Statistical Science: Some Current Challenges

David R. Cox^{1,2} Christiana Kartsonaki^{3,4} Ruth Keogh⁵

¹Department of Statistics, University of Oxford, Oxford, England, United Kingdom,

²Nuffield College, University of Oxford, Oxford, England, United Kingdom,

³Medical Research Council Population Health Research Unit, Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, England, United Kingdom,

⁴Nuffield Department of Population Health, Medical Sciences Division, Nuffield College, University of Oxford, Oxford, England, United Kingdom,

⁵Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, England, United Kingdom

Published on: Jul 30, 2020

DOI: <https://doi.org/10.1162/99608f92.a6699bda>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

ABSTRACT

A broad review is given of the role of statistical concepts in the design of studies, in various aspects of data collection and definition and especially in the analysis and interpretation of data. Outline examples are given from various fields of application with some emphasis on epidemiology and medical statistics. The role of probability in various stages of investigation is outlined, in particular its place in the assessment of uncertainty in the conclusions.

Keywords: interpretation, metrology, probability, statistics, study design, uncertainty

---

1. Introduction

Unexplained variation is everywhere from everyday life to the most specialized research areas in the natural and social sciences and their associated technologies. In some situations, such variation is virtually ignorable, in others it needs careful study for its control and for cautious assessment of its consequences. In each specific context these are primarily issues for discussion within the subject-field involved. Yet some features have strong common elements that may reasonably be called statistical. The long history of statistical work deals partly with implications for the design of observational and experimental studies, but for the most part concentrates on methods of analysis with a base in probability models. Central though that strand remains, there are many other features to the contribution of statistical thinking to investigations, experimental or observational, and this article aims to discuss and illustrate some of these.

The very welcome rich mix of perspectives captured in this new journal implicitly calls for clarification of the more traditional theme of statistical science. While demarcation of rigid boundaries is usually counterproductive, it is helpful to think of a number of broad components to statistical ideas (Cox et al., 2018). One such formulation is

- specification and clarification of questions for study
- planning of investigations
- issues of metrology
- various stages of data collection and analysis, which could well be richly subdivided
- interpretation and conclusions.

These themes have in each specific application a clear subject-matter focus. Nevertheless, the general ideas involved raise concepts and techniques of broad applicability, which it is in principle the objective of statistical theory to encapsulate and develop. In this article we comment first on the importance of clarification surrounding the questions for study, and then concentrate on aspects of the analysis of data, reflecting the preponderance of that aspect in the more formal literature. Our general emphasis is on studies relating one or

more explanatory variables, possibly themselves in a meaningful sequence, and possibly large in number, to one or more outcomes. The resulting analysis may be largely descriptive or, at the other extreme, may hinge on a specific stochastic formulation relating the data to the subject matter–generating process, typically by some form of Markov process, studied by computer simulation or where feasible by mathematical analysis.

2. Specification of Questions for Study

Statistical considerations often play a key role in the evolution of an initial question, typically posed by subject matter experts, which could be broadly defined, into one or more specific questions. A clear focus, even if it concerns general exploration, is essential to planning the collection, analysis, and interpretation of data. The whole process demands a collaborative approach.

Explicit specification of research questions is a starting point for many investigations. For example, a start might be the question: How does the strength of a particular material depend on the temperature and relative humidity in which it is used? In some contexts, a single, sharply focused question may be most fruitful, concentrating on a narrow band of temperatures and relative humidities and a specific material of immediate concern. In others, a more wide-ranging study, involving more than one material and a wider range of other features, would be much more enlightening. There is a powerful statistical literature on the special experimental designs needed to explore relatively complex patterns of dependence of outcome on several or indeed many features. In a medical setting a broad question of interest may be whether a particular medication is successful in lowering blood pressure. To turn this into a research question requires further clarifications, such as deciding the population of interest, whether there should be focus on a choice of dose and of appropriate duration of use, and whether the effect of the medication may differ according to other factors such as age or comorbidities.

In some situations, experimental and observational, posing a series of interlocking questions may be more fruitful than exclusive emphasis on a single point. This is in contrast with aiming to choose between, say, two sharply focused explanations of an effect, a so-called critical experiment.

Questions fall broadly into three types focused respectively on:

1. explaining the mechanistic, or causal, relationships between explanatory variables and one or more outcome variables,
2. predicting an outcome based on a set of explanatory variables based at least in part on an understanding of underlying process,
3. providing a preliminary essentially descriptive account of interdependencies among a set of variables, treating variables collected for different purposes appropriately.

Questions of the first type usually focus ultimately on a relatively small number of explanatory variables, though a larger set may play a role in the analysis to achieve secure conclusions about systematic relationships

with at least some underlying subject-matter, that is, mechanistic, explanation. The second and third types of question in some cases may involve a large set of explanatory variables. In the third type of question the roles of variables as explanatory or outcome may not be well defined. Exploratory analysis is often a base for moving to questions of the first two types. Schmueli (2010) provides a detailed discussion of the contrast between explanation and prediction. Hernan et al. (2019) classified data science tasks into those concerning description, prediction, and counterfactual prediction (meaning causality). See also Hand (2019) for a nice discussion on this topic.

3. Metrology

The ability to measure features reasonably accurately and reliably underpins many investigations. Striking examples are the various physics-based devices used in many fields, for instance, in preliminary assessment of patients and in guiding surgical intervention. Statistical issues receive particularly explicit prominence in such contexts where there is a contractual element involved. The sources of variability associated with the system need estimation and quality-control type procedures for their control established. The work of such organizations as the National Bureau of Standards and British Standards Institution depends on such themes. The broad statistical topic of components of variance is widely relevant in this and other contexts.

In many fields, the methods of measurement of key features are so well established that they may need little special attention, although a planned sampling scheme of occasional checks of underlying calibration may be very desirable. In some applications advances in technology lead to information not in directly analyzable form. For example, environmental or body-worn sensors, satellite data on weather and ambient air pollution, medical imaging and genomics or other biological data are often in unprocessed form, generating large collections of observations that cannot be represented by a single matrix. There is often more than one way in which such data can be collapsed into a form suitable for analysis. Specialized software is usually needed to process such data before analysis, that may or may not be specific to the devices that collect them, and because often the unprocessed data require a very large amount of storage, some reduction of data is also done. Choice between alternative methods for such preprocessing is not trivial, and in some fields where they have been compared, the choice has been shown to influence the final results of analysis substantially.

In some applications the uncertainty around the summary features used in subsequent analysis is ignored. In many contexts, however, more formal statistical analysis of the sources of variation involved is desirable. Preliminary exploration of the data may be needed to guide any further analysis. The data may be visualized and reviewed for completeness and unexpected values isolated that may not be detectable from the results of more complex analysis. For high-dimensional data, techniques such as principal component analysis or other dimension reduction methods may be used to aid visualization.

In some fields of study appreciable errors of measurement are pervasive. Measurements made on humans are often subject to error of some kind, including biological ones made in a laboratory and particularly those

obtained through questioning of a study participant or recorded by a doctor. Attention should be given to the presence of error or variation in the specific context of the investigation. For example, even though a blood pressure-measuring device may provide an accurate measure of blood pressure at a given moment, it may be the average blood pressure or the variation in blood pressure over some period that is of true interest. Other examples of error prone measurements, in rough order of increasing severity, include body mass index, dietary intake, and health-related quality of life.

The importance of errors of measurement, and more generally of sources of variability of no specific interest, depends on the research question. Two different aspects are the degradation of effectiveness induced by such errors and the allowance for such errors in reporting conclusions. When the aim is prediction, modest errors of measurement may not be of pressing concern if the error-prone measurements used in the development of a prediction model are the same as those that will be available in the intended population in which the prediction model is to be applied. On the other hand, where the aim is to understand mechanistic or causal relationships, errors of measurement in the variables at play can result in spurious findings. The bias may be in any direction and may be particularly insidious when the very extent of the data may give an illusion of security. That is, researchers should be aware of the danger of estimates that are made with high precision (narrow confidence intervals, small p -values), but that in fact are subject to substantial bias.

Measurement issues demand consideration at the study design phase. In some situations, additional measures of some kind (e.g., repeated blood pressure measurements over a period) may need to be incorporated to enable assessment of the impact of measurement error on the analysis. There is a large statistical literature on methods of analysis for understanding and mitigating the impacts of measurement error (Carroll et al., 2006 Lash et al., 2009).

4. Planning of Studies

In some contexts, the collection of new data is essential and its planning therefore crucial. If feasible, questions about mechanistic relationships are relatively likely to be addressed by one or more experiments, that is, controlled interventions. An example is the comparison of alternative medical treatments by randomized controlled trials. Such trials are based on ideas stemming originally from agricultural field trials, rapidly taken over into other fields, initially wide-ranging applications in other biological areas. Later, with appreciable further development, industrial applications became increasingly important. Mathematically, parts of the topic have important combinatorial links with coding theory and the theories of finite groups and Galois fields (Cox & Reid, 2000).

Some types of question may be tackled using existing data sources, for example, those arising from routinely collected medical records or more broadly from the routine operation of an administrative system. Data from more than one source may have to be linked in some cases. Issues of study design are far from redundant just because large amounts of data are involved. Understanding of how the data arose, and potential selection biases

are key to achieving secure findings. In some cases, it may be necessary or prudent to reduce the available data to a smaller subset for reasons of practical feasibility or computation. To obtain and analyze such samples appropriately and efficiently requires the principles of study design and appreciation of their implications for analysis. In some circumstances, routinely collected data may provide natural experiments that can be exploited to address certain questions. Casey et al. (2016) review the use of electronic health records in public health research and discuss their use in different types of investigation, such as those noted in Section 2, issues of sampling, and sources of bias. Gail et al. (2019) give guidance on study design in the context of observational studies of the effect of exposure on disease incidence, emphasizing some key issues in defining the research question and the implications for design. Madigan et al. (2013) highlight different sources of heterogeneity in findings from observational studies within large existing databases, including the role of study design. In the epidemiological literature, there is a recent emphasis on a technique of emulating a target randomized trial using observational data to facilitate investigations of causal questions and avoid some of the bias that can be pervasive in investigations using observational data (Hernan & Robins, 2016).

In some contexts, there is a reasonably secure tradition of careful investigation. In others where there is no such history, checks of the first stages of data collection are important, and in some situations, techniques with a long history in industrial quality control may be adaptable to ensure data quality. In any investigation taking an appreciable time to complete, intermediate stages of analysis are likely to be necessary and may raise delicate issues of accountability and the need in many cases to avoid biases by preserving due confidentiality. Especially in studies continuing over a long period, aspects of the data unanticipated in advance may demand attention but need special care, especially if there is a serious impact on the conclusions.

5. Some Specific Challenges

In this section we discuss specific challenges, with reference to the three broad types of research question discussed in Section 2.

5.1. Causality

An important type of question concerns the understanding of mechanistic or causal relationships. Not that long ago *causality* was, in much scientific work, regarded as a word best avoided. Usage has changed, and recent years have seen much explicit discussion of causality and of statistical methods and general reasoning for investigating relevant issues. There has been some emphasis on observational, as opposed to experimental, data, building on decades of earlier development. One major challenge of causal questions, particularly in the context of observational data, is to allow for possible confounding: that is, of whether what appears to be the dependence of an outcome on one specific explanatory variable is in fact the consequence of another

However, firm establishment of causality even from a controlled experiment may not be straightforward. For example, even in a well-conducted randomized controlled trial comparing a new treatment with the standard

care, issues of drop-out, adherence, and generalizability across different subgroups bring challenges for drawing causal conclusions.

Pioneering work on causality, initially in the context of agricultural field trials, was contained in R. A. Fisher's formulation of the principles of experimental design (Fisher, 1926, 1935), namely, replication, error control, randomization, and the notion of factorial design. So far as we know, Fisher in his published work used the word *causal* only rarely, occasionally in connection with his views on natural selection (Fisher, 1934) and later in the discussion of the observational evidence about smoking and lung cancer. This limit seems in line with mainstream usage of the word at the time. His ideas on experimental design spread rapidly into other fields.

In 1965 W. G. Cochran, who had worked with Fisher and his close colleague Yates, before emigrating, returned to the United Kingdom briefly from Harvard and read an important paper to the Royal Statistical Society stressing the need to study observational investigations in more depth. At the same time, Bradford Hill (1965) formulated his guidelines, pointers toward causality rather than necessary or sufficient conditions. Especially in those fields where studies designed from first principles are not possible and a relatively secure route to a causal interpretation is not available, Bradford Hill's guidelines retain their relevance, not least in helping pinpoint issues for study. Cox and Wermuth (1996, chap. 8) discussed the role of notions of causality in interpreting data, including the use of graphical representations. An influential review by Holland (1984) was followed by wide-ranging, if occasionally rather surprisingly negative, discussion. Pearl introduced a more explicit formulation in terms of his do-operator and its properties; for an enthusiastic account, see Pearl & MacKenzie (2019). A broad and original account of methods concerned with causal inference in a broadly epidemiological context is given by Hernan and Robins (2020). For a broad-ranging discussion from a rather different perspective, see Imbens and Rubin (2015).

In recent work, there is sometimes the implication that causality can be established securely by largely formal arguments. A cautious approach sees three broad meanings of causality. The first, Wiener-Granger causality, is essentially concerned with the prediction of one time series from past values of another and will not be considered further here. The second meaning, and the one implicit in most recent discussions, is based on the idea, roughly expressed, that had the potential cause C been different from how it is, other things being the same, then the outcome or response R would have been different. This requires precise formulation of the various stages, in particular of what is held fixed as hypothetically C is changed. The third version of causality requires the second plus some evidence-based explanation of the proposed causal effect in terms of underlying process. Evidence-based here means essentially having directly established relevance, not merely hypothesized relevance. See, for example, Cox and Donnelly (2011, chap. 9).

In many fields a single study is rarely sufficient to infer causality, especially if observational rather than experimental. Even results of a study having been reproduced in one or more different studies of a similar approach are not in general secure grounds for inferring causality as limitations or potential biases in one study are also likely to be present in other similar studies. For example, commonly in the medical sciences systematic

reviews often accompanied by meta-analyses are performed, which collate the available evidence on a particular question, subject to selected inclusion criteria. The extent of heterogeneity between estimates from identified studies may help identify whether there is one underlying association, but it can be difficult to combine estimates from studies with different characteristics. In network meta-analysis some relationships may be inferred indirectly. However, studies that are similar enough to be combined are also likely to have similar types of biases. In some fields evidence from different approaches, including from controlled experiments, is needed in order to conclude that a causal relation exists. However, it may not be straightforward to combine different approaches to estimate one underlying causal effect because these different approaches may be targeting different estimands.

Time is an inherent aspect of any question, data set and analysis. The planning and analysis to answer the question of interest should include consideration of the ordering of variables conceptually in time and, if different, the ordering in which they were observed or collected. The use of causal diagrams, such as directed acyclic graphs (DAGs), can be used to describe formally such aspects of the data and system under study. These approaches can help to clarify what is being estimated in an analysis, what the potential sources of bias are—for example selection bias and confounding—and help inform the formal analysis, which may require specialized methods or only standard methods such as regression, depending on the complexity of the situation and the questions of interest.

The econometric technique of instrumental variables may sometimes be used in which assumptions about potential conditional independences are used to enhance clarity and effective precision of interpretation. An adaptation of this to the use of genetic information is the basis of Mendelian randomization (example.g., Burgess and Thompson, 2015). There may, however, be wide limits of uncertainty. The idea behind the broader technique of instrumental variables is the use of a new variable that influences the outcome primarily via the explanatory variable of interest. Then the effect of the explanatory variable on the outcome can be inferred from the effect of the instrumental variable on the explanatory variable and from the effect of the instrumental variable on the outcome. More broadly, prior assumptions, often untestable, of some conditional independences within a complex system may enhance the estimation of primary features.

Another rather specialized aspect calling for a clear statement of objectives and careful assessment of the impact of haphazard variation concerns the analysis of a form of genetic data in which a very large number of potential explanatory variables are available for a limited number of individuals, say 10^5 variables for 100 or so individuals. Clearly, progress is possible only under an implicit or explicit assumption of sparsity, that is, that only a few of the explanatory variables have a nonnull effect. If empirical prediction is the objective, the lasso (Tibshirani, 1996) provides an effective answer. For interpretation and comparison with other sets of data, some explicit notion is usually needed that alternative models, virtually equally effective, are available, ideally leading toward a confidence set of potential explanations (Battey & Cox, 2018; Cox & Battey, 2017). The study of the relation between individual genomic data and disease is an important context.

5.2. Generalizability Versus Specificity

Two superficially contrasting but in fact intertwined ideas are those of generalizability and specificity. As to the first, suppose that secure conclusions are obtained, possibly on a large sample of study individuals. Why should the conclusions apply in some new somewhat different contexts? Possible justifications include

1. because there is an evidence-based underlying explanation of the conclusions that is likely to retain relevance,
2. because the conclusions of the initial study have been shown empirically to be stable over a range of conditions (in statistical jargon, there is no important interaction involving primary comparisons and individual characteristics),
3. because other studies of very different kinds point in the same direction,
4. because the data are an appropriately chosen random sample from the target population of concern.

The last of these would be relevant, for example, in those types of sampling inspection, auditing, and quality control in which carefully specified sampling schemes can be enforced. In many fields, however, the notion that the data are in a meaningful sense a random sample from the target population of future applications is fanciful. Formal discussion of the third aspect is difficult for explicit statistical discussions in which the primary emphasis is on individually secure investigations.

In apparent contrast, specificity is concerned with single instances of the applicability of conclusions. Why should the conclusions from an often broad-ranging investigation apply in this specific instance? In the special case of a clinical study, why should the treatment of proved overall benefit in a well-designed clinical trial apply to this specific patient awaiting a clinician's decision or recommendation. The first two considerations listed are key aspects.

Statistical thinking often places emphasis on the security of individual studies. Important understanding very often, however, stems from the combination of evidence of different kinds. Thus, as late as 1959, some years after the first clear evidence of a link between smoking and lung cancer, three leading statisticians from very different backgrounds, R. A. Fisher, J. Neyman, and J. Berkson were unconvinced of the evidence for a causal connection. Cornfield et al. (1959/2009) showed that when four different kinds of study were considered together, the argument for a causal dependence became very strong. This type of evidence may be called triangulation (Egger & Davey Smith, 1995). It is in line with the comment of Fisher to Cochran (1965) that one should "make one's theories elaborate," which Fisher clarified to mean encompass as rich a variety of types of evidence as feasible. This is distinct from the emphasis in much statistical discussion, which is on the design and analysis of individually secure studies. Cox and Donnelly (2011, chap. 9) give a wide-ranging discussion.

For discussion of these broad issues, see, for example, Keiding and Louis (2016) and Meng (2018).

6. Some Specific Applications

In this section we outline some examples from our own experiences highlighting some of the issues discussed above.

6.1. A Veterinary Example

Some of the issues likely to arise in experimental study of a relatively complex situation are illustrated by an investigation of the possible impact on bovine tuberculosis (TB) of that same disease in badgers. In outline, trial areas were designated in 10 sets of 3, the plots in each triplet being geographically close and similar. There were three ‘treatments’ randomly allocated within each triplet. They were intensive regular culling of badgers, localized culling in reaction to local new occurrence in cattle, and a control, no culling group. The incidence of bovine TB in the area was the outcome of interest. An initial assessment and plan of analysis suggested that a 20% reduction in incidence would have a good chance of detection.

The planning, implementation, and analysis of the trial took approaching 10 years. What may seem a long period included recruiting and training several teams of fieldworkers, negotiating with farmers to achieve cooperation in a randomized trial, and formulating a reasonably detailed proposal for analysis. An outbreak of foot and mouth disease was a major disruption of much agricultural activity for a period and partly accounts for the long time. An intermediate analysis after a few years showed an apparent effect of the expected direction in the intensive culling group but, entirely surprisingly, an adverse effect in the localized culling group; that is, they had higher incidence than the controls (Donnelly et al., 2003). The localized culling was abandoned. The planning was guided to some extent by a simple stochastic model representing the interaction of disease between two species incorporating infection between and within species (Cox et al., 2005).

In this particular context, after some further discussion, two contrasting plausible explanations were identified. One was that farmers in the control areas had illegally culled badgers and done so very effectively, that is, the supposedly control areas were in fact effectively culled. The other explanation was that badgers remaining in the culled area foraged more widely, a phenomenon that animal ecologists call perturbation. A subsidiary new investigation refuted the former and confirmed the latter explanation.

This illustrates a general situation. Data contradict prior information. This demands a search for explanation. It will usually be possible to construct one or more plausible retrospective explanations, but explanations arrived at this way tend to be very fragile. Their backing by totally independent evidence is mandatory.

6.2. Some Considerations in the Physical Sciences

While the method of least squares, a pivotal technique in much statistical analysis, had its origins from physical science issues, more recently the high precision achievable in much laboratory work has in some areas limited the need for deep statistical analysis. By contrast, developments in particle physics and in astrophysics have called for careful statistical discussion. In the investigation that led to confirmation of the existence of the

Higgs boson, essentially Poisson distributed noise was a key feature of the data and the analysis of the very large amount of data generated, while needing much caution, had a strong statistical core that received extensive preliminary discussion. See, for example, Heinrich and Lyons (2007). This is a conceptually unusual situation in that there is no special interest in how many Higgs might be produced in the collider, but concern is solely in whether there are enough to be detectable above the background noise.

In some engineering contexts statistical concepts have been prominent, for example, in the study of the strength of materials and structures. The recognition that strength may hinge on the smallest of a large number of features, the weakest link hypothesis, led to theoretical statistical development with wide applicability (Fisher & Tippett, 1928; Gnedenko, 1943). For a wide-ranging survey of statistical applications, see Coles (2001). There are links, largely unexplored, with survival analysis in a medical and epidemiological context.

6.3. Investigations Using a Disease Registry

Data arising from routine collection or observation are often used for wide-ranging investigations, as noted in Section 4. The U.K. Cystic Fibrosis (CF) Registry is one such example. This is a patient registry, established in 1996, that collects clinical and demographic data on individuals with CF, a chronic and progressive inherited disease, in the United Kingdom. Data are obtained at designated clinic visits approximately annually and recorded through a centralized data entry system and using standardized variable definitions. Information has been collected on more than 12,000 individuals on several hundred variables (Taylor-Robinson et al., 2017). The data have been used in a variety of investigations to address different types of research questions, including those of the three types noted in Section 2. An annual report on the registry contains a description of the patient population covering both demographic information and summaries of the distribution of clinical variables such as lung function. The methods used in producing the report include simple descriptive summaries such as percentages and means, alongside statistical analysis so as to provide updated estimates of life expectancy.

These data have also been used to develop prediction models and in studies aimed at understanding causal relationships. Keogh et al. (2019) derived a model for prediction of survival from a given age, given an individual's current health status, as summarized through clinical measures, alongside their demographic characteristics. The variables used in the model were identified from previous literature on factors associated with survival, as well through clinical input. The analysis used the landmarking approach for time-to-event and longitudinal data.

Another recent study using these data investigated questions concerning the causal impact of use of a commonly used treatment, DNase, on two important health outcomes, lung function (measured using FEV1%, a continuous variable) and number of days per year spent using intravenous antibiotics (a count variable) (Newsome et al., 2018), which are measured at the annual visits recorded in the registry database. DNase, which is administered via a nebulizer, is prescribed to aid in airway clearance and the decision to start

prescribing it to a patient depends on many factors, including pretreatment measures of the two outcomes of interest in this study. The focus was on estimating the expected difference in the two outcomes between individuals using DNase for 1 year up to 5 years compared to nonusers. Observations of treatment use, the outcomes, and a large number of other variables were available from the approximately annual visits made between 2007 and 2015. An understanding of how the treatment is assigned by clinicians and about the temporal ordering of different variables were key and directed acyclic graphs were used to visualize assumed relationships between variables over time. Questions about sustained treatment use over more than 1 year are particularly challenging to answer using longitudinal observational data due to time-dependent confounding of the association between treatment assignment and the clinical outcomes. This was addressed using techniques from the causal inference literature, including marginal structural models fitted using inverse probability of treatment weighting, g-computation, and g-estimation. Assumptions of the analyses were listed and some could be assessed through sensitivity analysis. For example, assumptions about temporal ordering of measures observed at the same time point were explored by examining what happened if orderings were reversed.

6.4. An Epidemiological Example

Diabetes is associated with a higher risk of pancreatic cancer. However, individuals with pancreatic cancer are typically diagnosed late when the disease is already advanced. There are at least three potential mechanisms underlying this observed association. Having diabetes may cause pancreatic cancer, via hyperglycaemia, insulin resistance, or other processes; individuals with pancreatic cancer may be more likely to have their previously undiagnosed diabetes diagnosed due to increased contact with health care professionals during consultations for symptoms caused by their cancer, or vice versa; and diabetes being a direct manifestation of their yet undiagnosed cancer. Pang et al. (2017) analyzed data from the China Kadoorie Biobank (CKB), a prospective cohort study of 0.5 million adults and meta-analyzed estimates from other prospective cohort studies. Participants of the CKB were asked, among other questions, if they have diabetes, and if so for how long, and were also screened for yet undiagnosed diabetes upon entry into the study. They were subsequently followed up for incidence of disease and death. This was done by linkage with health insurance, death, and disease registries. Pang et al. investigated the relationship between diabetes and risk of pancreatic cancer to help understand the underlying disease mechanisms. They assessed associations between prior self-reported and screen-detected diabetes at study baseline, as well as of incident diabetes during follow-up, with risk of pancreatic cancer. They also examined the association between duration of diabetes and pancreatic cancer, as association of recently diagnosed diabetes supports the last two potential mechanisms, whereas stronger association with longer duration would support the first. The association of blood glucose with risk of pancreatic cancer among individuals without diabetes was also assessed. The study design allowed investigation of how the relationship varies with time since diagnosis of diabetes, which can help understand to what extent each of the hypothesized mechanisms is supported. By comparing the associations of self-reported physician-diagnosed diabetes and of screen-detected diabetes with risk of pancreatic cancer they tried to

examine to what extent there were any differences between individuals with known and treated diabetes and those with yet unknown and untreated and perhaps more recently developed diabetes.

7. Discussion

There is an aspect that forms the core of the formal statistical literature but may be slightly less central in specific investigations. This is the use of probabilistic concepts in one or more of several different ways. One is to give a formal representation of the situation under study so as to focus discussion. In some situations, this may be by a specific stochastic process representing in suitable detail the data-generating process. Analysis of such models may be through a variety of methods, ranging from the rich mathematical theory available to direct simulation. Techniques from the design of experiments may in the latter case be helpful. More commonly, a suitable member of a broad family of models of interrelation is used, often methods that are in a sense a generalization of the linear models underpinning least squares regression. Such models may guide the approach to analysis and interpretation so that, for example, possibly complex patterns of interrelation can be captured in understandable form. Then there is the detailed choice of techniques of design and analysis, including especially the formal assessment of the security of the conclusions via such concepts as confidence intervals, posterior intervals, and tests of hypotheses, or with some shift of emphasis, significance tests, and so on. Vigorous discussion and differences of opinion over the best approach to these issues have continued for more than 200 years. This body of work implicitly raises important questions ranging from the differing purposes underlying research to concerns with the underlying interpretation to be attached to the notion of probability. The massive literature on theoretical statistics represents the very extensive and wide-ranging nature of academic writing in this field. Kolmogorov (1933) formulated the axioms of mathematical probability in the form most widely used, thus detaching the mathematics of probability into a vigorous and rigorous branch of pure mathematics, divorced from considerations of meaning. Kolmogorov himself, however, was deeply involved in applications and interested in various interpretations of probability. For statistical work most workers have, implicitly at least, seen the need for two interpretations of probability, one as a descriptor of random phenomena and one as concerned with assessing uncertainty.

Understanding the strengths and limitations of probabilistic arguments in the design of studies and more particularly in their analysis is central and although, as so often with foundational questions, the impact on day-to-day work is important, it may be indirect. The most likely to lead to serious misunderstanding concerns the distinction between the multiple possible roles of significance tests in the interpretation of data. Cox (2020) set out four distinct contexts for the use of such tests.

One message of the present article is that the general theoretical treatment of statistical models and, in particular, the role of probability in describing systems under study and assessing uncertainty is central. It has a long rich background and is evolving in a challenging way. In addition, there are other major and wide-ranging aspects to statistical thinking, these enhancing the importance and variety of the field.

Acknowledgments

We are grateful to referees for very constructive comments.

Disclosure Statement

David R. Cox, Christiana Kartsonaki, and Ruth Keogh have no financial or non-financial disclosures to share for this article.

References

- Battey, H., & Cox, D. R. (2018). Large numbers of explanatory variables: A probabilistic assessment. *Proceedings of the Royal Society A*, 474(2215), Article 20170631. <https://doi.org/10.1098/rspa.2017.0631>
- Burgess, S., & Thompson, S. G. (2015). *Mendelian randomization: Methods for using genetic variants in causal estimation*. CRC Press.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Chapman and Hall/CRC.
- Casey, J. A., Schwarz, B. S., Stewart, W. F., & Adler, N. E. (2016). Using electronic health records for population health research: A review of methods and applications. *Annual Review of Public Health*, 37, 61–81. <https://doi.org/10.1146/annurev-publhealth-032315-021353>
- Cochran, W. G. (1965). The planning of observational studies of human populations (with discussion) *Journal of the Royal Statistical Society*, 128(2), 234–266. <https://doi.org/10.2307/2344179>
- Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer.
- Cornfield, J., Haenszel, W., Cuyler Hammond, E., Lilienfeld, A. M., Shimkin, M. B., & Wynder, E. L. (2009). Smoking and lung cancer: Recent evidence and discussion of some questions. *International Journal of Epidemiology*, 38(5), 1175–1191. <https://doi.org/10.1093/ije/dyp289> (Reprinted from “Smoking and lung cancer: Recent evidence and discussion of some questions,” 1959, *Journal of the National Cancer Institute*, 22(1), 173–203)
- Cox, D. R. (2020). Statistical significance. *Annual Review of Statistics and its Application*, 7, 1–10. <https://doi.org/10.1146/annurev-statistics-031219-041051>
- Cox, D. R., & Battey, H. (2017). Large numbers of explanatory variables: a semi-descriptive analysis. *Proceedings of the National Academy of Science*, 114(32), 8592–8595. <https://doi.org/10.1073/pnas.1703764114>

- Cox, D. R., & Donnelly, C. A. (2011). *Principles of applied statistics*. Cambridge University Press.
- Cox, D. R., Donnelly, C. A., Bourne, F. J., Gettinby, G., McInerney, J. P., Morrison, W. I., & Woodroffe, R. (2005). Simple model for tuberculosis in cattle and badgers. *Proceedings of the National Academy of Science*, 102(49), 17588–17593. <https://doi.org/10.1073/pnas.0509003102>
- Cox, D. R., Kartsonaki, C., & Keogh, R. H. (2018). Big data: Some statistical issues. *Statistics & Probability Letters*, 136, 111–115. <https://doi.org/10.1016/j.spl.2018.02.015>
- Cox, D.R., & Reid, N. (2000). *The theory of the design of Experiments*. Chapman and Hall/CRC.
- Cox, D. R., & Wermuth, N. (1998). *Multivariate dependences*. Chapman & Hall.
- Donnelly, C. A., Woodroffe, R., Cox, D. R., Bourne, J., Gettinby, G., Le Fevre, A. M., McInerney, J. P. & Morrison, W. I. (2003). Impact of localized badger culling on tuberculosis incidence in British cattle. *Nature*, 426(6968), 834–837. <https://doi.org/10.1038/nature02192>
- Egger, M., & Davey Smith, G. (2001). Principles and procedures for systematic reviews. In *Systematic reviews in health care: Meta-analysis in context (2nd ed.)*. BMJ Public Library. <https://doi.org/10.1002/9780470693926.ch2>
- Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33, 503–513. <https://doi.org/10.23637/rothamsted.8v61q>
- Fisher, R. A. (1934). Indeterminism and natural selection. *Philosophy of Science*, 1(1), 99–117. <https://doi.org/10.1086/286308>
- Fisher, R. A. (1935). *Design of experiments*. Oliver and Boyd.
- Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2), 180–190. <https://doi.org/10.1017/S0305004100015681>
- Gail, M. H., Altman, D. G., Cadarette, S. M., Collins, G., Evans, S. J. W., Sekula, P., Williamson, E., & Woodward, M. (2019). Design choices for observational studies of the effect of exposure on disease incidence. *BMJ Open*, 9(12), Article e031031. <http://doi.org/10.1136/bmjopen-2019-031031>
- Gnedenko, B. V. (1943). Sur la distribution limite du terme maximum d’une serie aleatoire [On the limiting distribution of the maximum term of a random sequence]. *Annals of Mathematics*, 44(3), 423–453. <https://doi.org/10.2307/1968974>

- Hand, D. (2019). What is the purpose of statistical modelling? *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.4a85af74>
- Heinrich, J., & Lyons, L. (2007). Systematic errors. *Annual Review of Nuclear and Particle Science*, 67, 145–169. <https://doi.org/10.1146/annurev.nucl.57.090506.123052>
- Hernan, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology* 183(8), 758–764. <https://doi.org/10.1093/aje/kwv254>
- Hernan, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC.
- Hernan, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance*, 32(1), 42–49. <https://doi.org/10.1080/09332480.2019.1579578>
- Bradford Hill, A. (1965). The environment and disease: Association or causation. *Proceedings of the Royal Society of Medicine*, 58(5), 295–300. <https://doi.org/10.1177/003591576505800503>
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social science, and biomedical sciences*. Cambridge University Press.
- Keiding, N., & Louis, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society A*, 179(2), 319–376. <https://doi.org/10.1111/rssa.12136>
- Keogh, R. H., Seaman, S. R., Barrett, J. K., Taylor-Robinson, D., & Szczesniak, R. (2019). Dynamic prediction of survival in cystic fibrosis: A landmarking analysis using UK patient registry data. *Epidemiology*, 30(1), 29–37. <https://doi.org/10.1097/EDE.0000000000000920>
- Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung [Foundations of the theory of probability]*. Springer. <https://doi.org/10.1007/978-3-642-49888-6>
- Lash, T., Fox, M., & Fink, A. (2009). *Applying quantitative bias analysis to epidemiologic data*. Springer. <https://doi.org/10.1007/978-0-387-87959-8>
- Madigan, D., Ryan, P. B., Schuemie, M., Stang, M., Overhage, J. M., Hartzema A. G., Suchard, M. A., DuMouchel, W., & Berlin, J. A. (2013). Evaluating the impact of database heterogeneity on observational study results. *American Journal of Epidemiology*, 178(4), 645–651. <https://doi.org/10.1093/aje/kwt010>
- Meng, K. L. (2018). Statistical paradoxes and paradoxes in big data. I: Law of large populations, big data paradox and the 2016 Presidential election. *Annals of Applied Statistics*, 12(2), 685–726. <https://doi.org/10.1214/18-AOAS1161SF>

Newsome, S. J., Keogh, R. H., & Daniel, R. M. (2018). Estimating long-term treatment effects in observational data: A comparison of the performance of different methods under real-world uncertainty. *Statistics in Medicine*, 37(15), 2367–2390. <https://doi.org/10.1002/sim.7664>

Pang, Y., Kartsonaki, C., Guo, Y., Bragg, F., Yang, L., Bian, Z., Chen, Y., Iona, A., Millwood, I. Y., Chen, J., Li, L., Holmes, M. V., & Chen, Z. (2017). Diabetes, plasma glucose and incidence of pancreatic cancer: A prospective study of 0.5 million Chinese adults and a meta-analysis of 33 cohort studies. *International Journal of Cancer*, 140(8), 1781–1788. <https://doi.org/10.1002/ijc.30599>

Pearl, J., & MacKenzie, D. (2019). *The book of why: The new science of cause and effect*. Penguin.

Schmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>

Taylor-Robinson D., Archangelidi, O., Carr, S. B., Cosgriff, R., Gunn, E., Keogh, R. H., MacDougall, A., Newsome, S., Schlüter, D. K., Stanojevic, S., Bilton, D., & the CF-Epinet collaboration. (2018). Data Resource Profile: The UK Cystic Fibrosis Registry. *International Journal of Epidemiology*, 47(1), 9–10e. <https://doi.org/10.1093/ije/dyx196>

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, 58(1), 267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>

©2020 David Cox, Christiana Kartsonaki, and Ruth H. Keogh. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.