The Role of Expert Opinion and Judgement in Statistical Inference and Evidence-Based Decision-Making

Naomi C. Brownstein*
Florida State University
and
Thomas A. Louis
Johns Hopkins Bloomberg School of Public Health
and
Anthony O'Hagan
The University of Sheffield
and
Jane Pendergast, Duke University

March 14, 2018

Abstract

This article resulted from our participation in the session on the "role of expert opinion and judgement in statistical inference" at the October, 2017 ASA Symposium on Statistical Inference. Here, we present a strong, unified statement on roles of expert judgement and opinion in statistics, along with processes for obtaining input. Topics include the role of subjectivity in the cycle of scientific inference and decisions, followed by a clinical trial and a greenhouse gas emissions case study that illustrate the roles of opinions and the importance of basing them on objective information and a comprehensive uncertainty assessment. We close with a call for increased proactivity and involvement of statisticians in study conceptualization, design, conduct, analysis and communication.

Keywords: Inferential Cycle, Elicitation, Bayesian Paradigm, Clinical Trials, Carbon Flux, Scientific Method, Statisticians, Subjectivity, Collaboration, Team Science

*The authors gratefully acknowledge support, in part, from the following: NCB: N/A; TAL: NIH-NIAID, U19-AI089680; PMA2020 from the Bill & Melinda Gates Foundation; AO'H: N/A; and JP: NIA grant P30AG028716 . Authorship order is alphabetical.

1 Introduction

As participants in the October, 2017 Symposium on Statistical Inference (SSI), organized and sponsored by the American Statistical Association (ASA), we were challenged to host a session and write a paper inspired by the question, "Do expert opinion and judgement have a role in statistical inference and evidence-based decision-making?" While we work from different perspectives and in different statistical paradigms (both Frequentist and Bayesian), there was a resounding "yes!" among us all, and there was much common ground in our thinking about the role of opinion and judgement. This article is a distillation of that common ground.

The 2017 Symposium will not produce a single agreed position document on statistical practice like the "The ASA Statement on P-Values" that resulted from the 2015 ASA Board meeting (Wasserstein & Lazar 2016). Rather, the ASA commissioned this special issue of *The American Statistician* to stimulate "a major rethinking of statistical inference, aiming to initiate a process that ultimately moves statistical-science – and science itself – into a new age." We intend that our article represents a firm statement answering the question and title of our session, "What Should Be The Role of Expert Opinion and Judgement in Statistical Inference and Evidence-Based Decision-Making?" Whilst we might still disagree on specific details or the relative importance of the various points, it is important to present a strong, unified statement related to this infrequently discussed and often under-appreciated component of statistical and scientific practice. Additional literature on the topic may be found in a companion article (Brownstein 2018).

Our article is organized in the following sections where we share our thoughts (and opinions!) on when and how expert opinion have a legitimate and necessary role in scientific inquiry. Section 2 presents a graphic showing the various stages of scientific inquiry and the scientific method, with particular reference to cycles of inference and decision-making. In Section 3, the stages are examined in more detail, focusing on their needs for expert opinion and judgement. Two short case studies are presented in Section 4, with emphasis on the principled and scientific application of expert judgements. Finally, Section 5 summarizes our key conclusions.

2 The cycles of inference and decision in science

Figure 1 illustrates four stages in the perpetual process of scientific inquiry and evidence-based decision-making: Question, Study, Interpret, and Inform. We first describe activities that comprise each of these stages. Then, we briefly discuss how the perpetual cycle of the scientific method fits into our framework. Additional details on the stages of scientific inquiry may be found in another article submitted to this special issue (Pendergast 2018).

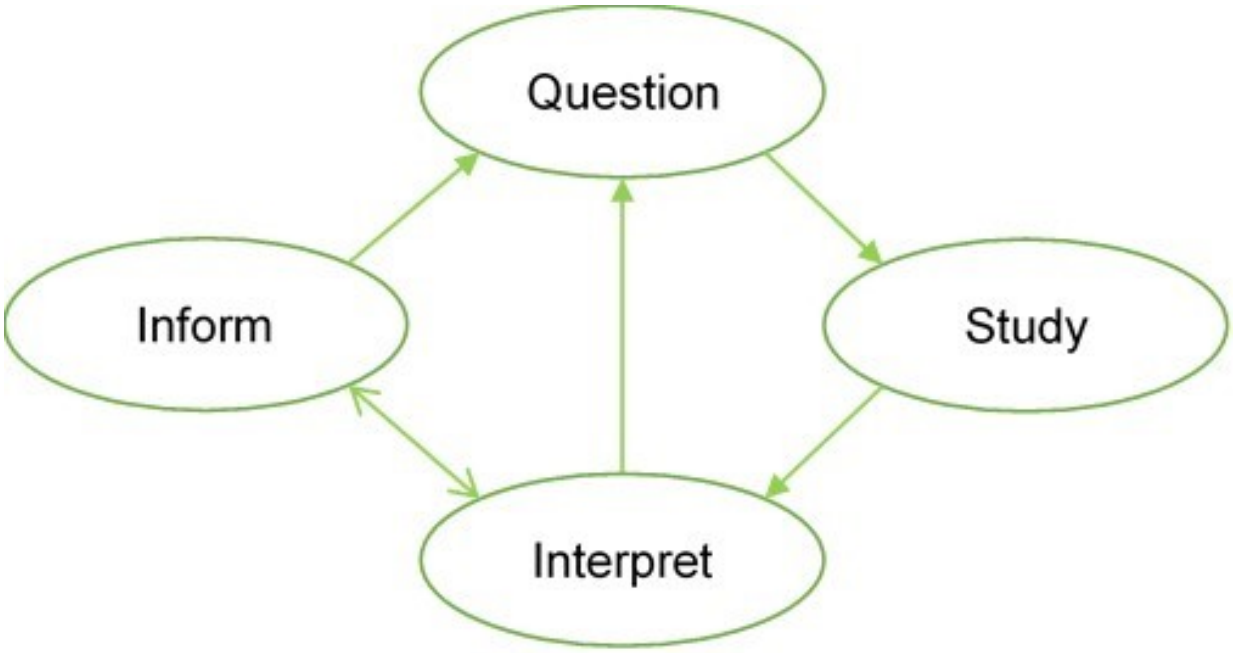

Figure 1: The cycles of inference and decision.

1. **Question**. Scientific inquiry can be characterized as beginning with one or more questions. A Question might be a formal scientific hypothesis arising either out of observation of real-world phenomena or from the current status of scientific knowledge and debate. On the other hand, a Question could be posed outside the scientific community, such as a request for evidence-based input to inform an impending policy decision. It might also simply express a wish to estimate more accurately certain quantities or parameters in current scientific theories.

2. **Study**. To address the Question scientifically, it is necessary to gather evidence. In the Study stage, we include all the activities of study design, including design

of observational studies, experiments, questionnaires, systematic literature reviews, and meta-analyses. The Study may also involve sequential design or the design of a number of distinct, possibly concurrent studies. We also include in this stage the conduct of the study, resulting in some form of data.

3. **Interpret**. In the Interpret stage, data resulting from the 'Study' are employed to address the Question. Typically, this may involve descriptive statistics and statistical inference, such as parameter estimation and hypothesis testing. In a Bayesian analysis, the primary result of the analysis might simply take the form of a posterior distribution. However, the 'Interpret' stage should also embed the findings of the analysis in the wider body of science to which it refers, thereby updating that body of knowledge. In doing so, the wider implications of those findings will emerge.

4. **Inform**. The Interpretation stage will often suggest new Questions, and a new cycle of scientific investigation will thereby be initiated. First, however, Interpretation usually will be followed by the Inform stage. For a formal scientific study, findings should be formally written and communicated in peer-reviewed outlets, such as conferences, journals, and books. In fact, peer-review may lead to revised Interpretation before formal publication of the findings. Subsequent examination and evaluation of published studies by the scientific community may in turn lead to new Interpretations of existing Studies and new Questions, leading to new Studies. Where the Question is a request for input to a decision, the Inform stage is when the results of the Study are communicated to facilitate the decision-making process. New Questions may arise based on the output produced in the Inform stage. Alternatively, the original Question may need to be revisited in one or more future Studies, especially when the evidence is not yet adequate to merit a robust conclusion.

The stages in Figure 1 are formulated in highly general terms, because the practice of scientific inquiry, and hence the remit of statistical analysis, is also very wide.

2.1 The Scientific Method

The practice of science has often been described in terms of the Scientific Method, which is defined by the English Oxford Living Dictionary (Oxford University Press 2018) as involving "systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses." A thorough review of the history of the scientific method is found in Anderson (2016). More generally, one might describe the scientific method as a collection of methodologies and processes of gathering observable, measurable evidence using experimentation and careful observation. The purpose may be to inform about pathways or mechanisms by which results are obtained or to aid in prediction or estimation of quantities of interest.

In the world of research and discovery, scholars utilize the scientific method as a prototypical guide, modifying or adding new elements to the process as needed in the particular area of exploration. Thus, we base our four-stage graphic on the depiction of Garland Jr. (2016), adding the Inform stage to accommodate decision-making and broadening the definitions of the other stages. While we believe that not all scientific inquiry, and certainly not all decision-making, falls naturally within that description, it is important to understand that the Scientific Method fits into our framework. Indeed, the Scientific Method can be thought of as the backbone of scientific rigor that in particular justifies our view of the progress of science in terms of cycles. In current terminology, we might add the term 'data-driven' when describing the scientific method, implying the use of data to make scientific conclusions.

3 Science and subjectivity

We now examine the many ways in which expert opinion and judgement enter into the four stages of Figure 1 and the roles that they play in each. We give particular attention to the role of statistical expertise, as opposed to *content* expertise, which refers to expertise in the discipline in which the Question arises.

3.1 The Question stage

When developing the Question, we rely heavily on the judgement and expert opinion of the content experts. The Question may arise from identification of a barrier or problem in need of an answer, a quest to understand the 'why' behind some phenomenon, event, or process, or simply to better quantify some parameter, effect or disturbance. For example, one might ask "Why are some people able to fend off the negative impacts of an HIV infection while others cannot?" Knowledge of the literature and what other experiments or studies others have done to address this Question or related Questions is critical. Where the Question arises from a request for scientific input to inform a decision, the content experts serve key roles in formulating specific questions, such as "What can we say about the toxicity of this pollutant for fish in European rivers?", or "Which areas in this catchment will be flooded if the catchment experiences a once-in-100-years weekly rainfall?"

While the content experts have primary responsibility to develop the Question, statisticians can elicit clarity on the framing of the Question by asking pertinent questions from their perspective. Inquiry from the statistician serves not only to establish and confirm the statistician's understanding of the Question and its scientific context, but also to translate the question to a statistical framework, which may guide analytic decisions in the Study Stage. Indeed, strong listening and communication skills are critical for both the content and statistical experts!

In addition, a Study may involve more than one Question. In this case, discussions are needed regarding which Questions are considered primary or secondary, the interrelatedness or independence of the Questions, and whether any of the Questions can be addressed jointly.

If the Question is seeking information on potential pathways or mechanisms, evidence for or against competing theories is presented, and decisions must be made on the rationale for how the Question will be pursued. The content experts play the main role here, but statistical expertise can be helpful when framing the Question to bring up potential statistical issues with the proposed approach.

Part of defining the Question is determining what evidence, measures and parameters would be useful and adequate to arrive at an answer. Properties of those measurements,

including validity, reliability, cost, and distributional properties are considered. This is an area in which both content and statistical experts can contribute. For example, content experts will be focused on what data would be needed and whether primary data collection would be needed or existing secondary data would be sufficient. If the study needs to collect primary data, there will be a need for discussion of exactly what information will be desired and how to elicit that information. In turn, the statistician will seek to better understand properties of the desired measurements. Moreover, the statistician may focus more on unmeasured influences, the impact of missing data, whether sources of bias or variability could be reduced or eliminated by appropriate study design or data collection methods. Thinking through the issues that can (and will!) arise when defining the Question and information needed to address it requires effective collaboration and team effort.

3.2 The Study stage

Once the Question and pertinent measures are defined, the approach to gathering information (data) must be developed. Can the Question be answered in one study, or will it take a series of planned studies? What studies have been done before? Could any component of prior studies be replicated with modification or improvement in this study? Did previous studies report unanticipated problems that could be avoided in this study?

Much of the work in planning a study involves a collaborative effort among all members of the research team. Those researchers who will be "on the ground" collecting information will have expertise on what is feasible and what is not. The statistician can offer their expertise and opinion on many aspects of study design and implementation, such as strengths and weaknesses of different study designs, questionnaire development, psychometric properties of data collection instruments, issues surrounding sources and control of error, replication, operational randomization, and blinding. The content experts will share their knowledge and opinions on the target scope of inference, such as whether to estimate the 100-year flood plane for a large geographic area or for just for one river; for all people with a disease or only those in a local area who meet defined inclusion criteria. Content experts bring to the discussion additional information that can be considered in the study design, perhaps stemming from theorized or understood pathways, mechanisms,

or concurrent influences by which observed outcomes can differ.

If working within the Bayesian framework, statisticians help elicit information from the content experts to feed into the prior distribution. Those approaching the problem from a Frequentist perspective will also be looking to prior studies and expert opinion when developing a study design. No matter what analysis approach is used, the ultimate goal remains the same: to collect enough high quality information to effectively address the Question. Here, the word quality can encompass reduced variability of measures, removal or control of sources of biases, and proper data collection and maintenance systems.

At the end of the Study stage, there should be enough information to create a study protocol, data monitoring plan, and a Statistical Analysis Plan (SAP), upon which everyone agrees (Finfer & Bellomo 2009, Ellenberg et al. 2003, Ott 1991). While not every study will require these formal documents, the idea is that there is a common understanding on how the study will be conducted, the data collected, monitored, and maintained, and the analytic approach used to address the Question. With the goal of transparency of the body of work to be accomplished, good documentation and data provenance are important components of scientific inquiry. A well thought-out study design and SAP can help safeguard against urges to reanalyze the data later after obtaining disappointing results. Unfortunately, forms of scientific malpractice, such as repeating the analysis with a slightly altered question, often called p-hacking, are commonplace in the scientific literature (Head et al. 2015). However, it should be noted that the data monitoring process itself involves judgment, as exemplified in Section 4.1.2 and discussed further elsewhere (Pocock 2006).

Often, before a study can begin, funding must be obtained. While there are commissioned studies, often there is a need to write a formal proposal to seek funding. A grant proposal provides an opportunity to present not only the justification for the study but also detail on the study process, analysis plan, and how the desired information will address the Question. The grant review process brings in a new external set of experts with judgements of their own on the merits of the proposal. Feedback for a proposal can identify strengths and weaknesses of the proposed work. Reviewers' expert opinions on whether a particular study should be funded are intended to weed out studies without strong support and justification for both the importance of the Question and the development of the Study.

3.3 The Interpret stage

The methodology used to analyze the data will feed into the interpretation of the data. Some analytic methods will produce estimated probabilities that relate directly to the Question; others may provide information that helps address the Question, but perhaps more indirectly. The chosen methodology may produce parameter estimates that need to be understood in context of the model. The expertise of the statistician is needed both to understand the nuances of proper interpretation of the analytic results in context of the executed study and modeling used and to guard against over-interpretation. For example, a statistician may help protect the team from common misconceptions and malpractice, such as the tendency to extend inferences to populations outside those under study, or to interpret association as causation. As the analytic results are interpreted in the framework of the Question and the Study protocol, the content expert blends the (properly interpreted) new findings into their existing knowledge and understanding.

3.4 The Inform stage

Once the analytic results have been interpreted within the framework of the study design and measures used to address the Question, it is time to assess what was learned, and share that information more broadly. Of course, those who developed the Question and often those who funded the study will need and expect a complete summary of the work done, describing how the results have informed the Question. Indeed, there may be interest in the work outside of academia, such as patients curious about new therapies for their conditions, policy-makers seeking to understand actions that may yield societal benefit, and others simply wondering about current topics and trends in various fields of science.

Scientific publications, where the process, methods, results, and conclusions of a study can be shared broadly, are important tangible outcomes of the Inform Stage. In fact, the process of creating a scientific manuscript and undergoing the peer-review process for publication is another place in which expert opinion and judgement enter into both the Interpret and Inform stages. Comments from others on drafts of the manuscript can lead to revised interpretation in light of new information or perspective from that feedback before submission of the manuscript for publication. Once submitted, additional expertise

and opinion is gathered from the peer reviewers, which again can impact the information presented in the final publication.

Strong communication skills in the Inform stage are paramount to ensure that all components of a study are presented clearly. In each Inform stage, we build on what we know and we learn from the findings, whether or not the results obtained were anticipated or exciting. To guard against publication bias, null results, in particular, must be communicated, despite disincentives for doing so (Franco et al. 2014, Easterbrook et al. 1991).

When the Question has arisen from a decision-making context, the primary purpose of the Inform stage is to convey the scientific evidence to the decision-maker, after it has been assembled and analyzed in the Study and Interpret stages. Here, too, communication skills are particularly important. Governments are increasingly but not exclusively reliant on evidence for policy- and decision-making (Oliver & de Vocht 2017, HM Treasury 2015, Oliver et al. 2014, LaVange 2014), and there is much current interest in the challenges of communicating uncertainty to decision-makers (Cairney & Oliver 2017, National Academies of Sciences, Engineering, and Medicine, and others 2017, National Research Council and others 2012*a*,*b*).

As depicted in our graphic of the Scientific Method (Figure 1), learning and discovery is cyclical. Once we address one Question, new Questions often arise. Researchers working in similar fields as the authors of one study may be interested in their work, perhaps using information from published studies to inform their next study. Sometimes the results obtained are not definitive, or not adequate for robust decision-making, and ways to redirect the next investigation of the same or a revised version of the Question are planned. Other times, a replication Study is conducted based on the same Question simply to see if the results remain qualitatively similar, despite inevitable lack of perfect duplicability in all aspects of the study environment (Lindsay & Ehrenberg 1993).

3.5 But is it science?

Expert opinions and judgements are subjective. When expressed as probabilities, those probabilities must also necessarily be interpreted according to the theory of subjective probability (Anscombe & Aumann 1963). Any expert opinions or judgements that are

used in any stage of a scientific inquiry imply that the inquiry itself, and its outcomes, also contain a subjective component. The inescapable conclusion is that science itself has an element of subjectivity.

We frequently encounter heated opposition to such a notion. Objectivity, we were taught, and some current students are still being taught, is fundamental to the scientific method; subjectivity is anathema to a true scientist. If, for instance, we seek to elicit prior distributions for parameters from a content expert, the request is not infrequently declined on the grounds that the expert is a scientist and therefore his or her subjective opinions are of no relevance. To such people, subjectivity is unscientific, practically synonymous with bias, prejudice, sloppy thinking, wishful thinking, or even superstition.

Science cannot be totally objective. On a daily basis, scientists propose new or amended theories, choose experimental designs or statistical analysis techniques, interpret data, and so on; the list, as we have highlighted, is almost endless. Indeed, making better subjective judgements is one of the key features that distinguish a top scientist from a lesser one. Statistics similarly involves subjective judgements, as others have recently argued (Gelman & Hennig 2017).

To reconcile these views, we must accept that, in practice, there is subjectivity in every stage of a scientific inquiry, but objectivity is nevertheless the fundamental goal. At every stage, we should seek to be as objective as we can be, to base opinions and judgements on evidence and careful reasoning, and wherever possible to eradicate bias, prejudice, sloppy thinking, etc.

We are firmly of the (subjective!) opinion that to recognize these facts openly can only be beneficial to the progress of science. Scientific activity encompasses many activities beyond the formal scientific method, many of which involve subjective opinions and judgements. Yet, the rigor of the scientific method is an ideal towards which all such activity should strive. In other words, while scientific activity includes numerous components that involve subjective opinions and judgements, objectivity is an ideal towards which all such judgements should strive.

4 Case studies

We present two case studies. While one is observational and one is experimental, both exhibit complexities that require careful input from content experts and statisticians for proper analysis. Examples focus on the roles that subjective judgements played in those studies and the steps that were taken to ensure that judgements were as objective, as 'scientific', as possible.

4.1 Example: Expert Judgement and Opinion in a Randomized Clinical Trial

The first case study was chosen for the subtlety and judgement involved in planning the statistical modeling approach and in the monitoring and interpretation of the results. Section 4.1.1 provides background on the planned statistical approach. Section 4.1.2 describes the study goals and design. Sections 4.1.3 and 4.1.4 provide the modeling details. Section 4.1.5 illuminates the situation that required extensive judgement. Finally Section 4.1.6 summarizes the results of the trial and the role of judgement throughout the process.

4.1.1 Background: Bayesian clinical trials

Bayesian approaches to clinical trial design, conduct and analysis have been shown to offer substantial improvements over traditional approaches in a variety of contexts. See Abrams et al. (2004) and Berry et al. (2010), for a range of examples. The Bayesian paradigm provides appropriate language for discussing interim and final results, for example by supporting statements such as, 'in the light of accruing data, the probability that treatment A is better than treatment B by 5 percentage points is 0.XY. (See Curtis et al. (2013), for an example.) When accruing data are consistent with the prior distribution, a trial can be stopped earlier than with traditional monitoring. However, it is also important to address situations wherein the prior and the data diverge, as is the case in the following example.

4.1.2 Protocol for the 'TOXO' Trial

The Unites States Community Programs for Clinical Research on AIDS conducted a randomized trial of prevention of Toxoplasmosis (TOXO) comparing pharmacological prevention (either Clindamycin or Pyrimethamine at a sub-therapeutic dose) with placebo, both with careful monitoring for symptoms and signs. The premise of such prevention studies is that a low dose of a pharmaceutical that is typically used to treat overt disease may also prevent or delay onset. However, even a low dose of a pharmaceutical may induce adverse effects, such as toxicities or resistance to treatment, and consequently, watchful waiting (the placebo 'intervention') may be better than the potentially preventative treatment. Eligibility for the trial included either an AIDS defining illness or $CD_4 < 200$ or a positive titer for toxoplasma gondii. While the trial was originally designed with four treatment groups (Clindamycin: Active vs placebo) and (Pyrimethamine: Active vs placebo) each with $2:1$ randomization to the Active arm, the Clindamycin arm was stopped after a few months. Hence, the example we present is based on data generated by the Pyrimethamine vs Placebo comparison, the primary Question on which we focus. Full details on the 'TOXO' trial may be found in Jacobson et al. (1994).

The Data and Safety Monitoring Board (DSMB) monitored the trial at pre-specified, calendar-determined dates using the O'Brien & Fleming (1979) boundaries, and the full database was available for our after-the-fact example of how monitoring might have proceeded using a Bayesian approach. Our analysis evaluated the Toxoplasmosis or death endpoint using the Cox (1972) proportional hazards model with adjustment for baseline CD_4 count. We 'stopped' the trial when the posterior probability of benefit or the posterior probability of harm became sufficiently high.

4.1.3 Model for the 'TOXO' Trial

We modeled the event 'toxoplasmosis or death' with covariates treatment group status ($z_{1j} = 1$, if participant j received Pyrimethamine; $z_{1j} = 0$, if placebo), and CD_4 cell count at study entry (z_{2j}). The log hazard-ratio is,

$$\log(\text{hazard ratio}) = \beta_1 z_{1j} + \beta_2 z_{2j}$$

with $\beta_1 < 0$ indicating a benefit for Pyrimethamine. We put a flat prior on the CD_4 effect (β_2), and a variety of priors for the Pyrimethamine effect (β_1). The choice of the Cox (1972) model and the use of a non-informative prior for β_2 were judgements of the statisticians. Although the Cox model is essentially considered the default choice when modeling the time to an event, it has important assumptions, namely that the hazard functions for both interventions are proportional and censoring is non-informative. As such, the model should only be adopted after careful consideration by experts, such as in this example. The choice of a noninformative prior distribution for β_2 reflects the statisticians' judgement that there was essentially no information available before the trial began on the association of CD_4 with TOXO incidence.

4.1.4 Elicitation in the 'TOXO' Trial

As described in Carlin et al. (1993)and Chaloner et al. (1993), we elicited prior distributions for β_1 from five content experts–three HIV/AIDS clinicians, one person with AIDS conducting clinical research, and one AIDS epidemiologist. We included two other priors in the monitoring–an equally-weighted mixture of the five elicited priors and a non-informative flat prior.

Our elicitation targeted potentially observable, clinically meaningful features and then mapped responses to the Cox model. (We did not elicit hazard ratios directly!). Specifically, we asked each expert elicitee to report their best estimate of the probability of TOXO or death in two years under placebo, P_0; then had each expert draw a picture of the distribution of the two-year probability under Pyrimethamine, conditional on the estimate under placebo, $[P_{pyri} \mid P_0]$. Then, for each elicitee, we converted these conditional distributions to a a prior distribution for the log(hazard ratio) using the transformation, $\beta_1 = \log(1 - P_0) - \log(1 - P_{pyri})$.

Figure 2 displays the elicitation results. At trial initiation, there was little known about the baseline incidence of TOXO, and so the content experts based their distributions on general expertise and analogy with other endpoints and contexts. The range of the five reported two-year incidence probabilities under placebo was wide, but elicitees believed that toxoplasmosis had a non-negligible baseline incidence. All elicitees were quite optimistic

regarding Pyrimethamine's benefit, with experts C and E the most optimistic, placing all of their probability distribution for incidence under Pyrimethamine way below their estimate of incidence under placebo.

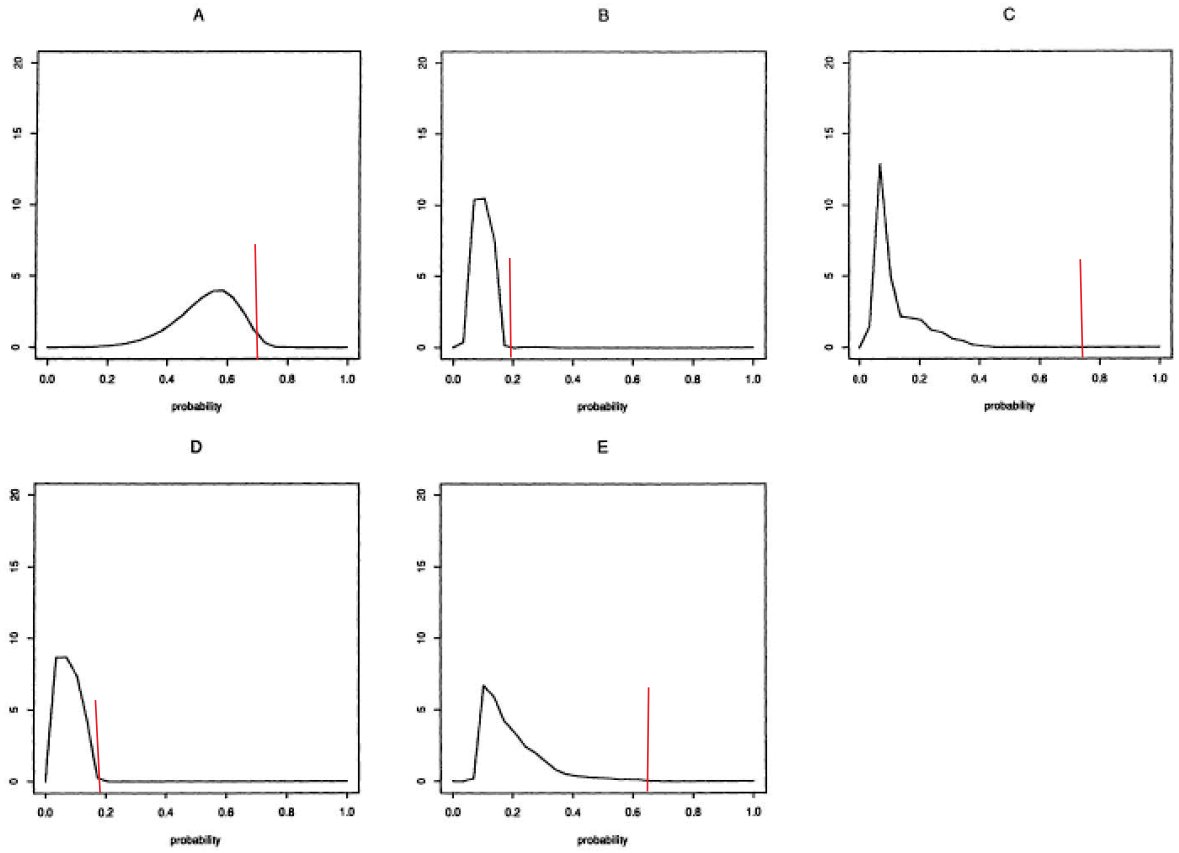

Figure 2: Elicited priors for the five elicitees. The red vertical line is at P_0 is the 'best guess' two-year incidence of TOXO or death under placebo. The smoothed densities are for two-year TOXO or death incidence under Pyrimethamine, conditional on the placebo rate.

4.1.5 Monitoring Results for the 'TOXO' Trial

The DSMB monitored the trial at calendar intervals. At its meeting on December 31, 1991, the DSMB recommended stopping the trial for futility, because the Pyrimethamine group had not shown significantly fewer events, and the low overall rate made a statistically

significant difference in efficacy unlikely to emerge. Additionally, there was an increase in the number of deaths in the Pyrimethamine group relative to the placebo (i.e. stopping due to harm).

Figure 4.1.5 displays posterior probabilities of benefit (hazard ratio $\leq 0.75, \beta_1 \leq \log(0.75)$) and harm (hazard ratio $> 1.0; \beta_1 > 0$) for an equally weighted mixture of the five prior distributions (E–exact, and N–normal approximation) and for monitoring based on the partial likelihood (L) which is equivalent to using a flat prior for β_1 (essentially 'Frequentist Bayes').

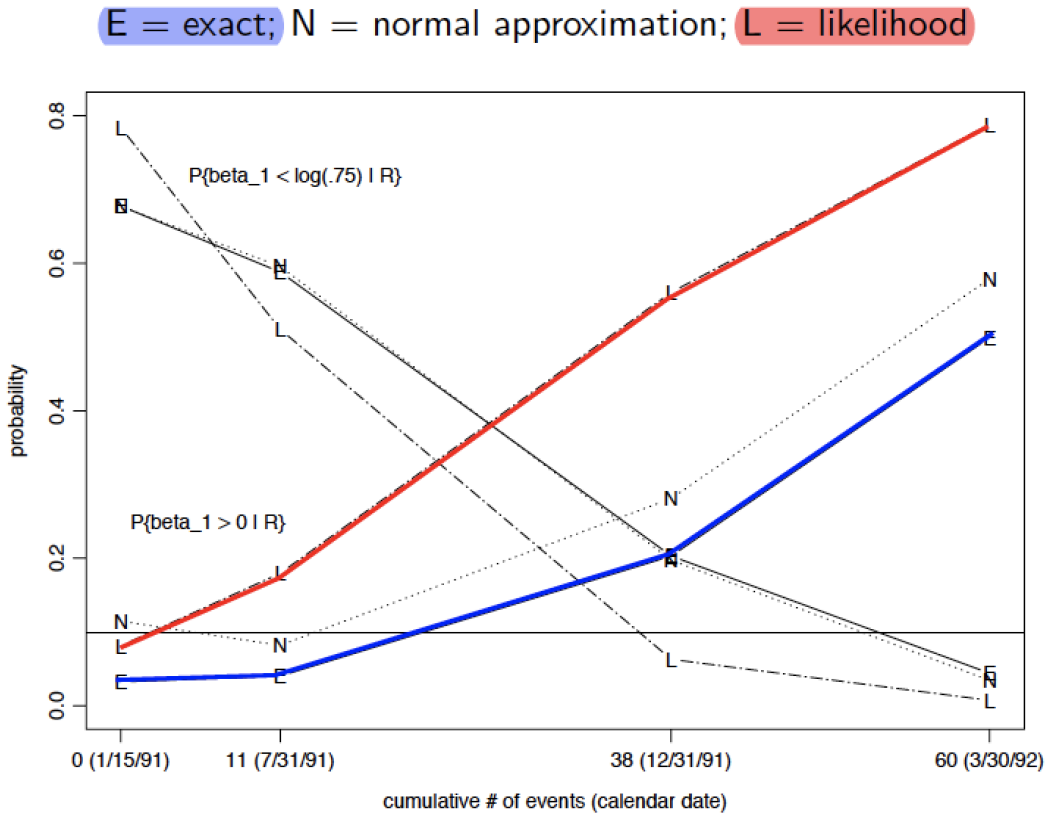

Figure 3: Posterior probability of benefit; hazard ratio $\leq 0.75, \beta_1 \leq \log(0.75)$; and harm; hazard ratio $> 1.0; \beta_1 > 0$; for the mixture prior (E–exact, and N–normal approximation) and for monitoring based on the partial likelihood (L). The blue line (—) plots the probability of harm for the mixture prior; the red line (—) plots probability of harm for the flat prior (partial likelihood). The abscissa is the monitoring date and number of events.

As displayed in Figure 2, elicitees believed that toxoplasmosis had a non-negligible

incidence and that Pyrimethamine would have a substantial prophylactic effect. Thus, each prior and their mixture are to varying degrees far from the accruing information in the trial. Consequently, the partial likelihood-based monitoring, which can be considered 'flat prior Bayes,' gives an earlier warning of harm compared to monitoring based on the mixture prior. It required considerably more information to overcome the *a priori* optimism of the elicitees. The information presented in Figure 2 also indicates that the probability of harm computed separately for each of the five elicited priors would have lagged behind that based on the partial likelihood, with use of prior A, B or D giving an earlier warning than use of priors C or E.

4.1.6 Discussion of the 'TOXO' Trial

Our after-the-fact analysis highlights ethical issues generated by real-time use of expert knowledge. Our example shows that if elicited priors were to be allowed in the actual clinical trial monitoring (of course, Institutional Review Board approval would be needed), trial stopping and other decisions could differ from those produced by traditional monitoring. Indeed, if that weren't the case, there would be no reason to do the hard work of elicitation!

In our example, stopping was delayed, highlighting the question of whether it would be ethical to continue beyond a traditional stopping point due to a prior belief that Pyrimethamine would have a strong, prophylactic effect. Of course, in other situations stopping could be earlier than that based on traditional methods, also raising an ethical issue, though possibly not as acute.

To summarize, in clinical trial monitoring, prior information can have two main effects.

a. If prior information conflicts with the available data at the time of monitoring, it may suggest continuing the trial, at least until the next review. This situation can arise when the experts are more optimistic than the emerging data, as in the TOXO trial, or when they are more pessimistic. In either case, continuing the trial long enough to obtain sufficient evidence to convince the content experts may have an added benefit that the results would translate to practice relatively rapidly.

b. If prior information supports the available data at the time of monitoring, it may suggest terminating early, on the grounds that sufficient evidence now exists to make

a decision.

An increasing number of trials are utilizing prior information (Abrams et al. (2004) and Berry et al. (2010)), and the advent of such trials emphasizes the importance of our condition that prior judgements must be made 'as carefully and as objectively as possible.' Expert opinion should be sought from enough content experts as is needed to encompass the range of opinion in the community, and prior distributions should be elicited carefully and rigorously, as described, for instance, in O'Hagan (2018).

4.2 Example: Expert Judgement and Opinion in Environmental Modeling

The second case study describes the estimation of a parameter in a complex environmental modeling problem. Section 4.2.1 describes the background of the problem and the nature of the complex model used to answer the Question. Section 4.2.2 describes quantification of the model inputs, and Section 4.2.3 describes the choices made in the computational techniques. Finally, section 4.2.4 summarizes and Interprets the results of the Study.

4.2.1 Background: UK carbon flux

As a party to the Kyoto protocol, the United Kingdom is committed to specific target reductions in net emissions of greenhouse gases. Monitoring progress towards these targets is challenging, involving accounting for numerous sources of emissions. One potential mitigating factor is the ability of vegetation to remove carbon dioxide from the atmosphere, thereby acting as a 'carbon sink.' However, accounting for this is also extremely complex. During the day, through the process of photosynthesis, vegetation absorbs carbon dioxide and releases oxygen, using the carbon to build plant material. Photosynthesis requires sunlight, chlorophyll in green leaves and water and minerals gathered by the plant's roots, so the efficiency of carbon removal depends on factors such as weather, leaf surface area and soil conditions. Conversely, carbon is released at night, carbon extracted from the atmosphere and converted to biomass will eventually be released as the plant ages and dies, and carbon in leaf litter is released by microbial action in the soil. The Sheffield Dynamic

Global Vegetation Model (SDGVM) was built with mathematical descriptions of all these processes to predict net carbon sequestration due to vegetation (Woodward & Lomas 2004). For a given site, the model takes many inputs describing the type of vegetation cover and soil at the site, together with weather data, to estimate Net Biosphere Production (NBP), i.e. the net decrease in atmospheric CO_2, from that site over a given time period.

This case study used SDGVM to estimate the total NBP for England and Wales in the year 2000. It is important to recognize that there is inevitably uncertainty about all the model inputs. Uncertainty about inputs induces uncertainty about model outputs, and the study sought to quantify the output uncertainty in the form of a probability distribution for the total NBP. Details are reported in Kennedy et al. (2008)and Harris et al. (2010).

Before considering the statistical aspects of this case study, we first note that considerable content expertise had already gone into the development of SDGVM. Based on the available scientific knowledge, expert judgements were made in choosing the structure of the model and the equations that describe each of its biological processes. Care went into these choices to ensure that they were reasonable and scientifically defensible. However, in such modeling, it is usually unrealistic to include every process to the greatest level of detail and complexity that is believed to be applicable. First, the more complex and detailed the model, the longer it will take to compute; for practical reasons, it may be necessary to compromise on complexity. Second, more complex models may be less robust and reliable in their predictions, because at the highest level of detail, there is often less scientific consensus about the equations and the parameters within them. In addition, models with a large number of parameters may suffer from overfitting (Hawkins 2004). Judgements about the optimal degree of complexity to achieve a computable, accurate and reliable representation of the phenomenon being modeled often demand a particularly high level of expertise.

4.2.2 Input distributions for the SDGVM

The model was run over 707 grid cells covering England and Wales. For each grid cell, given the land cover in that cell, and given input parameters describing the soil composition and properties of the vegetation types, the model was first 'spun-up' for 600 years using historic

climate data to stabilize the vegetation with respect to ages and heights of trees, leaf density, etc. The model was then run forward for one year using weather data from 2000, and the NBP for the year was computed for each grid cell. The NBP for England and Wales in 2000 is the sum of the NBP values across all the grid cells.

Care was taken to quantify the uncertainty in the various inputs as described previously (O'Hagan 2012). Briefly, details are described below.

- *Soil composition.* A publicly available soil map for England and Wales (Bradley et al. 2005) was used to provide estimates of the soil composition in each grid cell. Because the map was created at a higher resolution than the grid cells used for SDGVM, figures were averaged over each grid cell to provide an estimate. The variance of the same figures over a grid cell, divided by the number of map points in the cell, was used to quantify uncertainty around the estimates for a grid cell. However, the variance was increased by a factor to represent (a) additional uncertainty in the map data and (b) spatial correlation within the cell. The decision to use an increased estimate of uncertainty in the model was based on an expert judgement on the part of the statisticians, in consultation with content experts.

- *Land cover.* A map of land cover was also publicly available (Haines-Young et al. 2000), derived from satellite observation. The original analysis reported in Kennedy et al. (2008) did not quantify uncertainty in land cover. However, unlike the soil map, the content experts felt that the uncertainties in land cover were large and there could be biases in the process by which land cover is inferred from the satellite readings. In a subsequent analysis (Cripps et al. 2013, Harris et al. 2010), a statistical model was built to quantify uncertainty in land cover maps derived from remote sensing. The analysis of England and Wales NBP in 2000 was then extended to account for the additional uncertainty. The method makes use of the 'confusion matrix', which for the Haines-Young et al. (2000) map was given by Fuller et al. (2002). The confusion matrix is a contingency table based on a large survey of actual, ground-truth, land cover, and shows, for each ground-truth vegetation type, the frequency with which it was classified by the Haines-Young et al. (2000) map in each vegetation type. The statistical analysis required probabilistic inversion of the confusion matrix in order to

derive, conditional on the satellite land cover, the probabilities of the various ground-truth cover. The careful expert judgements of statisticians and content experts are delineated in Cripps et al. (2013).

- *Vegetation properties.* SDGVM classifies land cover into plant functional types (PFT). For England and Wales, we used four PFTs — evergreen needleleaf trees, deciduous broadleaf trees, crops and grassland. Each PFT is associated with a set of quantities, including maximum age, stem growth rate and leaf lifespan. However, most of these inputs were missing. Some properties had been estimated experimentally, but only for very few individual species within a given PFT. We therefore used expert elicitation to construct a probability distribution for each parameter. Elicitation is an area where it is particularly important to take care to avoid biases that commonly arise in subjective judgements of probabilities (Kynn 2008). Another article arising from the Symposium addresses issues related to elicitation in detail (O'Hagan 2018).

- *Weather.* Weather data, such as temperature, precipitation and cloud cover, were available for each grid cell for every day in 2000. There are no doubt errors in these data, due not only to errors in the underlying daily measurements, but also to the fact that those measurements have been interpolated to produce the data at the level of grid cells. Nevertheless, it was felt that uncertainty in these inputs was relatively small and could not be quantified without adding extra assumptions. The decision not to trade (potentially unreasonable) assumptions for (potentially increased) precision for weather data was a judgement made jointly by statisticians and content experts.

4.2.3 Propagating the input uncertainty in the SDGVM

In order to propagate input uncertainty through a mechanistic model such as SDGVM, the simplest and most direct approach is by Monte Carlo. A large sample of random draws are made from the probability distributions of the inputs, and the model is run for each sampled input set. The resulting set of outputs is then a random sample from the output uncertainty distribution. However, like many models that are built to describe complex physical processes, the computational load in running the SDGVM model is substantial, and it would not have been feasible to propagate parameter uncertainty through the model in

this way at even one of the grid cells. The problem of quantifying uncertainty in complex computer models has acquired the name Uncertainty Quantification (UQ), and various tools are available to enable uncertainty propagation. The analysis used Gaussian process emulation (O'Hagan 2006, Oakley 2016), which is probably the most popular UQ technique. Even so, it would not have been feasible to build emulators at every one of the 707 grid cells; 33 were chosen by content experts to represent the range of conditions across England and Wales, and GP techniques were adapted to infer the magnitudes of output uncertainty at the other 674 cells. The computational techniques and accompanying statistical theory to manage the analysis are set out in Kennedy et al. (2008) and Gosling & O'Hagan (2006). It is worth noting that here, too, expert opinion and judgement played a large role.

4.2.4 Results of the UK Carbon Flux Study

The content experts were certainly interested in knowing how much uncertainty in the total NBP would be induced by uncertainty in the inputs. Their instinctive approach to estimating the total NBP was to run the model just once with all of the inputs set to their expected values; we call their idea the plug-in estimate. Yet, the NBP output from SDGVM is a highly nonlinear function of its inputs. Therefore, the expected value of NBP, when we allow for uncertainty in the inputs, is not equal to the plug-in estimate. Statisticians incorporated the input uncertainty and computed the expected NBP for England and Wales in 2000 as 7.46 Mt C (megatonnes of carbon). By contrast, the plug-in estimate was 9.06 Mt C. Not only is this a substantial difference, but the standard deviation was estimated as 0.55 Mt C. Therefore, the total NBP was probably in the interval (6.36 Mt, 8.56 Mt C) and very likely to be less than the plug-in estimate.

The result was very surprising for the content experts. It seems that the explanation lies in their estimates of vegetation properties. In effect, the experts had estimated values that were more or less optimal for the plants to grow and absorb CO_2. Any deviation from these values led to lower NBP. For the total NBP to be even close to the plug-in estimate required all the parameter values to be close to their estimates, a joint event that had very low probability.

In this case study, the combination of expert opinion and judgements from both content

experts and statisticians, applied with as much care, rigour and objectivity as possible, led to a scientific result that certainly prompted new questions. For example, see Cripps et al. (2013). The cycle of scientific investigation was thereby renewed.

5 Conclusions

The use of expert judgement is essential in and permeates all phases of scientific and policy studies. Expert knowledge is information; to ignore it or fail to obtain it incurs considerable opportunity costs. Judgements should be as objective as possible, based on data when available, but deciding what data are relevant will always involve degrees of judgement and opinion. Documentation is key, so that stakeholders know how and in which study components (design, conduct, analysis, reporting, and decision-making) the principal judgements had impact.

We are all Bayesians in the design phase, because information is not yet available from the study being planned, and so design decisions must be based on (prior) external data and judgement. Designers employ varying degrees of formalism in developing the study design and statistical models, but until the data are in, all decisions are *pre-posterior* or *pre-analysis*. A formal Bayesian approach can be used to either develop a Frequentist design (Bayes for Frequentist) by, for example, finding a sample size and other design components that ensure

$$P(\text{power} > \text{goal} \mid \text{design \& assumptions}) > \gamma \ ,$$

(see, Shih 1995, for an implementation), or to ensure that Bayesian properties are in an acceptable region (Bayes for Bayes). All researchers, irrespective of their philosophy or practice, use expert judgement and opinion in developing the data model and interpreting results. We recommend additional use of Bayesian approaches, even if only to provide a vehicle for documenting the roles of judgements and as a platform for sensitivity analyses.

Elicitation of opinions can be conducted informally, but far better is to embed it in a formalism such as the approach in Section 4.2 and more fully discussed by O'Hagan (2018). Elicitation design requires determining which experts should be queried, how many and how their opinions should be combined or used separately.

It should be unsurprising that statisticians have essential roles as scientists, ideally serving as leaders or co-leaders in all study aspects. Indeed, virtually all aspects of a study have statistical content, though almost no aspects are solely statistical. Consequently, our expert advice is for pro-active 'design' in constituting study teams and roles to ensure that statisticians and statistical principles permeate all aspects of a research enterprise.

We echo the call in Gibson (2017) for statisticians to better advocate for the importance of their involvement throughout the scientific process. Finally, we applaud stakeholders, such as the National Institutes of Health (Collins & Tabak 2014) and the American Association for the Advancement of Science (McNutt 2014) for leading the call for increased statistical rigor, and we encourage additional funding agencies, journals, hiring and promotion committees, and others to join in the call for higher scientific standards, statistical and otherwise. Science in the twenty-first century and beyond deserves nothing less.

References

Abrams, K. R., Spiegelhalter, D. & Myles, J. P. (2004), *Bayesian Approaches to Clinical Trials and Health Care*, John Wiley & Sons. New York.

Anderson, H. & Hepburn, B. (2016), 'Scientific method', `https://plato.stanford.edu/archives/sum2016/entries/scientific-method`. [Online; accessed 29-January-2018].

Anscombe, F. J. & Aumann, R. J. (1963), 'A definition of subjective probability', *The annals of mathematical statistics* **34**(1), 199–205.

Berry, S. M., Carlin, B. P., Lee, J. & Müller, P. (2010), *Bayesian Adaptive Methods for Clinical Trials*, Chapman&Hall/CRC Press, Boca Raton, FL.

Bradley, R., Bell, J., Gauld, J., Lilly, A., Jordan, C., Higgins, A. & Milne, R. (2005), 'UK soil database for modelling soil carbon fluxes and land use for the national carbon dioxide inventory', *Report to Defra Project SP0511. London: Defra* .

Brownstein, N. C. (2018), 'Perspective from the literature on the role of expert judgment and opinion in scientific and statistical research and practice'. Submitted.

Cairney, P. & Oliver, K. (2017), 'Evidence-based policymaking is not like evidence-based medicine, so how far should you go to bridge the divide between evidence and policy?', *Health Research Policy and Systems* **15**(1), 35.
URL: *https://doi.org/10.1186/s12961-017-0192-x*

Carlin, B. P., Chaloner, K., Church, T., Louis, T. A. & Matts, J. P. (1993), 'Bayesian approaches for monitoring clinical trials, with an application to toxoplasmic encephalitis prophylaxis', *The Statistician* **42**, 355–367.

Chaloner, K., Church, T., Louis, T. A. & Matts, J. P. (1993), 'Graphical elicitation of a prior distribution for a clinical trial', *The Statistician* **42**, 341–353.

Collins, F. S. & Tabak, L. A. (2014), 'NIH plans to enhance reproducibility', *Nature* **505**(7485), 612.

Cox, D. R. (1972), 'Regression models and lifetables (with discussion)', *JR Stat Soc Ser B Mathodol* **34**, 187–220.

Cripps, E., O'Hagan, A. & Quaife, T. (2013), 'Quantifying uncertainty in remotely sensed land cover maps', *Stochastic Environmental Research and Risk Assessment* **27**(5), 1239–1251.

Curtis, A. B., Worley, S. J., Adamson, P. B., Chung, E. S., Niazi, I., Sherfesee, L., Shinn, T. & St. John Sutton, M. (2013), 'Biventricular pacing for atrioventricular block and systolic dysfunction', *New England Journal of Medicine* **368**(17), 1585–1593.

Easterbrook, P. J., Gopalan, R., Berlin, J. & Matthews, D. R. (1991), 'Publication bias in clinical research', *The Lancet* **337**(8746), 867–872.

Ellenberg, S. S., Fleming, T. R. & DeMets, D. L. (2003), *Data monitoring committees in clinical trials: a practical perspective*, John Wiley & Sons.

Finfer, S. & Bellomo, R. (2009), 'Why publish statistical analysis plans', *Crit Care Resusc* **11**(1), 5–6.

Franco, A., Malhotra, N. & Simonovits, G. (2014), 'Publication bias in the social sciences: Unlocking the file drawer', *Science* **345**(6203), 1502–1505.

Fuller, R., Smith, G., Sanderson, J., Hill, R., Thomson, A., Cox, R., Brown, N., Clarke, R., Rothery, P. & Gerard, F. (2002), 'Countryside Survey 2000 Module 7. Land Cover Map 2000. Final Report', **CSLCM/Final**.

Garland Jr., T. (2016), 'Scientific method as an ongoing process', `https://en.wikipedia.org/w/index.php?title=Scientific_method&oldid=822947033`. [Archived at U. C. Riverside; retrieved from the original on 19-August-2016 into Wikipedia; accessed 29-January-2018].

Gelman, A. & Hennig, C. (2017), 'Beyond subjective and objective in statistics', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **180**(4), 967–1033.
URL: *http://dx.doi.org/10.1111/rssa.12276*

Gibson, E. W. (2017), 'Leadership in statistics: Increasing our value and visibility', *The American Statistician* .
URL: *https://doi.org/10.1080/00031305.2017.1336484*

Gosling, J. P. & O'Hagan, A. (2006), Understanding the uncertainty in the biospheric carbon flux for england and wales, Technical Report 567/06, University of Sheffield.
URL: *http://tonyohagan.co.uk/academic/pdf/UUCF.pdf*

Haines-Young, R., Barr, C., Black, H., Briggs, D., Bunce, R., Clarke, R., Cooper, A., Dawson, F., Firbank, L., Fuller, R. et al. (2000), *Accounting for Nature: Assessing Habitats in the UK Countryside*, Natural Environment Research Council and Centre for Ecology and Hydrology. Department of the Environment, Transport and the Regions.
URL: *https://trove.nla.gov.au/work/17604300*

Harris, K., O'Hagan, A. & Quegan, S. (2010), 'The impact of satellite-derived land cover uncertainty on carbon cycle calculations', `tonyohagan.co.uk/academic/pdf/CTCDPaper_v6.pdf`.

Hawkins, D. M. (2004), 'The problem of overfitting', *Journal of chemical information and computer sciences* **44**(1), 1–12.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T. & Jennions, M. D. (2015), 'The extent and consequences of p-hacking in science', *PLoS biology* **13**(3), e1002106.

HM Treasury (2015), 'The aqua book: guidance on producing quality analysis for government', *HM Government, London, UK, nd https://www. gov. uk/government/publications/theaqua-book-guidance-on-producing-quality-analysis-for-government (accessed July 10, 2017)* .

Jacobson, M., Besch, C., Child, C., Hafner, R., Matts, J., Muth, K., Wentworth, D., Neaton, J., Abrams, D., Rimland, D. & et al. (1994), 'Primary prophylaxis with pyrimethamine for toxoplasmic encephalitis in patients with advanced human immunodeficiency virus disease: Results of a randomized trial', *The Journal of Infectious Diseases* **169**, 384–394.

Kennedy, M., Anderson, C., O'Hagan, A., Lomas, M., Woodward, I., Gosling, J. P. & Heinemeyer, A. (2008), 'Quantifying uncertainty in the biospheric carbon flux for england and wales', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**(1), 109–135.

Kynn, M. (2008), 'The heuristics and biases bias in expert elicitation', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **171**(1), 239–264.

LaVange, L. M. (2014), 'The role of statistics in regulatory decision making', *Therapeutic Innovation & Regulatory Science* **48**(1), 10–19.

Lindsay, R. M. & Ehrenberg, A. S. (1993), 'The design of replicated studies', *The American Statistician* **47**(3), 217–228.

McNutt, M. (2014), 'Journals unite for reproducibility', *Science* **346**(6210), 679–679.

National Academies of Sciences, Engineering, and Medicine, and others (2017), *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*, National Academies Press.

National Research Council and others (2012*a*), *Assessing the reliability of complex models: mathematical and statistical foundations of verification, validation, and uncertainty quantification*, National Academies Press.

National Research Council and others (2012*b*), *Communicating Science and Engineering Data in the Information Age*, National Academies Press.

Oakley, J. (2016), 'Introduction to uncertainty quantification and Gaussian processes'. **URL:** *http://gpss.cc/gpuqss16/slides/oakley.pdf*

O'Brien, P. C. & Fleming, T. R. (1979), 'A multiple testing procedure for clinical trials', *Biometrics* pp. 549–556.

O'Hagan, A. (2006), 'Bayesian analysis of computer code outputs: A tutorial', *Reliability Engineering & System Safety* **91**(10-11), 1290–1300.

O'Hagan, A. (2012), 'Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux', *Environmental Modelling & Software* **36**, 35–48.

O'Hagan, A. (2018), 'Expert knowledge elicitation: Subjective, but scientific', p. Submitted.

Oliver, K. A. & de Vocht, F. (2017), 'Defining 'evidence' in public health: a survey of policymakers' uses and preferences', *European Journal of Public Health* **27**(Supplement 2), 112–117.

Oliver, K., Lorenc, T. & Innvær, S. (2014), 'New directions in evidence-based policy research: a critical analysis of the literature', *Health Research Policy and Systems* **12**(1), 34. **URL:** *https://doi.org/10.1186/1478-4505-12-34*

Ott, M. G. (1991), 'Importance the of the study protocol in epidemiologic research.', *J. OCCUP. MED.* **33**(12), 1236–1239.

Oxford University Press (2018), 'Scientific method', `https://en.oxforddictionaries.com/definition/scientific_method`. [Online; accessed 29-January-2018].

Pendergast, J. (2018), 'The scientific method from a biostatisticians perspective'. Submitted.

Pocock, S. J. (2006), 'Current controversies in data monitoring for clinical trials', *Clinical trials* **3**(6), 513–521.

Shih, J. (1995), 'Sample size calculation for complex clinical trials with survival endpoints', *Controlled Clinical Trials* **16**, 395–407.

Wasserstein, R. L. & Lazar, N. A. (2016), 'The ASA's statement on p-values: Context, process, and purpose', *The American Statistician* **70**(2), 129–133.

Woodward, F. I. & Lomas, M. R. (2004), 'Vegetation dynamics simulating responses to climatic change', *Biological Reviews* **79**(3), 643–670.