Johns Hopkins University
Johns Hopkins University, Dept. of Biostatistics Working Papers

Year 2004 *Paper* 32

A Cox Model for Biostatistics of the Future

Scott L. Zeger* Peter J. Diggle†
Kung-Yee Liang‡

*The Johns Hopkins Bloomberg School of Public Health, szeger@jhsph.edu
†Lancaster University, United Kingdom, p.diggle@lancaster.ac.uk
‡Johns Hopkins Bloomberg School of Public Health, kyliang@jhsph.edu

A Cox Model for Biostatistics of the Future

Scott L. Zeger, Peter J. Diggle, and Kung-Yee Liang

Abstract

Professor Sir David R. Cox (DRC) is widely acknowledged as among the most important scientists of the second half of the twentieth century. He inherited the mantle of statistical science from Pearson and Fisher, advanced their ideas, and translated statistical theory into practice so as to forever change the application of statistics in many fields, but especially biology and medicine. The logistic and proportional hazards models he substantially developed, are arguably among the most influential biostatistical methods in current practice.

This paper looks forward over the period from DRC's 80th to 90th birthdays, to speculate about the future of biostatistics, drawing lessons from DRC's contributions along the way. We consider "Cox's model" of biostatistics, an approach to statistical science that: formulates scientific questions or quantities in terms of parameters gamma in probability models f(y; gamma) that represent in a parsimonious fashion, the underlying scientific mechanisms (Cox, 1997); partition the parameters gamma = theta, eta into a subset of interest theta and other "nuisance parameters" eta necessary to complete the probability distribution (Cox and Hinkley, 1974); develops methods of inference about the scientific quantities that depend as little as possible upon the nuisance parameters (Barndorff-Nielsen and Cox, 1989); and thinks critically about the appropriate conditional distribution on which to base infrences.

We briefly review exciting biomedical and public health challenges that are capable of driving statistical developments in the next decade. We discuss the statistical models and model-based inferences central to the CM approach, contrasting them with computationally-intensive strategies for prediction and inference advocated by Breiman and others (e.g. Breiman, 2001) and to more traditional design-based methods of inference (Fisher, 1935). We discuss the hierarchical (multi-level) model as an example of the future challanges and opportunities for

model-based inference. We then consider the role of conditional inference, a second key element of the CM. Recent examples from genetics are used to illustrate these ideas. Finally, the paper examines causal inference and statistical computing, two other topics we believe will be central to biostatistics research and practice in the coming decade. Throughout the paper, we attempt to indicate how DRC's work and the "Cox Model" have set a standard of excellence to which all can aspire in the future.

A Cox Model for Biostatistics of the Future

Scott L. Zeger[1], Peter J. Diggle[2,1], and Kung-Yee Liang[1]

The Johns Hopkins University Department of Biostatistics[1]

The University of Lancaster Department of Mathematics and Statistics[2]

1 Introduction

Professor Sir David R. Cox (DRC) is widely acknowledged as among the most important scientists of the second half of the twentieth century. He inherited the mantle of statistical science from Pearson and Fisher, advanced their ideas, and translated statistical theory into practice so as to forever change the application of statistics in many fields, but especially biology and medicine. The logistic and proportional hazards models he substantially developed, are arguably among the most influential biostatistical methods in current practice.

The goal of this paper is to look forward, say over the period from DRC's 80th to 90th birthdays, to speculate about the future of biostatistics. We have attempted to draw lessons from DRCs contributions as we look ahead.

As a BBC web-site recently published: "trying to predict the future is a mug's game". So we take the speculations below with a grain of salt as should you. Nevertheless, an 80th birthday is an occasion for celebration and some speculation.

Biostatistics is a term with slightly different meanings in Europe and the United States. Here we take it to refer to statistical science as it is generated from and applied to studies of human health and disease. We distinguish it from the broader term *biometrics* that refers to the interface of statistics and all biological sciences including for example agricultural science and from *medical statistics* that refers more narrowly to clinical than biomedical sciences.

The title of this paper refers to Cox's model of biostatistics. A cursory literature search of recent papers will find at least three frequently referenced Cox models (Box-Cox transformations; the Cox process in point process theory; and the Cox proportional hazards model). Within the confines of this chapter, we refer not to these specific models, but rather to an approach to statistical science that Professor Cox has forcefully advanced. We believe it is a model for the continued advancement of empirical science, at least for the period about which we speculate. This Cox Model (CM):

- formulates scientific questions or quantities in terms of parameters γ in probability models $f(y;\gamma)$ that represent in a parsimonious fashion, the underlying scientific mechanisms (Cox,1997);
- partition the parameters $\gamma = (\theta, \eta)$ into a subset of interest θ and other "nuisance parameters" η necessary to complete the probability distribution (Cox and Hinkley, 1974);

- develops methods of inference about the scientific quantities that depend as little as possible upon the nuisance parameters (Barndorff-Nielsen and Cox, 1989); and
- thinks critically about the appropriate conditional distribution on which to base inferences.

This paper starts with a brief review of what we believe are exciting biomedical and public health challenges that are capable of driving statistical developments in the next decade. We discuss how they may influence biostatistical research and practice. Section 3 then discusses the statistical models and inferences that are central to the CM approach. We contrast the CM with computationally-intensive strategies for prediction and inference advocated by Breiman and others (e.g. Breiman, 2001) and to more traditional design-based methods of inference (Fisher, 1935). We discuss the *hierarchical (multi-level) model* as an example of the future challenges and opportunities for model-based inference. In Section 3, we also discuss the role of increasingly "complex" models in science and statistics. A key question is how to quantify the uncertainty in target parameters, incorporating more than the usual sampling variability. In Section 4, we discuss conditional inference, a second key element of the CM. Recent examples from genetics are used to illustrate these issues.

Sections 5 and 6 discuss causal inference and statistical computing, two other topics we believe will be central to biostatistics research and practice in the coming decade. Throughout the paper, we attempt to indicate how DRC's work and the "Cox Model" have set a standard of excellence to which all can aspire in the future.

2 Biomedical and public health opportunities

2.1 The last 50 years

If we start the look forward with a brief look back at the past, what is clear is that we live in a revolutionary period for biological science, particularly for biomedicine and public health. It has been only 50 years since Watson and Crick elucidated the basic paradigm for molecular biology (Watson and Crick, 1953). Since then, the parallel advances in bio- and computer technology have produced a panoply of new measurements that are now driving statistical innovation.

In biomedical laboratories, we now routinely measure DNA sequences, gene and protein expression levels, protein structures and interactions. The new biotechnologies make it possible to control systems by: breeding genetically-designed laboratory animals, adding or removing putative genes, fabricating molecules, inserting human genes in other organisms and so on. Biomedical investigators have remarkable new tools to measure and to control systems permitting more informative studies of the mechanisms underlying health and disease.

The same 50 years has given rise to equally powerful technologies for population medical research. The modern randomized controlled trial has its roots in 1947 with the MRC Streptomycin in Tuberculosis Trial. There are now generally-accepted protocols for human experimentation such as the Nuremberg Code (Kious, 2001) and the controlled clinical trial

has become the single leading tool of clinical research. The advent of systematic testing of preventive and therapeutic therapies using controlled trials is arguably the most important discovery in clinical medicine during this period.

In 1952, Sirs Richard Doll and Bradford Hill also changed the practice of epidemiology by introducing the case-control study to investigate the association of smoking and lung cancer. Cornfield (1951) finished the foundation by developing the statistical framework for the analysis of case-control studies. Like the randomized trial in clinical research, the case-control study and derivative designs are now essential tools for every clinical and public health researcher.

The progress in biostatistics over the last few decades has reflected the advances in medical and public health research. The productivity in the sub-fields of statistical genetics, survival analysis, generalized linear models, epidemiologic statistics and longitudinal data analysis evidences the opportunities that medical and public health breakthroughs of the last half century have created for statistical science, and vice versa.

2.2 The next decade

Speculation about the near future of biostatistics best starts with consideration of the emerging opportunities in biomedicine and public health. The US NIH "Roadmap"(Zerhouni, 2003) and the UK MRC strategic plan (MRC, 2001) present similar visions for the future. Francis Collins (2003) provides an excellent review article of how genomics will change biomedical research and practice.

There is consensus that, while the 20th century has produced lists of genes and proteins, the priority for the next decade is to determine their functions. The goals are to identify and understand: the gene-protein networks central to normal biology; aberrations and interactions with the environment that produce disease; and corrective interventions that can prevent or treat disease. The MRC plan states: "The focus will shift to questions of protein structure, function and interactions, developments of physiological pathways and systems biology (MRC, 2001)." The NIH Roadmap identifies "Building Blocks, Biological Pathways and Networks" as the first of its five priorities (Zerhouni, 2003).

In the next decade, exhaustive gene lists will be refined for organisms at all levels. Once sequenced for a species, biologists will characterize the covariation across a population in key DNA markers and single nucleotide polymorphisms (SNPs). Parsimonious summaries of these patterns of association will be needed.

Associations (many of them false) of disease occurrence with SNPs, gene and protein expression levels will be discovered. Some associations will lead to the identification of biochemical pathways for normal biology and mistakes that cause disease. We should not be surprised if some aspect of the basic paradigm of molecular biology is supplanted or dramatically revised. Epigenetics (e.g. Feinberg, 2001) is an example of a major new idea that could radically change the nature of biomedical research in the next decade.

The biotechnology industry will continue to produce more powerful laboratory methods to quantify genes and their RNA and protein expressions. There will be a key role for statistical design and evaluation of the measurement process itself. For example, the early gene expression measurements by Affymetrix, a leading manufacturer of micro-arrays, have been

substantially improved by careful statistical analysis. Affymetrix quantified expression by comparing the binding of 25 base-pair oligonucleotides to the target sequence RNA ("perfect match") to a background rate of binding to a sequence with the middle base-pair changed. The idea was sensible: to correct for non-specific binding of RNAs with sequence similar to the target. But Irizarry and colleagues (2003) have shown there is an enormous price in variance to pay for the small improvement in bias when doing this ad-hoc baseline correction. Similar opportunities certainly exist in quantifying protein expression, protein-protein interactions and other fundamental measurements.

Measurement is also central to clinical research. For example, diagnosis and severity assessment for many psychiatric disorders such as depression and schizophrenia are based upon symptoms checklists. Quality of life or health status is also measured via longer lists of items. Professor Cox (1997) has discussed the opportunities for improved measurement in these situations where the basic data is a multivariate vector of item responses.

Having quantified tens of thousands of gene or protein expression levels, the natural question is which values best predict the incidence or progression of disease or the efficacy of a novel treatment. We should expect many more studies like the one by van de Vijver, et al. (2002) who followed 295 patients with primary breast carcinoma for time to metastases. Gene expression levels for 25,000 genes were used to develop a "risk profile" based upon roughly 70 genes. The profile was a strong discriminator of persons with early versus late progression. While cross-validation was used to estimate the prediction error, the number of models with only main effects and two-gene interactions is too numerous to contemplate. The potential for false discoveries is enormous. Effective methods for searching model space, informed by current biological knowledge are needed. Ruczinski and colleagues (2004) developed logic regression, one an exciting approach, to address exactly this problem of contending with high dimensional predictor space when predicting protein structure from amino acid sequences.

Similar problems are common in image analysis. See for example, a recent paper by Chetelat and Baron (2003)who predict onset of Alzheimers disease using data from functional magnetic resonance images with activation intensities at on the order of 10^5 spatial locations (voxels) in the brain.

The study of complex interactions or "systems biology" is an emerging theme in biomedical research. Harvard University Medical School has recently created a new Department of Systems Biology dedicated to "*bring together a new community of researchers - biologists, physicists, engineers, computer scientists, mathematicians and chemists* (note - no reference to statistics) - *to work towards a molecular description of complex cellular and multicellular systems, with an emphasis on quantitative measurement and mathematical and computational models.*"(Harvard University, 2004). An illustration is the work by von Mering, et al. (2003) who compares dendograms of protein-protein interactions across species to identify preserved associations and thereby to infer critical functions of protein systems.

The next decade of clinical and public health research will also be driven by the twin genomics and computing revolutions. In epidemiology, the study of gene and environmental interactions will remain center stage. For example, Caspi, et al. (2003) reported results of a prospective study that identified a polymorphism in the promoter region of the serotonin transporter (5-HT T) gene. Persons with one or two copies of the short allele exhibited more depression in response to stressful life events than individuals homozygous for the long

allele. Similar research on gene-environment interactions will be central to cardiovascular and cancer epidemiology.

The amount of data and computing power will make possible a new level of public health surveillance, as well (e.g. Dominici, et al., 2003). For example, in their most recent work, Dominici and colleagues have assembled public databases comprising: administrative medical records with diagnostic codes for the more than 40 million adults in the US Medicare population, hourly air pollution time series for 6 major pollutants from thousands of EPA monitors, hourly air and dew-point temperature from major weather center, and US Census data at the postal code level on socioeconomic status (SES), education, housing and other factors that influence health. They are using these data to quantify the deaths, cases of disease and medical costs attributable to air pollution and its possible interaction with SES and other community-level factors.

In summary, the twin bio- and computer technologies have created new windows into the workings of biological systems in the form of new measurements that are inherently high-dimensional. We have now assembled long lists of genes, proteins and other biological constituents. We are in the process of uncovering the interactions and networks that define normal and abnormal biology. These problems and similar ones in epidemiology and public health are likely to drive the development of statistical reasoning and methods for a decade or more.

3 Statistical models and statistical inference

The word "model" is over-used (see title of this paper). Within the statistics discipline, it can cover any mathematical representation of the genesis of a set of data, y, but this is such a general statement that it is essentially worthless.

From a scientific perspective, a useful distinction is between models which can be deduced from known or hypothesized scientific mechanisms underlying the generation of the data, and models which are simply empirical descriptions. Consider for example the homogeneous Poisson process. As a model for the time-sequence of positron emissions from a radio-active specimen, it has a well understood mechanistic justification; as a model for the time-sequence of seizures experienced by an epileptic patient, its value rests on empirical validation against observed data. Besag (1974) argued, cogently in our view but ultimately unsuccessfully, to use "scheme" rather than "model" when referring to an empirical model.

From a statistical perspective, the key modelling idea is that data, y, constitute a realization of a random variable, Y, in other words our models are probabilistic models. We then need to ask, "where does the randomness come from?" R A Fisher's answer, in the context of the analysis of agricultural field trials, was that the randomness was induced by the use of randomization in the allocation of experimental treatments to plots, leading to inferences which were essentially non-parametric. A second answer, which underlies most mainstream statistics teaching and practice circa 2004, is that the randomness is an inherent property of the process which produces each individual datum. Under this paradigm, the vector of plot-yields y in a field trial are modelled as a realization of a multivariate Gaussian random variable Y with appropriately structured mean vector and variance matrix, and inferences

are based on calculated probabilities within this assumed distributional model.

These two approaches to inference can be summarized as *design-based* and *model-based*, respectively. The connection between the two, in the context of the general linear model, was made in Kempthorne (1952). It is ironic that Fisher's analysis of variance methods are now taught almost exclusively under multivariate Gaussian distributional assumptions. The collection of papers published in the December 2003 issue of the *International Journal of Epidemiology* includes a fascinating discussion of the contrasting arguments advanced by Fisher and by Bradford Hill for the benefits of randomization in the contexts of agricultural field experiments and clinical trials, respectively (Chalmers, 2003; Armitage, 2003; Doll, 2003; Marks, 2003; Bodmer, 2003; Milton, 2003).

Most statisticians would accept that model-based inference is appropriate in conjunction with a mechanistic model. How one should make inferences when using an empirical model is less clear. In practice, design-based inference copes most easily with problems whose structure is relatively simple. The development of progressively more complex data-structures involving multi-dimensional measurements, together with temporal and/or spatial dependence, has therefore led to an increased focus on model-based inference, which then raises the awkward question of how sensitive our inferences might be to the assumed model. To quote Coombs (1964), "We buy information with assumptions," with the unspoken corollary that what is bought may or may not give good value.

The model-based approach lends itself equally well to classical likelihood-based or Bayesian modes of inference. The last decade has seen a strong shift towards the use of Bayesian inference in many branches of statistics including biostatistics. Whether this is because statisticians have been convinced by the philosophical arguments, or because the computing revolution and the development of Monte Carlo methods of inference has apparently reversed the historical computational advantage which classical methods held over their Bayesian rivals, is less clear. For a thought-provoking discussion of likelihood-based inference, including a critique of Bayesian inference, see Royall (1997).

In recent years, a third broad approach to inference has developed, which we might call algorithm-based. This sits naturally at the interface between statistics and computing, and can be considered as a computationally intensive version of John Tukey's early ideas concerning exploratory data analysis (Tukey, 1977), in which the traditional emphasis on formal inference is replaced by a less formal search for structure in large, multi-dimensional datasets. Examples include a wide range of non-parametric smoothing methods, many originating from the Stanford and Berkeley Departments of Statistics (e.g. Breiman, Friedman, Olshen and Stone, 1984; Friedman, 1991; Breiman and Friedman, 1997), and classification methods based on neural networks (Ripley, 1994). The cited examples illustrate the key contributions which statisticians continue to make to algorithm-based inference. Nevertheless, its emergence also represents a challenge to our discipline in that algorithm-based methods for data analysis are often presented under the computer-science oriented banner of "informatics," rather than as an integral part of modern statistical methodology.

3.1 Hierarchical models and model-based inference

A recurrent theme during the last fifty years of statistical research has been the unification of previously separate methods designed for particular problems. Perhaps the most important to date has been Nelder and Wedderburn's (1972) unification of a whole raft of methods including analysis of variance, regression, log-linear models for contingency tables, logit and probit models for binary data, within the class of generalized linear models (GLM's, Nelder and Wedderburn, 1972) For a comprehensive account, see McCullagh and Nelder (1989). The GLM class essentially represents the state of the art for modelling the relationship between a set of mutually independent univariate responses and associated vectors of explanatory variables.

Hierarchical models represent a comparable unification of models for dependent responses. Their defining feature is that observed responses $Y_i : i = 1, ..., n$ are modelled conditionally on one or more layers of latent random variables which are usually of scientific interest but are not directly observable. Using the notation $[\cdot]$ to mean "the distribution of," the simplest hierarchical model specification would take the form $[Y, S] = [Y|S][S]$, where S denotes a latent random vector. More elaborate examples are discussed below.

In contrast to classical GLM's, it is hard to identify a single, seminal paper on hierarchical models. Rather, the essential idea was proposed independently in different settings, and these separate contributions only later brought together under a common framework.

The Kalman filter (Kalman, 1960) is an early example of a hierarchical model for time series data. In its basic form, it models an observed univariate time-series $Y_t : t = 1, 2, ...$ in relation to an unobserved "state variable" $S_t : t = 1, 2, ...$ by combining a linear stochastic process model for S_t (the "state equation"), with a linear regression for Y_t conditional on S_t (the "observation equation") and uses the model to derive a recursive algorithm for computation of the minimum mean square error predictor for S_t, given the observed data $Y_s : s \leq t$. This results in the standard hierarchical factorization $[Y, S] = [Y|S][S]$. Within the time series literature, models of this kind became known as state-space models, or dynamic time series models (Harvey, 1989; West and Harrison, 1997).

Even earlier, Cox (1955) introduced a class of hierarchical models, now known as Cox processes, for point process data. A Cox process is an inhomogeneous Poisson process whose intensity, $\Lambda(x)$ is itself the realization of an unobserved, non-negative valued stochastic process. For example, a log-Gaussian Cox process (Moller, Syversveen and Waagepetersen, 1998) takes $\Lambda(x) = \exp\{S(x)\}$ where $S(x)$ is a Gaussian process. Models of this kind are likely to become increasingly important for modelling spatial point process data in which the local intensity of points is modulated by a combination of observed and unobserved environmental factors, and will therefore play a key role in environmental epidemiology. Although the Gaussian assumption is a strong one, it is attractive as an empirical model because of its relative tractability and the ease with which it can incorporate adjustments for observed, spatially referenced explanatory variables. Incidentally, Cox processes bear the same relationship to Poisson processes as do frailty models for survival data to the classic Cox proportional hazards model (Cox, 1972). In each case, an unobserved stochastic effect is introduced into the model to account for unexplained variation over and above the variation compatible with a Poisson point process model for the locations of events in space or time,

respectively.

Lindley and Smith (1972) were concerned not so much with developing new models as with re-interpreting the classical linear model from a Bayesian perspective. Hence, their use of the hierarchical model structure was to consider the linear regression relationships as applying conditional on parameter values which were themselves modelled as random variables by the specification of a prior distribution. The same formalism is used in classical random effects linear models.

Motivated primarily by social science applications, Goldstein (1986) used hierarchical models to explain how components of random variation at different levels in a hierarchy contribute to the total variation in an observed response. For example, variation in educational attainment between children might be decomposed into contributions arising from components of variation between schools, between classes within schools and between children within classes.

Finally, note that one way to extend Nelder and Wedderburn's GLM class for independent responses Y_i is to introduce a latent stochastic process into the linear predictor, so defining a generalized linear mixed model (GLMM). Breslow and Clayton (1993) review work on GLMM's up to that date, and discuss approximate methods of inference which have been largely superceded by subsequent developments in classical or Bayesian Monte Carlo methods (Clayton, 1996).

Hierarchical models of considerably greater complexity than the examples given above are now used routinely in a wide range of applications, but especially as models for spatial or longitudinal data. Spatial examples include models for smoothing maps of empirical measures of disease risk (Clayton and Kaldor, 1987; Besag, York and Molié, 1991), or for geostatistical interpolation under non-Gaussian distributional assumptions (Diggle, Moyeed and Tawn, 1998). Longitudinal examples include joint modelling of longitudinal measurements and time-to-event data in clinical trials (Wulfsohn and Tsiatis, 1997), in which stochastic dependence between the measurement process and the hazard function for the time-to-event process are induced by their shared link to a latent Gaussian process S.

Despite the widespread, and widely accepted, application of hierarchical models, their routine use is not without its attendant dangers. More research is needed on some fundamental issues concerning model formulation, computation and inference.

With respect to model formulation, a key issue is the balance between simplicity and complexity. Simple models tend to be well-identified. Complex models promise greater realism, but only if they are scientifically based. When models are empirical, simplicity (consistent with achievement of scientific goals) is a virtue. A challenge for the next decade is to develop a theory of model choice which appropriately combines empirical and subject-matter knowledge. In our opinion, it is doubtful whether such a theory could ever be reduced to a mathematical formalism, and we are skeptical about the role of automatic model selection algorithms in scientific research.

With respect to computation, we need a better understanding of when off-the-shelf Markov chain Monte Carlo algorithms are and are not reliable, and better tools for judging their convergence. At least for the near future, we also need to recognize that many applications will continue to need mechanistic models and tailored fitting algorithms, using the combined skills of subject-matter scientists, applied statisticians and probabilists.

Perhaps most challenging of all are questions concerning inference for hierarchical models. For Bayesian inference, how should we choose joint priors in multi-parameter models? How can we recognize poorly identified sub-spaces of the parameter space? And, whether or not we choose to be Bayesian, when might unverifiable modelling assumptions unduly influence our conclusions? In a sense, all of these questions have straightforward formal answers, but dealing with them thoroughly in practice is impossible because of the multiplicity of cases which need to be examined. Furthermore, there seems to be an increasing divergence between Bayesian philosophy, in which the notion of a "good" or "bad" prior has no meaning, and Bayesian practice, in which there appears to be a decreasing emphasis on topics such as prior elicitation through expert opinion, and a *de facto* shift to using computationally convenient families of prior within which location and/or scale parameters act as 'tuning constants" somewhat analogous to multi-dimensional band-width choices in non-parametric smoothing. The coming decade will provide better tools for the Bayesian data analyst to quantify the relative amounts of "information" provided by data and by prior specification, whether at the level of model choice or parameter values within a chosen model, and a better understanding of how these influence estimates and substantive findings.

3.2 Sources of uncertainty

The end-product of a piece of statistical work is the result of a long sequence of decision-making in the face of uncertainty. An over-simplified representation of this process is:

1. What is the *scientific* question?
2. What experimental *design* will best enable the question to be answered?
3. What statistical *model* will incorporate an answer to the question whilst dealing adequately with extraneous factors?
4. How should we *analyze* the resulting data using this model and interpret the estimates and measures of their uncertainty?

Currently, statistical theory deals explicitly and very thoroughly with the last of these, through the theory of statistical inference. In essence, inference allows us to say: "this is what we have observed, but we might have observed something different" and to moderate our conclusions accordingly. To a limited extent, techniques such as Bayesian model-averaging, or classical nesting of the model of interest within a richer class of possible models, allow us to take account of uncertainty at the level of model formulation, but here the methodology is less well developed, and less widely accepted. Rather, it is generally accepted that model formulation is, at least in part, a subjective process and as such not amenable to formal quantification. All statisticians would surely agree that careful attention to design is of vital importance, yet our impression is that in the formal training of statistical graduates, courses on design typically occupy a very small fraction of the syllabus by comparison with courses on inference, modelling and, increasingly, computation. We predict that some relatively old ideas in experimental design, such as the construction of efficient incomplete block designs, will soon enjoy a revival in their importance under the perhaps surprising stimulus of

bioinformatics, specifically gene expression microarray data. As discussed above, microarray data are in one sense spectacularly high-dimensional, but a typical microarray experiment is cost-limited to only a small number of arrays and these, rather than the many thousands of individual gene expression levels, are the fundamental experimental units at which design questions must be addressed. See, for example, Kerr and Churchill (2001), Yang and Speed (2002) and Glonek and Solomon (2004).

Finally, the best comment we can offer on the importance of addressing the right question is to quote DRC in an interview for the *International Statistical Institute Newsletter* (volume 28, issue 1, page 9). In answer to the question of what advice he would offer to the head of a new university department of statistics, he included "the importance of making contact with the best research workers in other subjects and aiming over a period of time to establish genuine involvement and collaboration in their activities."

3.3 Model complexity

The new measurements enabled by biotechnologies will certainly lead some to increasingly complex descriptions of biological systems. For example, Figure 1 in the paper by H. Jeong, et al (Nature, 2001) represents proteinprotein interactions for 1,870 proteins shown as nodes, connected by 2,240 direct physical interactions as determined by systematic two hybrid analyses for the yeast (Saccharomyces cerevisiae) proteome. They represent by the color of a node, the phenotypic effect of deleting the gene that encodes its corresponding protein (red, lethal; green, nonlethal; orange, slow growth; yellow, unknown). The analysis argues that proteins with more connections are more essential to yeast survival.

Biostatisticians have long recognized the trade-off of biological realism against parsimony in their models estimated from a finite set of observations. As the richness of data sets increases, more complex models will surely be contemplated. An important question for the next decade will be where to find the new balance between realism and parsimony.

DRC (1997) offerred an hierarchy of probability models according to their scientific basis that extends the two categories we discussed above:

1. purely empirical models
2. "toy models"
3. intermediate models
4. quasi-realistic models.

DRC's Types 1 and 4 correspond essentially to what we have called "empirical" and "mechanistic" models,respectively. Moving down the list, the scientific complexities are represented to an increasing degree. It seems plausible for at least two reasons that biostatistical models will strive for increasing biological content over the coming decade.

First, parameters in more realistic models are inherently more useful and interesting to scientist. The enormous success of the proportional hazards models (Cox, 1972) is partly because its parameters are relative risks with simple biological interpretations. The hierarchical models discussed above are especially popular among social scientists who study the effects

of covariates at the individual and community levels because their regression parameters have causal interpretations.

Second, imposing scientific structure is a partial solution to the dimensionality of our outcome measures. For example, in early statistical analysis with DNA microarrays, gene expressions were most often treated as exchangeable (e.g. Hastie, et al., 2000). But there is substantial scientific knowledge that can be used to arrange genes into prior clusters, for example because they encode proteins involved in a common metabolic pathway. Bouton, et al. (2003) have developed DRAGON, a software tool that queries genomic databases and creates such gene groupings. Biological knowledge can be used to reduce the number of variables from the order of 10^4 genes to 10^2 classes genes.

As is well known to statisticians, there is a price to pay for increasing "realism" in the face of limited data. Often, key parameters are difficult to identify. For example, Heagerty and Zeger (2000) have shown how parameter estimates in conditional random effects models (hierarchical models) depend quite strongly on difficult-to-verify assumptions about the variances of the unobserved random effects.

4 Conditional Inference

In 1958, Professor Cox published two seminal papers. The first, in the *Annals of Mathematical Statistics* addressed the statistical principle of *conditioning*. The second, in *JRSS, B*, applied this idea to the analysis of binary responses, in particular, using the logistic regression model. In the fifty years since their publication, these two articles established conditional inference as a cornerstone of the Cox Model (CM). In this section, we discuss three reasons why conditioning has had so much impact to date and consider its role in the future of biostatistics. As above, we will represent the full set of parameters by $\gamma = (\theta, \eta)$ where θ represents the subset of scientific interest and η represents the nuisance parameters.

4.1 Inducing the proper probability distribution

One fundamental argument in favor of conditioning is to insure that the probability distribution used for inference about θ is consistent with the data in hand (Cox, 1958a). Specifically, it calls for use of a probability distribution for the observed data conditional on ancillary statistics defined as those whose distributions are independent of θ. The idea is to focus on models for random variables that are directly relevant to θ without investing time in modelling less relevant ones. A simple example of an ancillary statistic is the sample size because, in most situations, the probability mechanism for an observation does not depend on the choice of sample size. Because the likelihood function and statistical uncertainty about θ does depend upon the sample size, this dependence must be made explicit as happens through conditioning.

The most compelling argument for conditional inference was a toy example offered by DRC in his 1958 paper (Cox, 1958a). Here, the parameter of interest is the mean value for a biological measurement. A coin is tossed to determine which one of two competing laboratories will be selected to run the biological assay. One lab has high precision (low

variance); the other low precision. The conditioning principle demands that inferences about the unknown mean be drawn conditional on the lab that actually made the measurement. That another lab, with very different precision, might have made the measurement is ancillary information.

4.2 Elimination of nuisance parameters

The closely related second objective of conditioning is to focus inferences on the parameters of interest θ, minimizing the impact of assumptions about the parts of the probability mechanism indexed by nuisance parameters η. The parameters η are necessary to jointly specify the mechanism but are not of intrinsic interest to investigators. Examples are common in biomedical studies. In case-control studies of disease etiology, θ is the odds ratio relating a risk factor to disease, and η is the probability of exposure for the control group. In the proportional hazards model (Cox, 1972), θ is the log of relative hazards comparing individuals with covariate value $X = x$ versus those with $X = 0$. The nuisance parameters comprise the hazard function for the reference group, which is left unspecified.

An issue at hand, as stated in Section 1, is how to develop methods of inference about the parameters of interest that depend as little as possible on the nuisance parameters. This is an important issue because it is well documented that the quality of inference for θ depends critically on how these two sets of parameters are intertwined with each others (e.g., Cox and Reid, 1987, Liang and Zeger, 1995). Neyman and Scott (1948) demonstrated that the maximum likelihood estimates for θ can be inconsistent when the dimension of η increases with the sample size. This is exactly the situation epidemiologists face in the matched case-control design, which is commonly used when confounding variables are difficult to measure. Examples include controlling for a case's neighborhood or family environment or genes. In this situation, the number of nuisance parameters is roughly equal to the number of cases. Cox (1958b) showed how to eliminate the nuisance parameters by conditioning on sufficient statistics for the nuisance parameters. The modern version of this approach, conditional logistic regression, is in routine use.

Conditional logistic regression has had important application in genetic linkage where the primary goal is to find the chromosomal region of susceptibility genes for a disease using phenotypic and genetic data from family members. One design that is attractive because of its simplicity is known as the case-parent trio design (Spielman et al., 1993). Blood samples are drawn from diseased offspring and both of their parents. Each parent-child dyad forms a matched pair. The "case" is defined to be the target allele of a candidate gene that was transmitted from the parent to the offspring. The "control" is the allele that was not transmitted to the offspring. Thus for each trio, two matched pairs are created and one can test the "no linkage " hypothesis by testing whether the target allele is preferentially transmitted to diseased individuals. The connection between the conventional one-to-one matched case-control design and the case-parent trio design enables investigators to address important questions that the new technologies for genotyping large numbers of markers have made possible. A key question for the coming decade is how to quantify evidence that the candidate gene interacts with environmental variables to cause disease. Here, the conditional approach to inference is likely to be essential.

4.3 Increasing efficiency

In the coming decade, genetic research will continue to focus on complex diseases that involve many genes and more complicated gene-environment interactions. Traditional genetic linkage methods are designed to locate one susceptibility gene at a time. Unless one has large numbers of homogeneous families, the power to detect the putative gene is small. Recently, Nancy Cox and colleagues (Cox, et al., 1999) applied statistical conditioning to address this issue in a search for genes associated with Type-I diabetes. Taking advantage of a previous linkage finding on chromosome 2, they carried out a genome-wide scan on the remaining chromosomes by conditioning on each family's "linkage score" from chromosome 2. Through this process, strong linkage evidence on chromosome 15 was discovered. This constitutes a sharp contrast to the previous analysis without conditioning on chromosome 15, which found no evidence of linkage. This approach was further developed for studies that use identical-by-descent or IBD sharing by Liang et al. (2001). They have similar findings in an asthma linkage study data reported by Wjst et al. (1999).

Examples above illustrate that proper conditioning can help to address the challenging issue of gene-gene interaction by increasing the statistical power in locating susceptibility genes. A limitation of this approach is that it is conditional on one region at a time. With thousands of SNPs, an important research question is how best to extend the conditional approach to higher-dimensional cases. Similar opportunities exist when searching for associations between gene or protein expression levels and disease outcomes or with voxel-specific image intensities and disease outcome or progression. What is certain, is that the CM approach to inference provides a way forward.

5 Determining cause

Most biostatisticians would agree that quantification of the evidence relevant to determining cause and effect is an essential responsibility of statistical science. For example, the practice of medicine is predicated upon the discovery of treatments that cause an improvement in health by reversing a disease process or by alleviating its symptoms.

The meaning of the term "cause is generally accepted among statisticians to be close to the Oxford English Dictionary definition: "That which produces an effect (OED, 1989)." But how this idea is best implemented in empirical research has been, is today and will be for the next decade, the subject of important work.

In the biomedical sciences, Koch's postulates were among the earliest attempts to implement causal inference in the modern era. Robert Koch, a nineteenth century physician who discovered the anthrax and tubercle bacilli, proposed four criterion by which an infectious agent would be established as the cause of an illness. While the Koch postulates are inconsistent with current biological theories, they remain an important example of the integration of biological reasoning with experimental data to establish cause. Sir Bradford Hill (Hill, 1965), in addition to establishing the modern randomized clinical trial and case-control study, posited nine attributes when determining whether an association is causal: strength, consistency, specificity, temporality, biologic gradient, plausibility, coherence, experimental

evidence, and analogy. Like Koch, Hill proposed a qualitative, inter-disciplinary process to establish cause. His ideas grew out of the debate in the 1950s about whether smoking caused lung cancer. His attributes are relied upon today in text books (e.g. Rothman and Greenland, 1998) and court rooms.

Biological reasoning about mechanisms is central to the Koch and Hill approaches to causal inference. Is there a biological mechanism that explains the statistical association? Are there alternate mechanisms that would give rise to the same observations? Statistical evidence from scientific studies is central. But the evidence is assembled through a qualitative process in which current biological and medical knowledge plays a prominent role.

A second line of causal reasoning is based upon the formal use of "counterfactuals". The causal effect of a treatment for a person is defined as the difference in outcome between two other-wise identical worlds: one in which the treatment was taken, the other where it was not. The counterfactual definition of "cause" dates back to the eighteenth century or earlier (Rothman and Greenland, 1998) and is prevalent in economics and more recently statistical sciences. In formal counterfactual analysis, the pair of responses $Y_i(t = 1)$ and $Y_i(t = 0)$ for unit i with $(t = 1)$ and without $(t = 0)$ "treatment" (or risk factor), is directly modelled.

Statistical methods for causal inference have frequented the leading journals for more than two decades (e.g. Rubin, 1974; Holland, 1986). The formal use of counterfactuals in probability models is particularly effective in randomized clinical trials and other similar studies where the assignment of treatments occurs according to a known or well-approximated mechanism. Counterfactual models allow us to study the effects of departures from this leading case, for example to study the effects of drop-outs (e.g. Scharfstein, Rotnitzky, Robins, 1999) or failure to comply (Frangakis and Rubin, 1999).

It has also become an central idea in epidemiologic research where randomized trials are not possible for the study of most risk factors. See for example papers by Greenland, Robins and Pearl (1999) and Kaufman, Kaufman, and Poole (2003) and references therein.

The formal analysis of counterfactual variables is now commonplace in problems far from the randomized trial. For example, Don Rubin, an early exponent and key researcher on formal causal inference, is also the statistical expert for the tobacco industry in their suits against the states and the United States Justice Department (Rubin, 2002). He has testified that the medical costs caused by smoking in a population are properly determined by comparing two worlds: one in which smoking occurred, and the other in which it did not. To quantify the effects of the tobacco companies alleged fraudulent behavior, he calls for a comparison of two worlds: one with, the other absent those behaviors that took place over several decades in thousands or perhaps millions of discrete acts (Rubin, 2001).

It is hard to argue in the abstract with these causal targets for inference whether in a randomized controlled trial, an epidemiologic study or an assessment of a complex industrial behavior. But an important question is what role statistical models will have in causal inference. Should statistical models quantify evidence for key components such as relative risks which are then combined through a qualitative process with background knowledge and theory to establish cause? The process of determining that smoking causes lung cancer, cardiovascular disease and premature death is an example of one such process. Or should formal causal models be relied upon to organize the evidence, prior knowledge and beliefs about mechanisms, alternative explanations and other relevant biological and medical factors?

DRC (1986) pointed to the multi-layer process of establishing cause in his discussion of Holland (1986) saying: "Is not the reason that one expects turning a light switch to have the result it does not just direct empirical observation but a subtle and deep web of observations and ideas - the practice of electrical engineering, the theory of electrical engineering, various ideas in classical physics, summarized in particular, in the Maxwell's equations, and underneath that even ideas of unified field theory? "

The statistical science approach to causal inference will continue to be refined in the coming decades. We will learn how to more accurately quantify the contributions to inference from statistical evidence versus from difficult-to-verify assumptions, particularly when making inference requires information about a complex world that did not occur. We can envision the evolution of an hierarchy of "causal inference". The term *causal estimation* will be reserved for the randomized studies and for modest departures from them. Inferences about causal targets that critically depend on unverifiable assumptions will be termed *causal extrapolations* while inferences about the causal effects of the behaviors of the tobacco industry might best be called *causal speculations*. These qualitative distinctions represent points along a continuum. How best to quantify the respective roles of statistical evidence and model assumptions in causal inference is, in our opinion, an important topic for further research.

6 Statistical Computing

The explosion of information technologies during the last few decades has changed forever the way in which empirical science is conducted. The collection, management and analysis of enormous data sets has become routine. The standard responses in biostatistical research have radically changed. Journal articles 20 years ago dealt mainly with binary, count or continuous univariate response variables. Generalized linear models (McCullagh and Nelder, 1989) was a breakthrough because it unified regression methods for the most common univariate outcomes.

But in today's studies, the response is commonly of very high dimension: an image with a million discrete pixels; a micro-array with a continuous measure of messenger RNA binding for each of 30,000 genes; a schizophrenia symptoms questionnaire with 30 discrete items. The intrinsically multivariate nature of data has been made possible by fast computers with inexpensive storage. The emerging fields of computational biology, bioinformatics and data mining are attempts to take advantage of the exponential growth in digitally-recorded information and in the computing power to deal with it.

It is safe to predict that biostatistics research in the coming decades will become increasingly interwoven with computer science and information technologies. We have already discussed the role of computationally intensive algorithms and resampling-based inference. A second major change is the near instantaneous interconnection among scientists that the internet makes possible. International research groups now work closely together on "big science" projects. The development of Linux and the "Human Genome Project" are leading examples.

In statistics, internet groups, operating in the Linux model, have created R, a statistical

computing and graphics language that now dominates most centers of statistical teaching and research. It is organized to permit continuous expansion as unconnected researchers produce R-packages for specialized methods.

In bioinformatics, the package "Bioconductor" is a leading tool for managing and analyzing gene-expression array data (Gentleman and Carey, 2003). Bioconductor is created and managed by a group of investigators from around the world. They have implemented protocols for extending the package themselves and for accepting extensions offered by others. In just three years, it has grown to have substantial influence on the practice of biostatistics in molecular genetics laboratories everywhere. Bioconductor demonstrates how the Internet can re-shape biostatistical research to involve larger teams of statisticians, computer scientists and biologists, loosely organized to achieve a common goal. A cautionary note is that an understandable focus on the computational challenges of bioinformatics brings with it a danger that the continuing importance of fundamental statistical ideas such as efficient design of incomplete block experiments may be forgotten.

The specialty of bioinformatics also points towards another trend. Because genetic information is collected and shared through web-based data systems, methods of analysis are also shared in the same way. Traditionally in statistics, when a new methods was developed, the researcher would make available a difficult-(or impossible)-to-use program for others to try. Over a decade or more, the method would became established and eventually added to commercial software. For example, generalized additive models (Hastie and Tibshirani, 1990) took about 10 years to appear in Splus and are still not available in other commonly used packages.

For a method to be influential among molecular biologists, it must be easy to access and use from the beginning. The standard is to create a web-site where other laboratories can bring their data and conduct the analysis as a visitor, taking away the results. See examples by Bouton et al. (2003) or Colantuoni, et al. (2003). The success of biostatistical research centers will be increasingly dependent on their ability to disseminate their methods and papers to the end-users at a faster rate than has been traditionally attempted. The risks of adopting for general use, innovative methods are non-trivial and not to be ignored. But given the appropriate cautions, we expect the trend toward more rapid dissemination of biostatistical methods to continue.

7 Summary

We take great pleasure in congratulating Professor Sir David Cox on the occasion of his 80th birthday. We hope that this brief tribute has not severely misrepresented any of his ideas.

We hope our readers agree that we live in exciting times for biomedical and public health scientists and the biostatisticians who create and apply novel research methods. New kinds of measurements abound. The trend toward high-dimensional responses makes acute the need for better statistical tools (and statisticians). We believe the biostatistics will in the decade to come make important contributions to:

- improve biological measurement

- develop biologically-relevant statistical models with parameters that represent key targets for inference
- develop designs and analyses that make efficient use of the new measurements and valid assessments of the strength of evidence
- disseminate new methods more widely to biologists that will rely upon them.

DRC's approach to statistical science, here called the Cox Model, will continue to bear scientific fruit for the decades ahead. Of this we are certain. On this special occasion, we say thanks not only for what is done but for all the future progress for which you have already paved the way.

References

Armitage, P. (2003). Fisher, Bradford Hill, and randomization. *International Journal of Epidemiology*, **32**:925-8.

Barndorff-Nielsen, OE and Cox, DR. (1989). *Asymptotic Techniques for Use in Statistics.*Chapman & Hall, pp 252.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., B*, **36:** 192-236.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with Discussion). *Annals of the Institute of Statistical Mathematics*, **43**:1–59.

Bodmer, W. (2003). RA Fisher, statistician and geneticist extraordinary: a personal view. *International Journal of Epidemiology*, **32**:938–942.

Bouton, C., Henry, G. Colantuoni, C., Pevsner, J. (2003). Dragon and Dragon View: methods for the annotation, analysis and visualization of large-scale gene expression data. in Parmigiani, G., Garrett, ES., Irizarry, RA., Zeger, SL. *The Analysis of Gene Expression Data*. Springer, New York.

Box, GEP. and Cox, DR. (1964). An analysis of transformations. *J Roy Stat Soc, Series B.*, **26**:211–252.

Breiman, L. (2001). Statistical modeling: the two cultures (with discussion). *Stat Sci*, **16**:199–215.

Breiman, L. and Friedman, J. (1997). Predicting multivariate responses in multiple linear regression (with discussion). *J Roy Stat Soc, Series B*, **59**:3–37.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth.

Breslow, NE. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *J Am Stat Assoc*, **88**:9–25.

Caspi, A., Sugden, K., Moffitt, TE., et al. (2003). Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*, **301**:386-389.

Chalmers, I.(2003). Fisher and Bradford Hill: theory and pragmatism? *International Journal of Epidemiology*, **32**: 922–924.

Chetelat G, Baron JC. (2003). Early diagnosis of Alzheimer's disease: contribution of structural neuroimaging. *Neuroimage*, **2**:525–541.

Clayton, DG. (1996). Generalized linear mixed models. In *Markov Chain Monte Carlo in Practice*, ed. W.R. Gilks, S. Richardson and D.J. Spiegelhalter, 275–301. London : Chapman and Hall.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics*, **43**:671–681.

Colantuoni, C., Henry, G., Bouton, C., Zeger, SL., Pevsner, J. (2003). Snomad: biologist-friendly web tools for the standardization and normalization of microarray data. in Parmigiani, G., Garrett, ES., Irizarry, RA., Zeger, SL. *The Analysis of Gene Expression Data*. Springer, New York.

Collins, FS., Green, ED.,Guttmacher, AE., Guyer, MS. (2003). A vision for the future of genomics research. *Nature*. **422:** 835-847.

Coombs, CH. (1964). *A Theory of Data*. New York: Wiley.

Cornfield, J. (1951). A method for estimating comparative rates from clinical data: application to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute*, **11:** 1269-1275.

Cox, DR. (1955). Some statistical methods related with series of events (with Discussion). *J Roy Stat Soc, Series B*, **17**:129–157.

Cox, DR. (1958a). Some problems of connected with statistical inference. *Annals of Math. Stat.*, **29**:357-372.

Cox, DR. (1958b). The regression analysis of binary sequences (with discussion). *J Roy Stat Soc, Series B*, **20**:2151-242.

Cox, DR. (1972). The statistical analysis of point processes, in PA Lewis (ed.), *Stochastic Point Processes*, Wiley, New York, pp. 55-66.

Cox, DR. (1972). Regression models and life tables (with Discussion). *J Roy Stat Soc B*, **34**:187–220.

Cox, DR. (1986). Comment on Holland, W. Statistics and causal inference.*J Am Stat Assoc*, **81**:945-960.

Cox, DR. (1997). The current position of statistics: a personal view, (with discussion). *International Statistical Review*, **65:** 261-276.

Cox, DR., Hinkley, DV. (1974). *Theoretical Statistics*, London: Chapman and Hall.

Cox, DR., Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J Roy Stat Soc, Series B*, **49**:1-39.

Cox NJ, Frigge M, Nicolae DL, Concanno P, Harris CL, Bell GI, Kong A. (1999). Loci on chromosome 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genet*, **21**:213–215.

Diggle, P.J., Moyeed, R.A. and Tawn, J.A. (1998). Model-based Geostatistics (with discussion). *Applied Statistics*, **47**:299-350.

Doll, R. (2003). Fisher and Bradford Hill: their personal impact. *International Journal of Epidemiology*, **32**:929-931.

Dominici, F. McDermott, A. Zeger, SL., Samet, JM. (2003). National maps of the effects of PM on mortality: exploring geographical variation. *Environmental Health Perspectives*, **111**:39-43.

Feinberg, AP. (2001). Cancer epigenetics take center stage. *Proc Nat Acad Sci*, **98**:392-394.

Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd.

Fisher,R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, **33**, 503-513.

Fisher, R.A. (1935). *The Design of Experiments.* Edinburgh: Oliver and Boyd.

Frangakis, CE., and Rubin, DB. (1999). Addressing complications of intention-to-treat analysis in the presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, **86**:365-379.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, **19**:1-141.

Gentleman, R., Carey, V. (2003). Visualization and annotation of genomic experiments. in Parmigiani, G., Garrett, ES., Irizarry, RA., Zeger, SL. *The Analysis of Gene Expression Data*. Springer, New York.

Glonek, G.F.V. and Solomon, P.J. (2004). Factorial and time course designs for cDNA microarray experiments. *Biostatistics*, **5**:89-111.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares. *Biometrika*, **73**:43-56.

Greenland, S., Robins, J. M., Pearl, J. (1999). Confounding and collapsibility in causal inference. *Statistical Science*, **14**:29-46.

Harvard University Department of Systems Biology website: (http://sysbio.med.harvard.edu/). February 20, 2004.

Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall: London.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

Hastie, T., Tibshirani, R., Eisen, MB., Alizadeh, A., Levy, R., Staudt, L., Chan, WC., Botstein, D., Brown, P. (2000). Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **1**.

Hill, AB. (1965). The environment and disease: association or causation? *Proc R Soc Med*, **58**:295-300.

Heagerty, PJ., Zeger, SL. (2000). Marginalized multilevel models and likelihood inference (with discussion). *Statistical Science*, **15**:1-26.

Holland, W. (1986). Statistics and causal inference.*J Am Stat Assoc*, **81**:945-960

Irizarry, RA. Gautier, L. and Cope, LM. (2003). An R package for analysis of Affymetrix oligonucleotide arrays. in Parmigiani, G., Garrett, ES., Irizarry, RA., Zeger, SL. *The Analysis of Gene Expression Data*. Springer, New York.

Jeong, H., Mason, SP., Barabsi, A-L., Oltvai, ZN. (2001). Lethality and centrality in protein networks, *Nature*, **411**:41.

Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, **82**, 35-45.

Kaufman, JS., Kaufman, S., Poole, C. (2003). Causal inference from randomized trials in social epidemiology. *Soc Sci Med*, **7**:2397-2409.

Kious, BM. (2001). The Nuremberg Code: its history and implications. *Princet J Bioeth*, **4**:7-19.

Kempthorne, O. (1952). *The Design and Analysis of Experiments*. New York: Wiley.

Kerr, M.K. and Churchill, G.A.(2001). Experimental design for gene expression microarrays. *Biostatistics*, **2**:183-201.

Liang, K-Y., Zeger, SL. (1995). Inference based on estimating functions in the presence of nuisance parameters (with discussion). *Statistical Science*, **10**:158-172.

Liang. K-Y., Chiu, Y-F,, Beaty, TH, Wjst, M. (2001). Multipoint analysis using affected sib pairs: incorporating linkage evidence from unlinked regions. *Genet Epidemiol*, **21**:105-122.

Lindley, DV., Smith, AFM. (1972). Bayes estimates for the linear model (with discussion). *J Roy Stat Soc, Series B*, **34:**1-41.

Marks, H.M. (2003). Rigorous uncertainty: why RA Fisher is important. *International Journal of Epidemiology*, **32**:932-937.

Medical Research Council Streptomycin in Tuberculosis Trials Committee. Streptomycin treatment for pulmonary tuberculosis. (1948). *British Medical Journal*, 769-782.

Medical Research Council Research Funding Strategy and Priorities, 2001-2004. (2001). Medical Research Council, London.

Milton, J. (2003). Spies, magicians, and Enid Blyton: how they can help improve clinical trials. *International Journal of Epidemiology*, **32**:943-944.

Moller, J., Syversveen, A. and Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, **25**:451-482.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *J Roy Stat Soc, Series A*, **135**:370-384.

Oxford English Dictionary. Ed. J. A. Simpson and E. S. C. Weiner. 2nd ed. (1989). Oxford: Clarendon Press, OED Online. Oxford University Press. 4 Apr. 2000. ¡http://dictionary.oed.com/cgi/entry/00181778¿

Ripley, B.D. (1994). Neural networks and related methods for classification (with Discussion). *J Roy Stat Soc, Series B*, **56**:409-456.

Rothman, KJ., Greenland, S. (1998). *Modern Epidemiology, Second Edition.* Philadelphia: Lippincott Williams and Wilkins.

Royall, R. (1997). *Statistical Evidence: a likelihood paradigm.* London: Chapman and Hall.

Rubin, DB. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**:688-701.

Rubin, DB. (2001). Estimating the causal effects of smoking. *Stat Med*, **20**:13951414.

Rubin, DB. (2002). The ethics of consulting for the tobacco industry. *Stat Methods Med Res*, **11**:373-380.

Ruczinski, I., Kooperberg, C., LeBlanc, M. (2004). Logic regression. *Journal of Computational and Graphical Statistics*, to appear.

Scharfstein, DO., Rotnitzky, A., Robins, JM. (1999). Adjusting for non-ignorable drop-out using semi-parametric non-response models, *J Am Stat Assoc*, **94**:1096-1146.

Spielman RS, McGinnis RE, Ewens WJ. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, **52**:506-516.

Tukey, JW. (1977). *Exploratory Data Analysis.* Reading: Addison-Wesley.

Watson, JD. and Crick, FHC. 1953. Molecular structure of Nucleic Acids. *Nature*, **171**:737–738.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models.* New York: Springer.

Wjst, M., Fischer, G., Immervoll, T., et al. (1999). A genome-wide search for linkage to asthma. German Asthma Genetics Group. *Genomics*, **58**:1–8.

Wulfsohn, M.S. and Tsiatis, A.A (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**:330–339.

Van de Vijver, MJ., Yudong, DHE, Van T Veer, LJ. et al. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, **347**:1999-2009.

von Mering, C., Zdobnov, EM., Tsoka, S. (2003). Genome evolution reveals biochemical networks and functional modules. *Proc Nat Acad Sci*, **100**:1542815433.

Yang, Y.H. and Speed, T. (2002). Design issues for cDNA microarray experiments. *Nature Reviews: Genetics*, **3**:579-88.

Zerhouni, E. (2003). The NIH roadmap. *Science*, **302**:(5642):63-72