicted data—i.e., in which our choice of $\pi$ maximises the sensitivity with which $y$ depends upon $\theta$. Given that a periodic (cosine) function is used for the effect of how we chose to sample ($\pi$) the data ($y$), there are multiple peaks which coincide with the peaks and troughs of the periodic function.

**Table 1.** Expected information gain *.

| Prior | Likelihood | Expected Information Gain |
|---|---|---|
| $s \sim Cat(\mathbf{s})$ | $o\|s,\pi$ $\sim Cat(\mathbf{A}(\pi)s)$ | $I[\pi]^s = diag(\mathbf{A}(\pi)\cdot \ln \mathbf{A}(\pi))\cdot \mathbf{s}$ $-(\mathbf{A}(\pi)\mathbf{s})\cdot \ln(\mathbf{A}(\pi)\mathbf{s})$ |
| $\mathbf{A} \sim Dir(\mathbf{a})$ | $o\|\mathbf{A},\pi$ $\sim Cat(\mathbf{A}\pi)$ | $I[\pi]^{\mathbf{A}} \approx \frac{1}{2}\sum_i(\psi(\mathbf{a}_{i\pi}) - \psi(\sum_i \mathbf{a}_{i\pi}))$ $\times\left(\mathbf{a}_{i\pi}^{-1} - (\sum_i \mathbf{a}_{i\pi})^{-1}\right)$ |
| $x \sim \mathcal{N}(\eta, \Sigma_x)$ | $y\|x,\pi$ $\sim \mathcal{N}(M(\pi)x, \Sigma_y(\pi))$ | $I[\pi]^x = \frac{1}{2}\ln\left(M(\pi)\Sigma_x M(\pi)^T + \Sigma_y(\pi)\right)$ $-\frac{1}{2}\ln\Sigma_y(\pi)$ |

* *Cat* indicates a categorical distribution, *Dir* a Dirichlet distribution, and $\psi$ indicates a digamma function. See Appendix B of [26] for details of the approximation used in the expected information gain in the second row.

## 4. Function Approximation

We next turn to a generic supervised learning problem—that of trying to approximate some function based upon known inputs and the observable outcomes they (stochastically) cause. Our generative model for this is straightforward:

$$\begin{array}{c} P(y,\theta|\pi) = P(y|\theta,\pi)P(\theta) \\ P(\theta) = \mathcal{N}(\eta_\theta, \Sigma_\theta) \\ P(y|\theta,\pi) = \mathcal{N}(\phi_\pi\theta, \Sigma_y) \end{array} \qquad \Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \end{bmatrix} \tag{7}$$

The matrix ($\Phi$) comprises a set of concatenated row vectors ($\phi_i$). The columns of $\Phi$ are elements of some basis set. For the simulations that follow, these are Gaussian radial basis functions. In the MATLAB demo routines that accompany this paper, alternative basis sets (including cosine and polynomial) can be chosen. We can see this model as an approximation of an unknown function whose (discretised) input is the row index of $\Phi$ and whose output is the expected value of $y$ corresponding to that row:

$$\mathbb{E}[y|\theta,\pi] = \textstyle\sum_i \theta_i \Phi_{\pi,i} \triangleq \sum_i \theta_i \Phi_i(x=\pi) = f_\theta(x=\pi) \tag{8}$$

Our choice of $\pi$ is used to test alternative values of the input ($x$) to see the output ($y$) that the unknown function ($f$) generates. Here, $\pi$ indexes the discrete intervals of $x$ implicit in the columns of $\Phi$. Figure 3 illustrates a depiction of this model as a Bayesian network and a visual representation of the data-generating process. In this simulation, we move beyond our worked example in Figure 2 and consider multiple samples. In this setting, we replace our prior beliefs after each observation with our posterior beliefs, allowing a gradual refinement of our approximated function and the future choices on which this depends.

The posterior probability of the parameters following an input and the marginal probability of the data are:

$$\begin{array}{c} P(\theta|\pi,y) = \mathcal{N}\left(\Upsilon_\pi(\Sigma_\theta^{-1}\eta_\theta + \phi_\pi^T\Sigma_y^{-1}y), \Upsilon_\pi\right) \\ P(y|\pi) = \mathcal{N}\left(\phi_\pi\eta_\theta, \phi_\pi\Sigma_\theta\phi_\pi^T + \Sigma_y\right) \\ \Upsilon_\pi \triangleq (\phi_\pi^T\Sigma_y^{-1}\phi_\pi + \Sigma_\theta^{-1})^{-1} \end{array} \tag{9}$$

The middle-right panel of Figure 3 shows the predictive (marginal) probability distribution after 28 random choices for $\pi$. The lower-left panel shows the posterior probabilities of the parameters. The specific choices are shown in the lower-right panel. Even with these random choices, the 28 samples provide a reasonably good approximation to the shape of the function, as can be seen comparing the middle-left (large number of samples drawn from the function) and right panels. However, a closer look at the choices made reveals some clear inefficiencies. For instance, choices 2 and 4 and choices 6 and 7 are very close to

one another (circled in red). Intuitively, we would expect that for any (smooth) function, it will almost always be preferable to sample locations far away from one another, as some uncertainty will have been resolved by samples nearby.

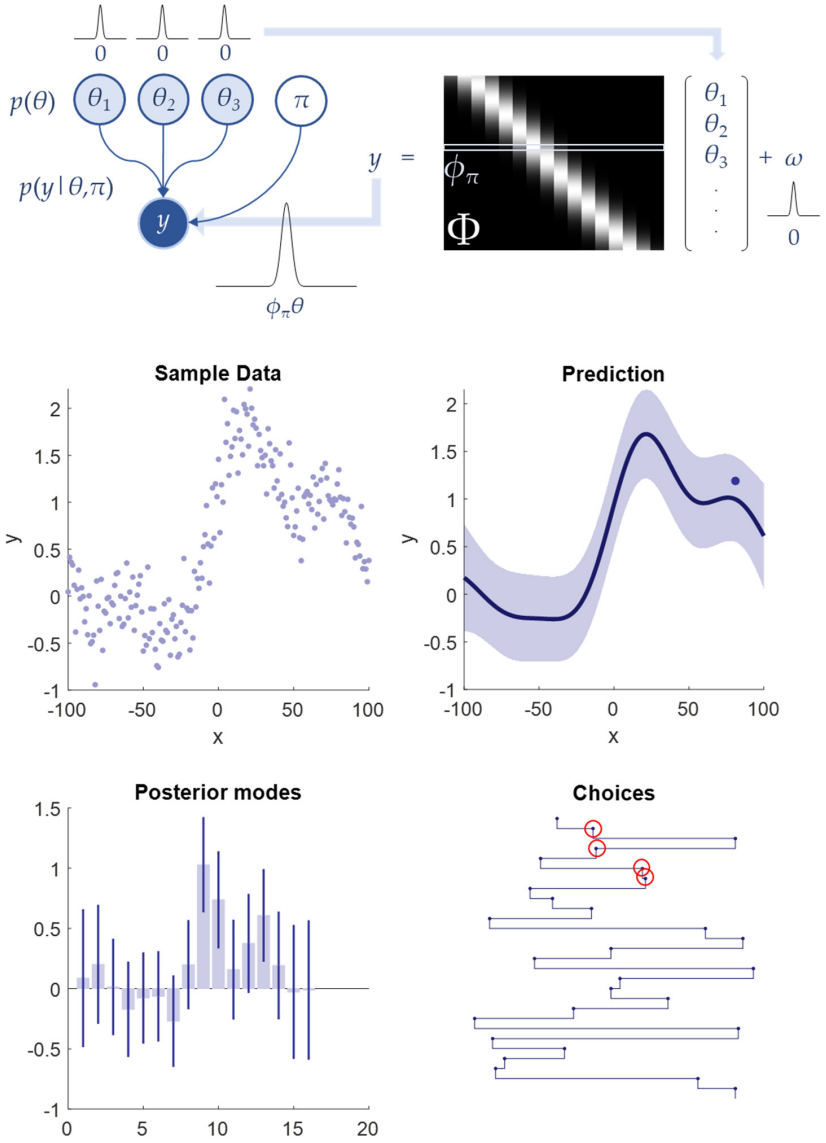


**Figure 3.** Function approximation with random sampling. The upper part of this figure illustrates the generative model schematically in terms of a Bayesian network. Here, a set of 16 parameters with zero-centred normally distributed priors weight the columns of a matrix of Gaussian basis functions (only three are shown in the schematic). The action we choose when generating our data selects which row of the matrix is used to generate that outcome. The middle-left panel shows the result of taking 200 or so samples from the underlying function that we seek to approximate. The middle-right panel shows the predictive distribution following the final (28th) sequential sample—shown as a blue dot—in terms of mode and 90% credible intervals. The lower-left panel shows the posterior parameter modes and 90% credible intervals, while the lower-right panel shows the sequence of choices (from top to bottom) aligned with the $x$-axis of the plot above. The red circle samples illustrate redundant or inefficient sampling under this random scheme.

This is where the information gain becomes relevant. Substituting Equation (7) into Equation (3), we have

$$I[\pi]^{\theta} = \frac{1}{2}\ln\left(\phi_{\pi}\Sigma_{\theta}\phi_{\pi}^{T} + \Sigma_{y}\right) - \frac{1}{2}\ln\Sigma_{y} \tag{10}$$

Figure 4 shows the same set-up as in Figure 3, but now samples are drawn from a distribution whose log is proportional to the information gain in Equation (10). During each iteration, a sample location is chosen from this distribution (a large proportionality constant makes this almost deterministic; i.e., the location with the highest information gain is almost always chosen). A datapoint is then sampled from the generative model under this choice. This is then used to update priors to posteriors using Equation (9). The process is then repeated up to 28 iterations.

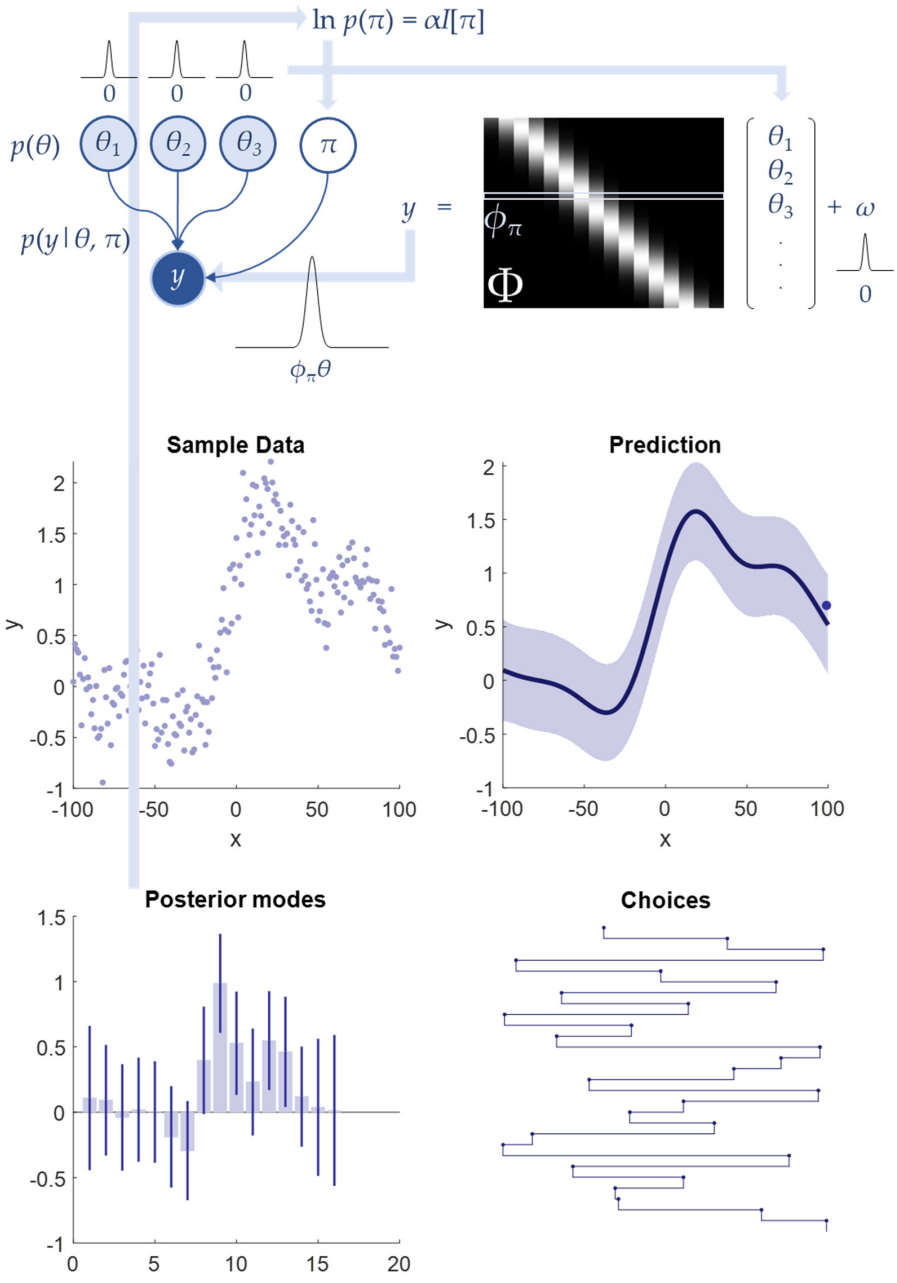


**Figure 4.** Function approximation with intelligent sampling. This figure uses the same format as Figure 3, but now each choice is sampled to maximise anticipated information gain. Note the more accurate approximation to the region between −100 and −50 in the middle-right panel. The $\alpha$ (i.e., precision or inverse temperature) parameter here is set to 64 to ensure almost deterministic sampling of the most informative locations.

Unlike the random sampling, it is not until late in the simulation that we revisit a previously sampled location. The implication is that we can achieve better inferences with the same, or perhaps even a smaller, number of samples when we select these samples in a smart way. This raises the question as to how many samples we should collect.

An answer to this question can be drawn from behavioural neuroscience and the so-called exploitation–exploration dilemma [27,28]. This deals with the situation in which certain parts of our environment may be particularly aversive or attractive—perhaps due to the presence of food or predators—but we do not know where these locations are to begin with. Initially, we must explore, seeking information about our environment, and then at some stage switch to exploiting the knowledge acquired, to ensure we occupy preferred locations. If all locations were equally preferrable, we could continue exploring forever, developing ever more precise beliefs about our world. The same is true when sampling data or performing experiments if there is no cost to performing this sampling. However, once we acknowledge the computational and resource costs to sampling, we conclude that if it were not for the potential information gain, it would be preferable to stop acquiring new data.

To accommodate this cost of sampling, we can simply augment the expected information gain with a prior belief about our propensity to stop sampling:

$$G = [I[\pi], 0] + C \tag{11}$$

Equation (11) treats the information gain as a vector of all the possible sampling locations and concatenates this with a zero element, which reflects the information gained if we were to stop sampling. The relative preference for sampling or not is then expressed in the $C$ vector, which assigns a prior (in the form of a log probability) to each choice of sampling location. In neurobiology, this expression is sometimes referred to as an expected free energy—reflecting its analogous form (KL-Divergence and log marginal) to the variational free energy functionals used in approximate inferences [29–31]. It allows us to combine information-seeking and cost-aversive imperatives into the same objective function. Figure 5 shows what happens when we select our data according to $G[\pi]$, where the preference for stopping sampling takes the value $1/4$ (i.e., $C = [0,\ldots,0,1/4]$ includes zeros for all but the 'stop sampling' option, which is assigned a value of $1/4$). This means that when the expected information gain is sufficiently low the chance of deciding to stop sampling increases.

In Figure 5, we see that the same first 15 choices are made as in Figure 4. However, at this point, a decision is made to stop sampling. This is because sufficient uncertainty has been resolved that the cost of sampling is greater than the potential information gain by sampling further. Depending upon how costly the sampling is assumed to be, this decision to stop will move earlier or later, and the quality of the function approximation will vary.

A reasonable question to ask at this stage is why bother with the full information-seeking objective? It is clear from Figures 4 and 5 that all we need to do is choose to sample from the regions with greatest predicted variance, sequentially reducing that variance until it is minimal throughout the domain of the function being approximated [32]. This suggests that the predictive entropy from Equation (2) is sufficient on its own (c.f., maximum entropy sampling [33]). However, Figure 6 offers a rebuttal to this with a process whose measurement noise increases in variance from the centre of the function domain. This means the amount of unresolvable uncertainty is heterogeneous throughout the domain of potential sampling. Were we to ignore this, the uncertainty left to resolve will always be greater than the cost of sampling. Using the full expected information gain tells us that, although there is uncertainty left, there is no information to gain, and allows for earlier termination of our sampling procedure. Furthermore, it leads to a methodical sampling strategy: sampling starts from the minimally ambiguous central location and fans out as local uncertainty is reduced until the capacity to resolve further uncertainty drops below the cost of acquiring further samples. This avoidance of sampling in ambiguous locations is sometimes referred to as a 'streetlight effect' [34] (see [35] for further discussion with

models comprising categorical distributions), which occurs only in the presence of a full expected information gain objective. The streetlight effect is the tendency to search where data are generated in a minimally ambiguous way—i.e., under a streetlamp compared to searching elsewhere on a darkened street.

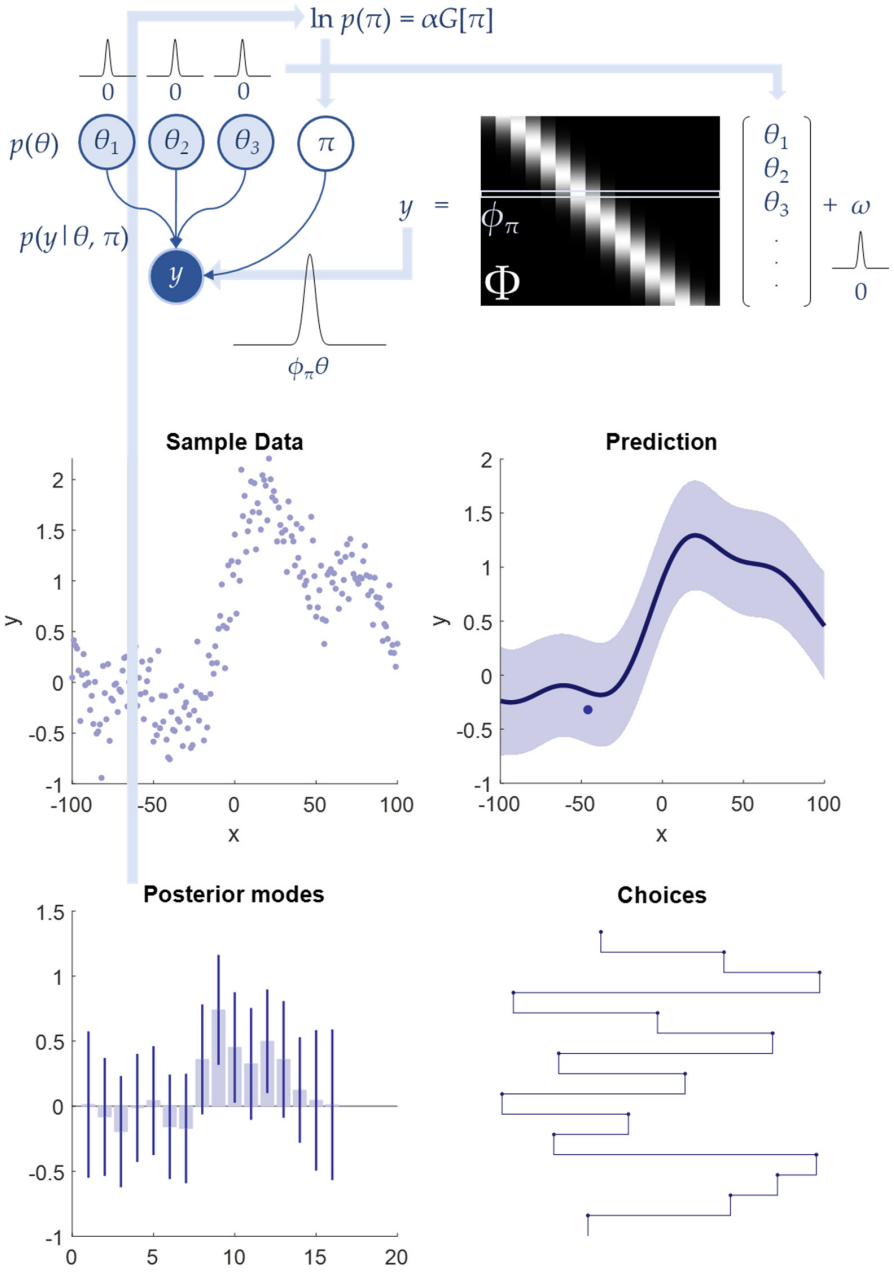


**Figure 5.** Function approximation with cost of sampling. Using the same format as in Figure 3, we here see the effect of equipping sampling with a cost. In the absence of any potential information gain, there is a preference for not sampling. The result of this is that the process of gathering data terminates after a smaller number of acquired samples. This termination occurs once the potential information gain falls below the cost of acquiring further samples. This means that the quality of the function approximation will depend upon the cost associated with sampling. In this example, a reasonable approximation to the function has been attained—inferior to that of Figures 3 and 4 but still capturing a broad outline of the function.

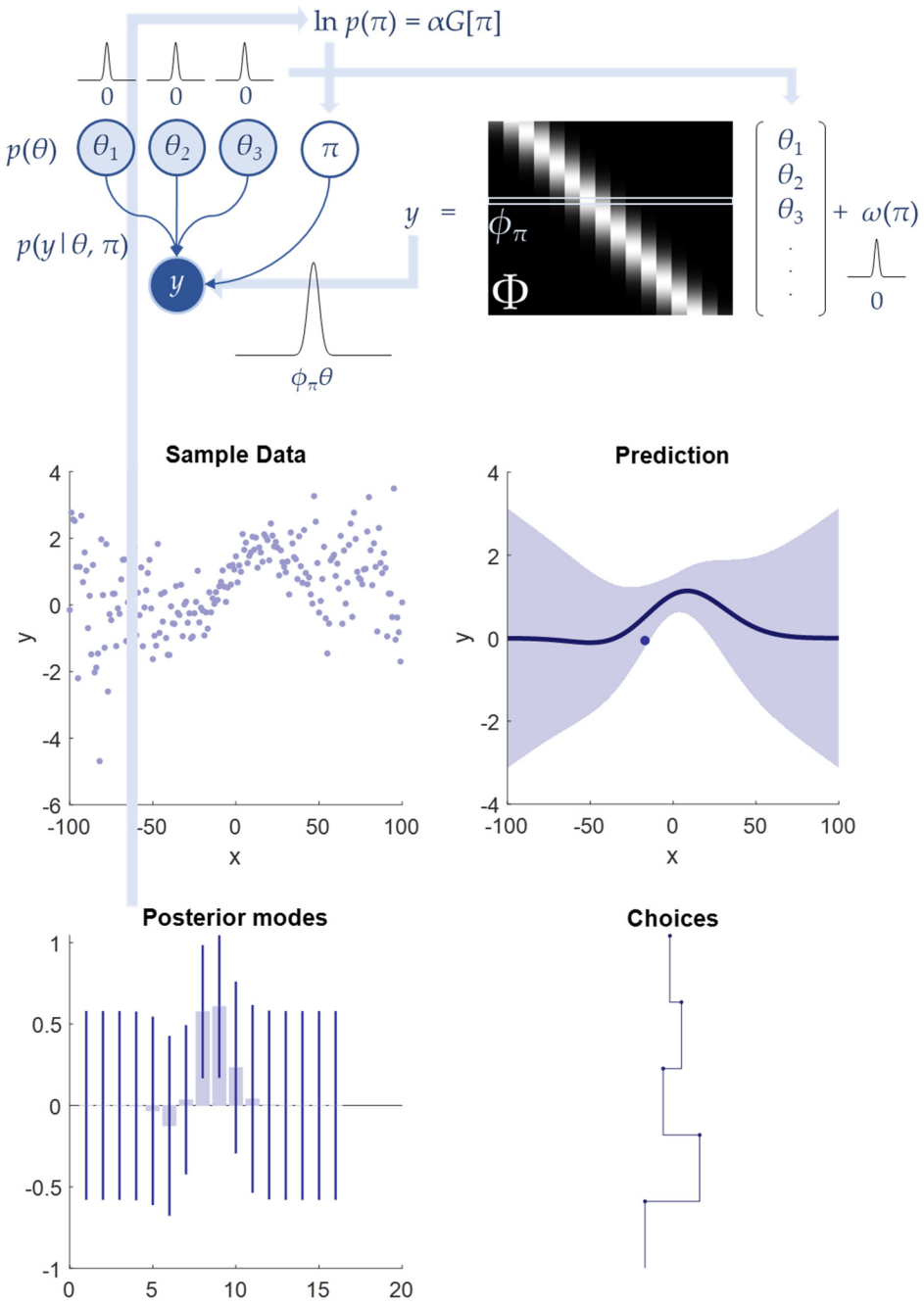


**Figure 6.** Function approximation with unresolvable uncertainty. This figure deals with the situation in which the variance associated with the likelihood distribution is heterogeneous along the *x*-axis. Specifically, it increases quadratically with distance from zero. The result of this is that less information gain is available at the extremes of the *x*-axis. This leads to termination of sampling despite the relatively larger uncertainty about the function in these extremes. This is optimal in this context as there is a limit to the amount of additional uncertainty that can be resolved.

## 5. Dynamic Processes

In this section, we consider processes that evolve in time. For example, while the approach in Section 3 might be appropriate for mapping out the heights of a mountain range, it would not be suitable for measuring (for example) weather conditions or tectonic activity in different locations across that terrain, as these will change with time (although see [36,37] for closely related approaches to placing static sensors for environmental monitoring). To move beyond a temporal snapshot, we must extend our models to include dynamics.

There are many ways to do this. One could set up a hidden Markov model, whose states probabilistically evolve according to discrete transition probabilities [38]. An alternative would be to formulate a model in terms of a differential equation that determines the time evolution of hidden states generating data [39]. A final option is to extend the model we used in Section 4 to include a temporal dimension. Specifically, in place of our one-dimensional set of basis functions, we now use two-dimensional basis functions with both temporal and spatial components. While we employ the third of these methods in this paper, each of these three approaches is perfectly valid.

$$\begin{array}{c} P(y,\theta|\pi) = P(y|\theta,\pi)P(\theta) \\ P(\theta) = \mathcal{N}(\eta_\theta, \Sigma_\theta) \\ P(y|\theta,\pi) = \mathcal{N}\big((\phi_\tau \otimes \phi_\pi)\theta, \Sigma_y\big) \end{array} \qquad \Phi = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \vdots \end{bmatrix} \tag{12}$$

Equation (12) can be interpreted similarly to Equation (8), in which the expectation of the data is treated as a function approximation, which now includes a time argument.

$$\begin{array}{r} \mathbb{E}[y|\theta,\pi,\tau] = (\Phi\Theta\Phi^T)_{\tau,\pi} = \sum_{i,j} \Phi_{\tau,i}\Theta_{i,j}\Phi_{\pi,j} \triangleq f_\theta(x=\pi, t=\tau) \\ vec(\Phi\Theta\Phi^T) = (\Phi \otimes \Phi)vec(\Theta) \\ \theta = vec(\Theta) \end{array} \tag{13}$$

However, the form of Equation (12) means we can use the same inferential machinery, and the same information gain calculations, as we did for the static inference examples. Figure 7 shows a graphical representation of the matrices involved in generating our data and the inferences obtained after sampling. Every eight time steps, a sample is chosen to maximise the expected information gain, and this is used to update posterior beliefs exactly as previously. As can be seen from the lower-right panel in Figure 7, an accurate inference has been obtained of the evolution of our function over time. The choices made when sampling provide a relatively even coverage of the spatial dimension. Note the implicit inhibition of return, which prevents resampling near recently sampled locations. However, once sufficient time has elapsed, it is worth returning to previously sampled locations as they may have changed since previously being sampled.

Figure 8 shows several snapshots of the inferences drawn at different time points during the simulation reported in Figure 7. This is designed to provide some intuition as to the filling in of predictions as more observations are made. Note that because this is a dynamic process—with a degree of smoothness both in space and time—predictions based upon the current data can be used to inform predictions about nearby spatial locations and to both predict and postdict the values of the function at different points in time.

In this and the previous section, we have demonstrated the way in which smart or optimal sampling may be used to select data in a manner that balances the cost of sampling or of performing further experiments against the information gained from those samples or experiments. Each of these examples has relied upon relatively simple, and analytically comfortable, linear Gaussian systems. Next, we address active sampling in a situation where analytical solutions are no longer possible.

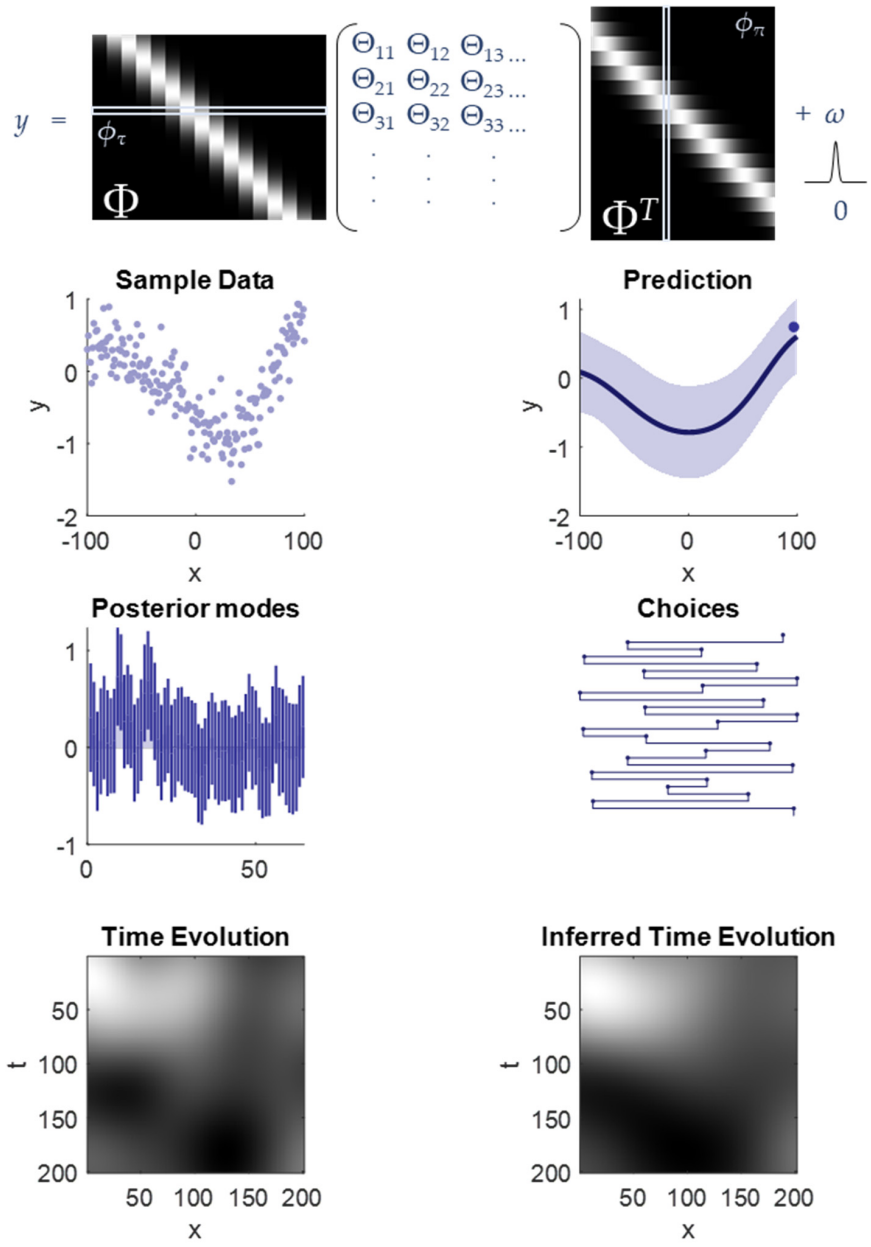


**Figure 7.** Dynamic information seeking. This figure deals with the situation in which the function we are trying to map out changes with time. The generative model is summarised in the upper part of this figure as the product of a matrix of coefficients with two matrices of basis functions. The data generated now depend both upon our choice and upon time. The plot of sample data shows data we could have sampled from the function at the final time point by taking the final row of the plot in the lower left and generating samples from this for each possible sampling location. This is to afford a comparison between a relatively exhaustive sampling of the data and the inferred shape of the function in the prediction plot based upon only one sample from this time point. The prediction plot shows the predicted data at this time point. The plot of posterior modes shows the estimated coefficients with 90% credible intervals at the final time point. The plot of choices is like that shown in previous figures. However, each choice occurs eight time steps after the previous choice. The lower two plots show both the time evolution of the underlying function generating the data and the inferred (expected) values of this function at each time point. We see that the shape of this function at different time points, despite only the sparse sampling of the data, is well captured by this scheme.

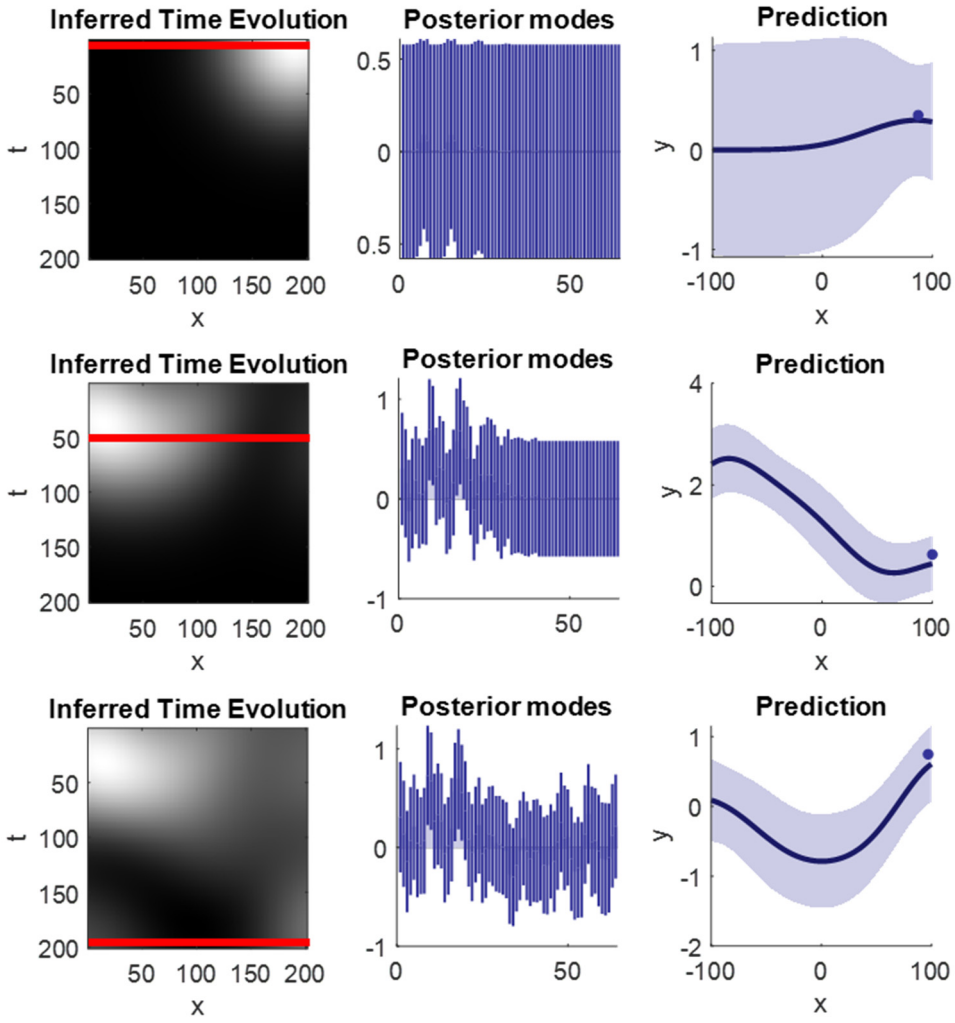


**Figure 8.** Trajectory of beliefs. This figure captures several snapshots of the time-evolving inference process summarised in Figure 7. The first row shows the inferences made following the first data sample. The second row shows the inferences made at time step 50. The third shows the inferences made by the end of the time period. Note the improvement in predictive precision in the second and third rows, corresponding to the use of prior information from previous time steps.

## 6. Clinical Trials

In our final example, we demonstrate the potential utility of the ideas outlined above in the context of a more concrete example. Adaptive Bayesian clinical trials [40] have been increasingly popular in efficiently answering questions about medical interventions in rapidly evolving healthcare situations. A prominent example of this was the use of an adaptive Bayesian methodology to assess interventions during the 2014 West African Ebola outbreak [41]. A key argument made in favour of this design was that in a rapidly developing epidemic it is necessary to be able to efficiently compare several possible treatments and to be able to adapt as new information emerges [42]. For instance, we may wish to stop one treatment arm, or perhaps reallocate participants to another treatment arm, if there is early evidence of the inferiority or superiority of this treatment, respectively. These benefits are not restricted to the management of epidemics but may apply more broadly. See, for example [43–49].

We have two reasons for choosing a clinical trial design as our final example. The first is that they are common hypothesis-driven experimental designs with resource limitations. The second is that they offer us an opportunity to go beyond the illustrative linear models used above and force us to deal with more expressive model structures that require

pragmatic approximations to be able to apply the methods outlined above efficiently. Our hope is to demonstrate that use of expected information gain does not restrict us only to the simplest model classes but has broad applicability in real-world problems.

The active sampling approach advocated in this paper offers two main opportunities to augment adaptive trial designs. First, it allows us to adapt the design (e.g., inclusion criteria and assignment to a treatment arm) to maximise the information we obtain about treatment efficacy. Second, it allows us to balance this information gain against various costs, namely, stopping the trial when the potential information gain no longer exceeds the cost of acquiring further data. Alternatively, it might include a preference for good clinical outcomes and aversion for bad outcomes—prompting adjustments of the allocations between treatment arms to maximise benefit as we infer more about the effect of the intervention. This blurs the line between clinical trial and public health intervention and can be seen as analogous to animal behaviour that is never fully exploitative or explorative but is a balance between the two.

Our setup is as follows: for each new cohort of participants, we decide upon the randomisation ratio to adopt and the time(s) at which we follow them up. To simplify for illustration purposes, we assume each cohort is followed up only once, noting that the same principles can be applied to multiple follow-up points. Our generative model can be expressed as follows:

$$\begin{aligned} p(y|\theta, v, d, \pi_\tau) &= Cat([\rho, 1-\rho]) \\ p(d) &= Cat\left(\left[\tfrac{1}{2}, \tfrac{1}{2}\right]\right) \\ p(v|\pi_r) &= Cat(\pi_r) \\ p(\theta) &= \mathcal{N}(\eta_\theta, \Sigma_\theta) \\ \rho &= \prod_i^{\pi_\tau} \tfrac{1}{1+e^{-X_i\theta}} \\ X_i &= [\phi_i, d, v] \end{aligned} \tag{14}$$

In this model, the data ($y$) are binary values reflecting survival up to the measurement time or death before this point. The probability of survival to a given time is the product of the survival probabilities for all previous time steps. Our probability for survival for a given time step is a logistic function that determines the effect of time since enrolment in the trial, the effect of demographic (here, sex), and the treatment effect. As in our previous examples, the time evolution is modelled using a Gaussian basis set. Patient demographics ($d$) are sampled from a categorical distribution. There are two choices that can be made. The first is the time at which follow up is performed (subscript $\tau$). The second is the randomization ratio (subscript $r$). There are assumed to be three options for this ratio. These are $1/3$, $1/2$, or $2/3$ allocated to the treatment group. The allocation is represented by the symbol $v$. We assume these decisions are taken before each cohort, where each cohort consists of eight participants. The parameterisation of the treatment effect (i.e., the final element of $\theta$) can be interpreted as a hazard ratio, as might be estimated in a Kaplan–Meier [50] analysis, and is consistent with Cox proportional hazards assumptions [51].

This model differs from those in previous sections in that it includes highly nonlinear functions and is not based upon conjugate priors. This makes analytical solutions to the inference problem intractable. In place of this, we employ a variational approximation to arrive at posterior estimates. This involves a Newton optimization scheme (also known as Variational Laplace) applied to maximise a lower bound on the log marginal likelihood. The relevant gradient expression for this scheme is as follows:

$$\begin{aligned} \nabla_\theta \rho &= \rho \textstyle\sum_j \frac{X_j^T}{1+e^{X_j\theta}} \\ p(y|\ldots) &= y\rho + (1-y)(1-\rho) \\ \nabla_\theta p(y|\ldots) &= (2y-1)\nabla_\theta \rho \\ \nabla_\theta \ln p(y\Big|\ldots) &= p\,(y|\ldots)^{-1} \nabla_\theta p(y\Big|\ldots) \end{aligned} \tag{15}$$

The Hessian is as follows:

$$\begin{aligned}\mathbf{H}_\theta[\rho] &= \rho^{-1}\nabla_\theta\rho\nabla_\theta\rho^T - \rho\textstyle\sum_i \frac{X_i X_i^T e^{X_i\theta}}{(1+e^{X_i\theta})^2}\\ \mathbf{H}_\theta[p(y|\ldots)] &= (2y-1)\mathbf{H}_\theta[\rho]\\ \mathbf{H}_\theta[\ln p(y|\ldots)] &= p\left(y|\ldots\right)^{-1}\mathbf{H}_\theta[p(y|\ldots)] - p\left(y|\ldots\right)^{-2}\nabla_\theta p(y\Big|\ldots)\nabla_\theta p\left(y|\ldots\right)^T\end{aligned} \tag{16}$$

The variational optimization scheme then takes the form (where superscripts indicate iteration number)

$$\begin{aligned}\vartheta^{i+1} &= \vartheta^i + \left(\Sigma_\theta^{-1} - \mathbf{H}_{\vartheta^i}\left[\ln p(y\Big|\vartheta^i, v, d)\right]\right)^{-1}\left(\nabla \ln p(y\Big|\vartheta^i, v, d) + \Sigma_\theta^{-1}(\eta_\theta - \vartheta^i)\right)\\ q(\theta) &= \mathcal{N}\left(\vartheta^{\max(i)}, \left(\Sigma_\theta^{-1} - \mathbf{H}_{\vartheta^{\max(i)}}\left[\ln p(y\Big|\vartheta^{\max(i)}, v, d)\right]\right)^{-1}\right)\end{aligned} \tag{17}$$

Figure 9 shows a simulation of this design with random sampling of both follow-up time and randomization ratio. The variational inference scheme here terminates following 16 iterations. The upper part of the figure shows the form of the generative model as a Bayesian network. The process used to simulate data is summarised in the plots of survival against time for each of the combinations of treatment versus placebo and male versus female. In this example, the treatment is harmful, leading to a lower survival in the treatment compared to the placebo group. This pattern is captured by the inference scheme following 16 cohorts and a relatively even sampling throughout time and randomization patterns.

Next, we wish to employ the information gain prior, to deliberately choose follow-up times and randomization ratios to maximise our information about the trajectories of each patient group. As above, the first step in doing so is to identify the form of the messages required for Equation (3).

$$I[\pi]^{\theta,d,v} = \Lambda_{\theta,d,v}\left[\mu^{\varnothing}_{\theta,d,v}\Lambda_y\left[\mu^{y}_{\theta,d,v}\ln\mu^{y}_{\theta,d,v}\right]\right] - \Lambda_y\left[\mu_y^{\theta,d,v}\ln\mu_y^{\theta,d,v}\right] \tag{18}$$

We can simplify the estimation of these messages using a local quadratic approximation for the continuous parameters. This technique is sometimes referred to as Variational Laplace and is implicit in Equation (17) [52,53].

For the messages in the first term of Equation (18), we employ a quadratic expansion of the entropy around the posterior mode of the parameters:

$$\begin{aligned}\mu^{\varnothing}_{\theta,d,v} &= p(\theta)p(d)p(v\Big|\pi_r)\\ \mu^{y}_{\theta,d,v} &= y\rho + (1-y)(1-\rho)\\ \Lambda_{\theta,d,v}\left[\mu^{\varnothing}_{\theta,d,v}\Lambda_y\left[\mu^{y}_{\theta,d,v}\ln\mu^{y}_{\theta,d,v}\right]\right] &\approx \textstyle\sum_{d,v} p(d,v\Big|\pi_r)\Bigg(\begin{bmatrix}\rho_{\theta=\vartheta}\\ 1-\rho_{\theta=\vartheta}\end{bmatrix}\cdot\begin{bmatrix}\ln\rho_{\theta=\vartheta}\\ \ln(1-\rho_{\theta=\vartheta})\end{bmatrix} +\\ &\quad \tfrac{1}{2}tr\left(\Sigma_\theta\left(\mathbf{H}_\theta[\rho]\ln\tfrac{\rho}{(1-\rho)} + \tfrac{1}{\rho(1-\rho)}\nabla\rho\nabla\rho^T\right)_{\theta=\vartheta}\right)\Bigg)\end{aligned} \tag{19}$$

The second term of Equation (18) has the following form when the $\rho$ function has been expanded to the second order:

$$\begin{aligned}\mu_y^{\theta,d,v} &= \begin{bmatrix}\mathbb{E}_{p(\theta,d,v|\pi_r)}[\rho] & 1-\mathbb{E}_{p(\theta,d,v|\pi_r)}[\rho]\end{bmatrix}^T\\ \mathbb{E}_{p(\theta,d,v|\pi_r)}[\rho] &\approx \textstyle\sum_{v,d} p(v\Big|\pi_r)p(d)\left(\rho_{\theta=\vartheta} + \tfrac{1}{2}tr(\Sigma_\theta\mathbf{H}_\theta[\rho]_{\theta=\vartheta})\right)\\ \Lambda_y\left[\mu_y^{\theta,d,v}\ln\mu_y^{\theta,d,v}\right] &\approx \begin{bmatrix}\mathbb{E}_{p(\theta,d,v|\pi_r)}[\rho]\\ 1-\mathbb{E}_{p(\theta,d,v|\pi_r)}[\rho]\end{bmatrix}\cdot\begin{bmatrix}\ln\mathbb{E}_{p(\theta,d,v|\pi_r)}[\rho]\\ \ln(1-\mathbb{E}_{p(\theta,d,v|\pi_r)}[\rho])\end{bmatrix}\end{aligned} \tag{20}$$

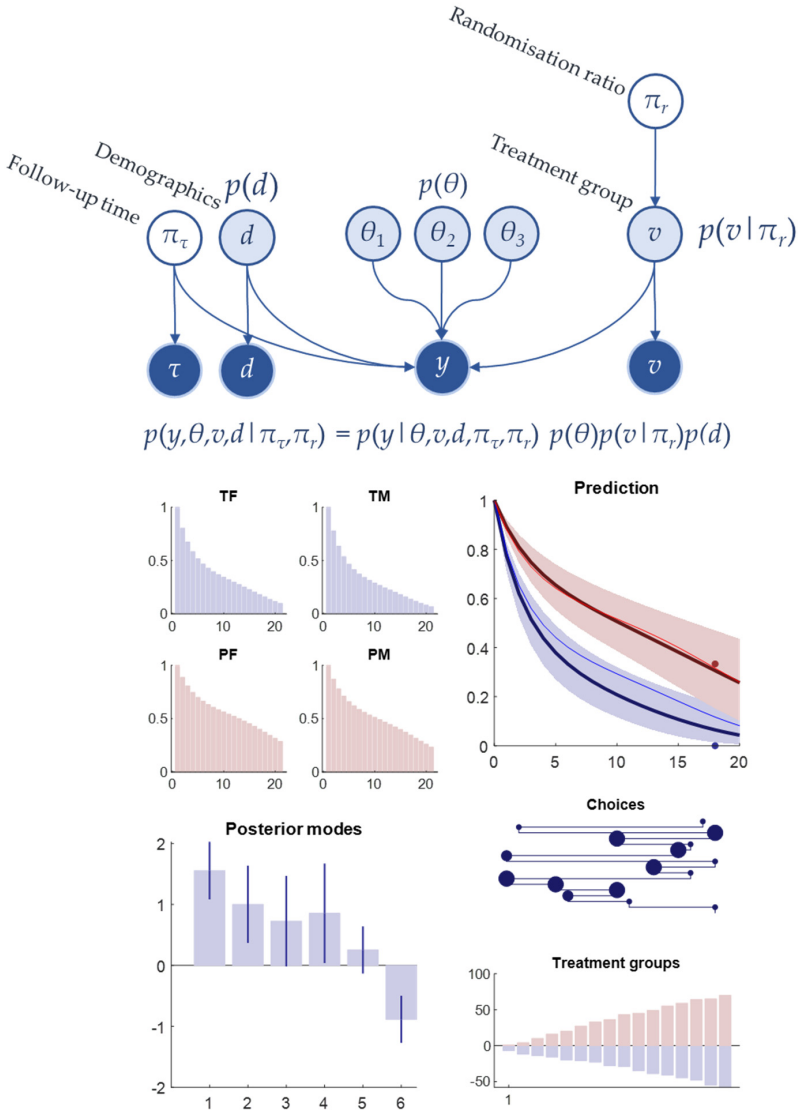


**Figure 9.** Random sampling for a Randomised Control Trial. The upper part of this figure shows the form of the generative model as a Bayesian network with the factors shown below. The underlying dynamics of the simulated trial are shown in the four plots on the middle left. These show the expected survival curves for each combination of demographics and treatment group. The letter T indicates the treatment group, the letter P indicates the placebo group, the letter F indicates females, and the letter M indicates males. The middle-right plot shows predicted survival from the time of enrolment (the units of the time axis are arbitrary). The thin red and blue lines give the true survival curves based upon the underlying generative process. The predictions are shown as thicker lines with accompanying 90% credible intervals. Below this is a plot of choices which takes the same format as in previous figures but now also reports the choice randomization ratio. This is shown by the size of the markers for each choice. There are three sizes of marker representing, from small to large, 2/3 allocated to a treatment group, 1/2 allocated to a treatment group, and 1/3 allocated to a treatment group. The lower-right plot shows the cumulative number of people allocated to the treatment (blue) and placebo (red) groups. Finally, the lower-left plot shows the posterior estimates of the parameters, with parameter five representing the effect of demographic on survival and parameter six representing the effect of treatment on survival.

Please see Equations (15) and (16) for the forms of the relevant gradients and Hessians. Figure 10 illustrates the same setup as in Figure 9 but now using the expected information gain from Equation (18) to guide sampling of data. Due to the approximations made to ensure tractability, we reduce the temperature (i.e., $\alpha$ from Figure 4) to four to acknowledge

that there is some uncertainty as to whether the best choices are those consistent with the maximal value of Equation (18). There are some notable differences between the choices made in Figure 10 compared to Figure 9. The most obvious of these is that the follow-up times selected have been moved later once optimal sampling is employed. This makes intuitive sense as a later follow-up time is informative about the survival probabilities at all prior times, whereas an earlier follow-up time is not informative about survival probabilities at later times. This has led to a marginal improvement in estimation of the hazard functions. In both cases, the number of people assigned to each treatment group appears approximately equivalent, as would be expected under random sampling, with an expectation of 50% assignment to each group overall.

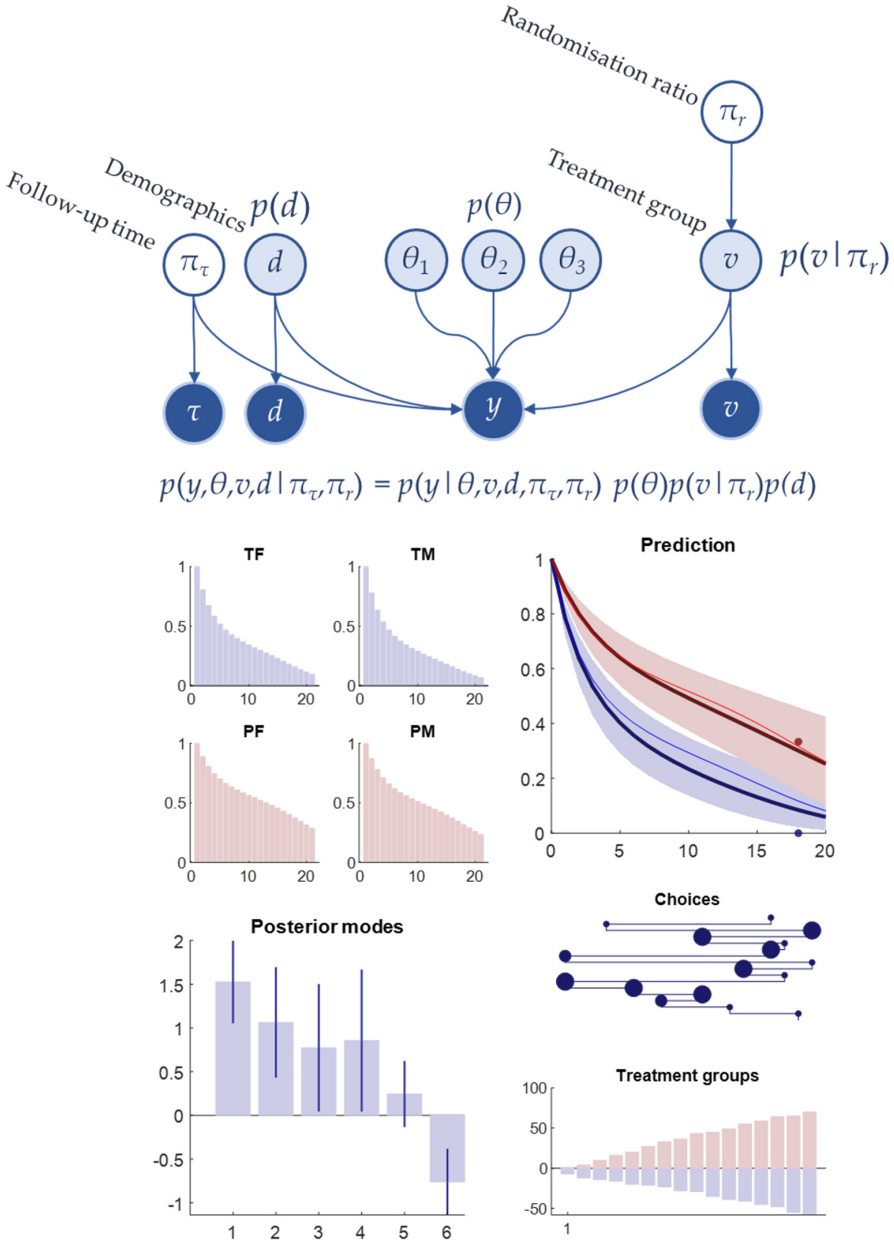


**Figure 10.** Intelligent sampling for a Randomised Control Trial. Using the same format as Figure 9, this figure shows the result of active sampling that maximises (subject to some random noise) the expected information gain about the parameters. This leads to a slightly better predicted survival curve, which appears to be due to a tendency to select later follow-up times. Later times tend to have greater potential information gain as survival to a later time is informative about previous survival probabilities.

Our final simulation now incorporates the preferences discussed in previous sections. Here the preferences are expressed in terms of a prior probability that survival is more likely than death. This must be nuanced as a preference for observing survival alone would favour observations at the point of enrolment where the probability of survival is 100%. To finesse this problem, we specify a preference for survival that increases exponentially in time. In other words, it is preferable to observe somebody having survived until the end of the potential follow-up period compared to having observed their survival shortly after enrolment. As shown in Figure 11, this preference for long-term survival results in longer follow-up times and, interestingly, once an inference that the treatment is potentially harmful has been made, there is a gradual shift towards randomising more people towards the placebo group. This implies a similar reasoning to when clinical trials are terminated early when a clear benefit of membership of one of the two treatment arms has been identified.

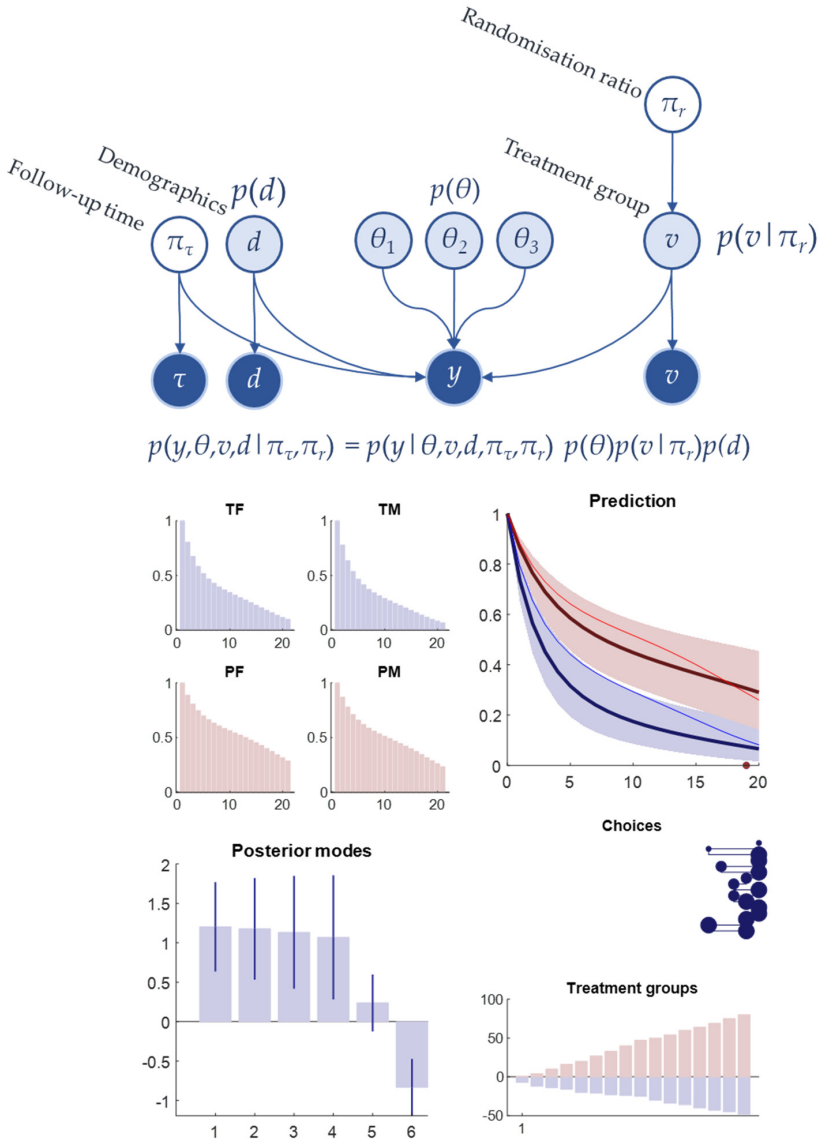


**Figure 11.** Sampling with preference for long-term survival. Again, using the format of Figure 9, this figure demonstrates the effect of preferring long-term survival. As might be expected, this bias is the follow-up time towards the end of the trial period. Perhaps more interestingly, although initially randomising to both treatment arms, once it becomes apparent that long-term survival is compromised in the treatment group, these preferences bias randomization towards the placebo group.

## 7. Discussion

This paper's focus has been on illustrating how we might make use of information-seeking objectives—augmented with costs or preferences—to choose the best data to optimise our inferences. While we have highlighted specific examples of active sampling or optimal Bayesian design, we have not provided a systematic analysis of the benefits of this approach here—instead favouring demonstrations of application to different sorts of models and problems. We have considered the problem of function approximation when there is a cost to data acquisition, or to the computational demands when sampling data from large datasets. We extended this to consider dynamic inference when the data obtained from the same experiment, address, or abstract location may change as a function of the time at which they are acquired. Finally, we illustrated the use of these methods even when the problem is not reducible to a general linear model.

It is worth noting that several of our examples would yield identical results had we adopted a data-sampling approach based upon maximum entropy sampling [54,55]. However, the key differences emerge when the variance around predicted outcomes is inhomogeneous. We illustrated this both in the static inference example in Figure 6 and also in our clinical trial example. Note that the covariance associated with the likelihood distribution is a function of the Hessian of the log likelihood. Equation (15) shows that this is not constant in the nonlinear setting considered in our final example.

There are several technical points worth considering for how we might advance the concepts reviewed in this paper. These are broadly: (1) refinement of the active selection process, (2) empirical evaluation of active versus alternative sampling methods, and (3) identifying the appropriate cost functions. Sophisticated inference is a method that might address the first of these [56]. This deals with the situation in which there is a dependence between the information gained from one choice and the potential information that can be gained following future decisions. In addition to this, it can help prioritise those sources of information that minimise cost in the future. Essentially this involves solving a recursive planning problem, like the approach used to solve Bellman optimality problems. The utility of sophisticated recursions is obvious when we consider how we would approach a clinical trial with multiple follow-up points, as the timing of the second follow-up will depend upon the information gain we anticipate from the first.

The issue of empirical validation is one that can be addressed through two routes. The first is in selecting a limited number of samples from a large dataset through either an active sampling approach or through some alternative data selection strategy. By using the posterior probability distributions under each sampling strategy as priors and comparing the marginal likelihood for the two sets of priors—assessed against the remainder of the dataset—we could evaluate the sampling strategy that leads to the best model. The second route to assessing the utility of active sampling is through comparing the energy costs and computing time required to achieve a reasonable approximation of a function using a limited dataset compared to that required to take account of a larger and more comprehensive dataset.

The identification of appropriate cost functions may be a trickier problem as these will vary from application to application and with the resources available. This has particular significance for clinical trial design, where the relative preference for different outcomes clearly has important ethical connotations. One approach would be to reverse-engineer the implicit cost functions employed in previous experimental designs (for example, the cost functions that would lead to a given early stopping criteria). An alternative is to find a way to synthesise and formalise views from the relevant stakeholders. For an interesting example of this in the context of determining outcomes for cystic fibrosis trials, see [57,58].

## 8. Conclusions

In this paper, we have sought to demonstrate that an appeal to the action–perception cycles seen in biology may be helpful in addressing the difficulties of managing very large datasets or in acquiring costly experimental data. The key ideas involved an appeal to

foveal-like sampling of small portions of the total available data to minimise computational cost, the ways in which we can select the most informative data, and the way in which we can trade this off directly against the costs of sampling further datapoints. Drawing from problems of exploitation and exploration—and making use of the same expected information gain measures that seem to be used in biological cognition—we have seen that it is possible to trade off the benefits of uncertainty resolution against its costs. A possible area of application for this is in the realm of Bayes-adaptive clinical trial design. Healthcare research inherently balances desired health outcomes with the resolution of uncertainty about how to achieve these. However, the broader imperative to balance the costs and benefits of how we sample data is likely to become more acute in an era of very large datasets and energy-hungry models.

**Author Contributions:** Conceptualization, T.P., K.F. and P.Z.; software, T.P.; formal analysis, T.P., K.F. and P.Z.; writing—original draft preparation, T.P.; writing—review and editing, K.F. and P.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** TP is supported by an NIHR Academic Clinical Fellowship (ref: ACF-2023-13-013).

**Data Availability Statement:** The MATLAB scripts used to generate the figures in this paper are available from https://github.com/tejparr/Active-Data-Selection/ (accessed 9 March 2024). These have been tested with MATLAB 2023a.

**Conflicts of Interest:** The authors declare no conflicts of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

## References

1. Mirza, M.B.; Adams, R.A.; Mathys, C.D.; Friston, K.J. Scene Construction, Visual Foraging, and Active Inference. *Front. Comput. Neurosci.* **2016**, *10*, 56. [CrossRef]
2. Yang, S.C.-H.; Wolpert, D.M.; Lengyel, M. Theoretical perspectives on active sensing. *Curr. Opin. Behav. Sci.* **2016**, *11*, 100–108. [CrossRef] [PubMed]
3. Zweifel, N.O.; Hartmann, M.J.Z. Defining "active sensing" through an analysis of sensing energetics: Homeoactive and alloactive sensing. *J. Neurophysiol.* **2020**, *124*, 40–48. [CrossRef] [PubMed]
4. Bajcsy, R. Active perception. *Proc. IEEE* **1988**, *76*, 966–1005. [CrossRef]
5. Crimaldi, J.; Lei, H.; Schaefer, A.; Schmuker, M.; Smith, B.H.; True, A.C.; Verhagen, J.V.; Victor, J.D. Active sensing in a dynamic olfactory world. *J. Comput. Neurosci.* **2022**, *50*, 1–6. [CrossRef]
6. Itti, L.; Baldi, P. Bayesian surprise attracts human attention. *Vis. Res.* **2009**, *49*, 1295–1306. [CrossRef]
7. Denzler, J.; Brown, C.M. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 145–157. [CrossRef]
8. Fuster, J.N.M. Upper processing stages of the perception–action cycle. *Trends Cogn. Sci.* **2004**, *8*, 143–145. [CrossRef]
9. Patterson, D.; Gonzalez, J.; Le, Q.; Liang, C.; Munguia, L.-M.; Rothchild, D.; So, D.; Texier, M.; Dean, J. Carbon emissions and large neural network training. *arXiv* **2021**, arXiv:2104.10350.
10. Henderson, P.; Hu, J.; Romoff, J.; Brunskill, E.; Jurafsky, D.; Pineau, J. Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.* **2020**, *21*, 10039–10081.
11. Rillig, M.C.; Ågerstrand, M.; Bi, M.; Gould, K.A.; Sauerland, U. Risks and Benefits of Large Language Models for the Environment. *Environ. Sci. Technol.* **2023**, *57*, 3464–3466. [CrossRef]
12. Strubell, E.; Ganesh, A.; McCallum, A. Energy and policy considerations for deep learning in NLP. *arXiv* **2019**, arXiv:1906.02243.
13. MacKay, D.J.C. Information-Based Objective Functions for Active Data Selection. *Neural Comput.* **1992**, *4*, 590–604. [CrossRef]
14. Lindley, D.V. On a Measure of the Information Provided by an Experiment. *Ann. Math. Statist.* **1956**, *27*, 986–1005. [CrossRef]
15. Zeidman, P.; Kazan, S.M.; Todd, N.; Weiskopf, N.; Friston, K.J.; Callaghan, M.F. Optimizing Data for Modeling Neuronal Responses. *Front. Neurosci.* **2019**, *12*, 986. [CrossRef]
16. Manohar, S.G.; Husain, M. Attention as foraging for information and value. *Front. Hum. Neurosci.* **2013**, *7*, 711. [CrossRef]
17. Friston, K.J.; Lin, M.; Frith, C.D.; Pezzulo, G.; Hobson, J.A.; Ondobaka, S. Active inference, curiosity and insight. *Neural Comput.* **2017**, *29*, 2633–2683. [CrossRef]
18. Lindley, D.V. Theory and Practice of Bayesian Statistics. *J. R. Stat. Society. Ser. D (Stat.)* **1983**, *32*, 1–11. [CrossRef]
19. Wainwright, M.J.; Jordan, M.I. Graphical Models, Exponential Families, and Variational Inference. *Found. Trends Mach. Learn.* **2008**, *1*, 1–305. [CrossRef]
20. Loeliger, H.A.; Dauwels, J.; Hu, J.; Korl, S.; Ping, L.; Kschischang, F.R. The Factor Graph Approach to Model-Based Signal Processing. *Proc. IEEE* **2007**, *95*, 1295–1322. [CrossRef]

21. Dauwels, J. On variational message passing on factor graphs. In Proceedings of the 2007 IEEE International Symposium on Information Theory, Nice, France, 24–29 June 2007; pp. 2546–2550.
22. Wu, Y.; Mascaro, S.; Bhuiyan, M.; Fathima, P.; Mace, A.O.; Nicol, M.P.; Richmond, P.C.; Kirkham, L.-A.; Dymock, M.; Foley, D.A.; et al. Predicting the causative pathogen among children with pneumonia using a causal Bayesian network. *PLoS Comput. Biol.* **2023**, *19*, e1010967. [CrossRef] [PubMed]
23. Marcot, B.G.; Holthausen, R.S.; Raphael, M.G.; Rowland, M.M.; Wisdom, M.J. Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *For. Ecol. Manag.* **2001**, *153*, 29–42. [CrossRef]
24. Yedidia, J.S.; Freeman, W.T.; Weiss, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory* **2005**, *51*, 2282–2312. [CrossRef]
25. Forney, G.D. Codes on graphs: Normal realizations. *IEEE Trans. Inf. Theory* **2001**, *47*, 520–548. [CrossRef]
26. Parr, T.; Pezzulo, G.; Friston, K.J. *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*; MIT Press: Cambridge, MA, USA; London, UK, 2022.
27. Sajid, N.; Tigas, P.; Zakharov, A.; Fountas, Z.; Friston, K. Exploration and preference satisfaction trade-off in reward-free learning. *arXiv* **2021**, arXiv:2106.04316.
28. Marković, D.; Goschke, T.; Kiebel, S.J. Meta-control of the exploration-exploitation dilemma emerges from probabilistic inference over a hierarchy of time scales. *Cogn. Affect. Behav. Neurosci.* **2021**, *21*, 509–533. [CrossRef] [PubMed]
29. Pezzulo, G.; Cartoni, E.; Rigoli, F.; Pio-Lopez, L.; Friston, K. Active Inference, epistemic value, and vicarious trial and error. *Learn. Mem.* **2016**, *23*, 322–338. [CrossRef] [PubMed]
30. Friston, K.; Rigoli, F.; Ognibene, D.; Mathys, C.; Fitzgerald, T.; Pezzulo, G. Active inference and epistemic value. *Cogn. Neurosci.* **2015**, *6*, 187–214. [CrossRef]
31. Millidge, B.; Tschantz, A.; Buckley, C.L. Whence the Expected Free Energy? *Neural Comput.* **2021**, *33*, 447–482. [CrossRef]
32. Koudahl, M.T.; Kouw, W.M.; de Vries, B. On Epistemics in Expected Free Energy for Linear Gaussian State Space Models. *Entropy* **2021**, *23*, 1565. [CrossRef]
33. Shewry, M.C.; Wynn, H.P. Maximum entropy sampling. *J. Appl. Stat.* **1987**, *14*, 165–170. [CrossRef]
34. Demirdjian, D.; Taycher, L.; Shakhnarovich, G.; Grauman, K.; Darrell, T. Avoiding the "streetlight effect": Tracking by exploring likelihood modes. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Beijing, China, 17–21 October 2005; Volume 351, pp. 357–364.
35. Parr, T.; Friston, K.J. Uncertainty, epistemics and active inference. *J. R. Soc. Interface* **2017**, *14*, 20170376. [CrossRef]
36. Sun, C.; Yu, Y.; Li, V.O.K.; Lam, J.C.K. Multi-Type Sensor Placements in Gaussian Spatial Fields for Environmental Monitoring. *Sensors* **2019**, *19*, 189. [CrossRef]
37. Krause, A.; Singh, A.; Guestrin, C. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.* **2008**, *9*, 235–284.
38. Rabiner, L.; Juang, B. An introduction to hidden Markov models. *IEEE ASSP Mag.* **1986**, *3*, 4–16. [CrossRef]
39. Friston, K.; Stephan, K.; Li, B.; Daunizeau, J. Generalised filtering. *Math. Probl. Eng.* **2010**, *2010*, 621670. [CrossRef]
40. Berry, D.A. Bayesian clinical trials. *Nat. Rev. Drug Discov.* **2006**, *5*, 27–36. [CrossRef]
41. The PREVAIL II Writing Group. A Randomized, Controlled Trial of ZMapp for Ebola Virus Infection. *N. Engl. J. Med.* **2016**, *375*, 1448–1456. [CrossRef]
42. Proschan, M.A.; Dodd, L.E.; Price, D. Statistical considerations for a trial of Ebola virus disease therapeutics. *Clin. Trials* **2016**, *13*, 39–48. [CrossRef] [PubMed]
43. Broglio, K.; Meurer, W.J.; Durkalski, V.; Pauls, Q.; Connor, J.; Berry, D.; Lewis, R.J.; Johnston, K.C.; Barsan, W.G. Comparison of Bayesian vs Frequentist Adaptive Trial Design in the Stroke Hyperglycemia Insulin Network Effort Trial. *JAMA Netw. Open* **2022**, *5*, e2211616. [CrossRef] [PubMed]
44. Backonja, M.; Williams, L.; Miao, X.; Katz, N.; Chen, C. Safety and efficacy of neublastin in painful lumbosacral radiculopathy: A randomized, double-blinded, placebo-controlled phase 2 trial using Bayesian adaptive design (the SPRINT trial). *Pain* **2017**, *158*, 1802–1812. [CrossRef]
45. Berry, D.A. Adaptive clinical trials in oncology. *Nat. Rev. Clin. Oncol.* **2012**, *9*, 199–207. [CrossRef]
46. Warner, P.; Whitaker, L.H.R.; Parker, R.A.; Weir, C.J.; Douglas, A.; Hansen, C.H.; Madhra, M.; Hillier, S.G.; Saunders, P.T.K.; Iredale, J.P.; et al. Low dose dexamethasone as treatment for women with heavy menstrual bleeding: A response-adaptive randomised placebo-controlled dose-finding parallel group trial (DexFEM). *eBioMedicine* **2021**, *69*, 103434. [CrossRef] [PubMed]
47. Ryan, E.G.; Couturier, D.-L.; Heritier, S. Bayesian adaptive clinical trial designs for respiratory medicine. *Respirology* **2022**, *27*, 834–843. [CrossRef]
48. Hong, W.; McLachlan, S.-A.; Moore, M.; Mahar, R.K. Improving clinical trials using Bayesian adaptive designs: A breast cancer example. *BMC Med. Res. Methodol.* **2022**, *22*, 133. [CrossRef]
49. Connor, J.T.; Elm, J.J.; Broglio, K.R. Bayesian adaptive trials offer advantages in comparative effectiveness trials: An example in status epilepticus. *J. Clin. Epidemiol.* **2013**, *66*, S130–S137. [CrossRef] [PubMed]
50. Kaplan, E.L.; Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **1958**, *53*, 457–481. [CrossRef]
51. Cox, D.R. Regression Models and Life-Tables. *J. R. Stat. Society. Ser. B (Methodol.)* **1972**, *34*, 187–220. [CrossRef]
52. Zeidman, P.; Friston, K.; Parr, T. A primer on Variational Laplace (VL). *NeuroImage* **2023**, *279*, 120310. [CrossRef]

53. Friston, K.; Mattout, J.; Trujillo-Barreto, N.; Ashburner, J.; Penny, W. Variational free energy and the Laplace approximation. *NeuroImage* **2007**, *34*, 220–234. [CrossRef]
54. Sebastiani, P.; Wynn, H.P. Maximum Entropy Sampling and Optimal Bayesian Experimental Design. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2000**, *62*, 145–157. [CrossRef]
55. Ko, C.-W.; Lee, J.; Queyranne, M. An Exact Algorithm for Maximum Entropy Sampling. *Oper. Res.* **1995**, *43*, 684–691. [CrossRef]
56. Friston, K.; Da Costa, L.; Hafner, D.; Hesp, C.; Parr, T. Sophisticated Inference. *Neural Comput.* **2021**, *33*, 713–763. [CrossRef] [PubMed]
57. McLeod, C.; Wood, J.; Mulrennan, S.; Morey, S.; Schultz, A.; Messer, M.; Spaapen, K.; Wu, Y.; Mascaro, S.; Smyth, A.R.; et al. Preferred health outcome states following treatment for pulmonary exacerbations of cystic fibrosis. *J. Cyst. Fibros.* **2022**, *21*, 581–587. [CrossRef] [PubMed]
58. Charlie, M.; Richard, N.; Jamie, W.; Siobhain, M.; Sue, M.; André, S.; Mitch, M.; Kate, S.; Matthew, S.; Yue, W.; et al. Novel method to select meaningful outcomes for evaluation in clinical trials. *BMJ Open Respir. Res.* **2021**, *8*, e000877. [CrossRef]