J. R. Statist. Soc. A (2015)
178, *Part* 4, *pp.* 1–18

Statistics: a data science for the 21st century

Peter J. Diggle

Lancaster University, UK

[*The address of the President delivered to* The Royal Statistical Society *on Wednesday, June 24th, 2015*]

Summary. The rise of data science could be seen as a potental threat to the long-term status of the statistics discipline. I first argue that, although there is a threat, there is also a much greater opportunity to re-emphasize the universal relevance of statistical method to the interpretation of data, and I give a short historical outline of the increasingly important links between statistics and information technology. The core of the paper is a summary of several recent research projects, through which I hope to demonstrate that statistics makes an essential, but incomplete, contribution to the emerging field of 'electronic health' research. Finally, I offer personal thoughts on how statistics might best be organized in a research-led university, on what we should teach our students and on some issues broadly related to data science where the Royal Statistical Society can take a lead.

Keywords: Data science; Electronic health research; Health surveillance; Informatics; National Health Service prescribing patterns; Reproducible research; Statistical education

1. The rise of data science: threat or opportunity?

The first thing to say is that we have been here before. I began my career in 1974, at which time statistical software packages were beginning to become widely available. This was seen by some of my colleagues as an existential threat. If useful statistical methods could be implemented in software, surely would not the need for statisticians diminish? In fact, the reverse happened, for at least three reasons. Firstly, if something is impossible it is easy to convince yourself that you can get by without it. Packages enabled scholars of many disciplines who might previously have considered statistics irrelevant to their subject to begin to appreciate its power. Secondly, packages enabled *statisticians* to do more things routinely, again increasing the reach of statistics to other disciplines. Thirdly, packages could not design studies—a point to which I shall return.

Having seen off the threat of packages, should we feel threatened by the rise of data science? Undoubtedly, there *is* a threat, but it is one that has been with us for a very long time, namely that any numerate scholar can operate as an amateur statistician within their own substantive discipline. This may explain why some people continue to view statisticians as technicians rather than as wholehearted collaborators. However, we are still here, many years after Rutherford may or may not have said what is attributed to him; the variant I favour is 'If your result needs a statistician then you should design a better experiment,' and my only quibble with the sentiment expressed is the implicit exclusion of the statistician from the design phase.

So what exactly *is* data science, and how does it relate to its close cousins, information science and statistics? *Wikipedia* definitions may not be authoritative, but they are often illuminating. Dated December 30th, 2014, we find the following entries.

Address for correspondence: Peter J. Diggle, Lancaster Medical School, Lancaster University, Lancaster, LA1 4YB, UK.
E-mail: p.diggle@lancaster.ac.uk
 0964–1998/15/178000

(a) '*Data science* is... the extraction of knowledge from data.... It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, and information technology....'
(b) '*Information science* is an interdisciplinary field primarily concerned with the analysis, collection, classification, manipulation, storage, retrieval, movement, dissemination, and protection of information.'
(c) '*Statistics* is the study of the collection, analysis, interpretation, presentation, and organization of data.'

If nothing else, these definitions show very considerable overlap. But, for me at least, the *Wikipedia* headline definition of data science comes closer to my definition of statistics than does its definition of statistics, whereas its definition of information science seems to me to be much more concerned with technology than with science. So if data science is a close relation of statistics, its increasing popularity must surely present us with an opportunity. We should embrace data science, proudly assert what we can offer it and humbly acknowledge what we can learn from it.

What can we offer?

Crucially, we can assert that uncertainty is ubiquitous and that probability is the correct way to deal with uncertainty (Lindley, 2000, 2006). We understand the uncertainty in our data by building stochastic models, and in our conclusions by probabilistic inference. And on the principle that prevention is better than cure we also minimize uncertainty by the application of the design principles that Fisher laid down 80 years ago (Fisher, 1935), and by using efficient methods of estimation.

Also, context matters. Borrowing from the *Wikipedia* headline definition of data science, the extraction of knowledge from a given set of data depends as much on the context in which the data were collected as on the numbers that the data set contains.

And what can we learn?

Principally, we learn that a published article is no longer a complete solution to a practical problem. We need our solutions to be implemented in software, preferably open source so that others can not only use but also test and, if need be, improve our solutions. We also need to provide high quality documentation for the software. And in many cases we need to offer an accessible, bespoke user interface.

At one time, I would have argued that data science *is* just a new name for statistics. I would now agree with Professor Iain Buchan (University of Manchester) that this misses an essential ingredient, namely informatics (information science by another name), a term that encompasses the hardware and software engineering that is needed to convert routinely recorded data into usable formats and to build bespoke software solutions for non-expert users. To paraphase a remark that Iain made to me recently, informatics seeks to maximize the utility of data, whereas statistics seeks to minimize the uncertainty that is associated with the interpretation of data.

On a related topic, the provision of open source software seems to me also to be fundamental to the goal of achieving reproducibility of research findings that rely on computational methods. This issue has acquired particular prominence in the context of biological research based on modern high throughout technologies. See, for example, Baggerly and Coombes (2009, 2011).

Developing protocols to ensure that scientific findings, and in particular their associated statistical analyses, are reproducible, has become a substantial area of methodological research in its own right; see, for example, Gentleman and Lang (2007) and the special issue of *Computing in Science and Engineering* guest edited by Fomel and Claerbout (2009). Reproducibility of computational results falls short of the traditional view of scientific reproducibility by

independent replication of substantive findings, but it seems to me unexceptionable as a minimal standard and is becoming accepted as such; see, for example, Laine *et al.* (2007) and Peng (2011). The journal *Biostatistics* has promoted computational reproducibility since 2009, initially in an editorial (Diggle, *et al.*, 2009) and subsequently in a discussion piece introduced by Keiding (2010), who emphasized that computational reproducibility of an analysis is no guarantee of its scientific usefulness.

2. Statistics and information technology: a very short history

There seems general agreement that the world's first electronic digital computer was the 'Colossus' machine that was developed at the Bletchley Park code breaking centre during the Second World War, and first used in February 1944 (Copeland, 2006). At this time, statistical computations relied on the use of mechanical calculators. A famous example is Fisher's 'millionaire' calculator, which features in some well-known images of Fisher, and of his successor at Rothamsted, Frank Yates (Ross, 2012).

Fisher and Yates were very much 'hands on' in their use of mechanical calculators. In Australia, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) Division of Mathematical Statistics took a different approach. Dr Peter Thorne (the Pearcey Foundation), speaking on an Australian Broadcasting Corporation science programme, recalled that, in the 1940s,

> 'if you wanted to do mathematical calculations in Australia, you hired a person, usually a woman, who used a calculating machine—either mechanical or hand-cranked'

(`http://www.abc.net.au/science/articles/2015/05/07/4184086.htm`). At the Division of Mathematical Statistics headquarters in Adelaide they used women plural, who were called 'computers' and whose collective job, in production line style, was to turn a data set into an analysis of variance, each computer having been trained to carry out a specific task.

As an undergraduate in the late 1960s, I was taught computer programming as a self-contained skill, but in my parallel courses in statistics I continued to use mechanical or (brave new world) electronic desk-top calculators. In the early 1970s, programming was beginning to enter the statistics curriculum and the first statistical packages were becoming available; GenStat was developed, initially at the Waite Institute in Adelaide and later at Rothamsted, in the late 1960s (Payne, 2009); around the same time in the USA, SAS was developed at North Carolina State University, and SPSS by Bent and Hull (1970) with a specific focus on social science applications.

Not everyone was convinced of the pedagogical merits of this development. At a meeting of the Royal Statistical Society in November 1972 my former Newcastle University colleague Dr Dennis Evans, an early advocate for the use of computers in the teaching of statistical methods, could not hide his frustration in responding to one of the discussants of his paper (Evans, 1973), remarking that

> 'I would like to take issue with ... when he assures us that students understand more about multiple regression when they invert a 5×5 matrix using a desk calculator rather than a computer package.'

By the 1980s, experience of hands-on statistical computing would form an integral part of a standard statistics degree syllabus. For me, a key driver of this was Nelder and Wedderburn's (1972) breakthrough paper on generalized linear models, and its dissemination through the GenStat and GLIM packages. This development offered, for the first time, a transparent path from the theory of the exponential family, through the unifying framework of the iteratively weighted least squares algorithm to the implementation of a wide range of statistical methods in a single piece of software.

The now ubiquitous Markov chain Monte Carlo (MCMC) methods were already being used in the 1970s for particular statistical tasks; see, for example, Ripley (1979). Gelfand and Smith (1990) brought MCMC methods into the statistical mainstream. 3 years later, the Royal Statistical Society held a discussion meeting around MCMC methods with papers by Smith and Roberts (1993), Besag and Green (1993) and Gilks *et al.* (1993). Packaged software implementations followed. Perhaps the best known, and certainly one of the first, was the BUGS project, which began in 1989 and embraced both a language and its associated software implementation (Gilks *et al*., 1994). As described in Lunn *et al*. (2009), early versions of the BUGS software were running from 1991 onwards, before the first stable version was released in 1995.

Arguably the most transformational development in statistical software since the 1990s has been the R project (`www.r-project.org`). The R language, which had its origins in the S language (Becker *et al*., 1988), was developed by Ross Ihaka and Robert Gentleman, working at the University of Auckland in the mid-1990s (Ihaka and Gentleman, 1996). One important aspect of R is that it is open source; the project is overseen by the R Foundation, which is a not-for-profit organization hosted by the Vienna University of Economics and Business. However, for most users its crucial feature is its extendibility through a plethora of 'contributed packages,' all of which (6637 on May 12th, 2015) can be downloaded from the project's Web site, and some of which operate as interfaces to other systems, e.g. the `R2WinBUGS` package. An R package has become the standard vehicle for disseminating novel statistical methodology, and almost a pre-requisite for new methodology becoming widely used in practice.

3. Case-studies in 'e-health' research

The last two decades have seen a transformation in the importance of cutting-edge statistical and computational methods to research in the life sciences, to the extent that many biostatisticians now publish their original research in life science journals rather than in statistics journals. Much of the focus of this activity has been motivated by the emergence of powerful new technologies focused on molecular level problems in biomedical research. The journal *Biostatistics*, which was launched in 2000, illustrates this increasing focus. Over its first 10 years, with an unchanging editorship, the proportions of its published papers that dealt explicitly with genetics or 'omics,' i.e. excluding papers on high dimensional data without a specific area of application, rose from 0.09 in 2000 to 0.31 in 2009 (Fig. 1).

More recently, there has been increasing interest in capitalizing on the parallel opportunities for statistical and computational innovation in the population health sciences, under the label of '*e-health*' *research*. A major boost to e-health research in the UK was a recent call for a network of e-health research centres led by the Medical Research Council. The call referred to 'the wealth of electronic health data within the NHS' and the opportunities for using these data

> 'to identify more effective treatments, improve drug safety, assess risks to public health and study the causes of diseases and disability.'

I see abundant scope for statisticians to contribute to e-health research, in the same way that they have contributed to bioinformatics research. Risking a charge of self-indulgence, I shall illustrate this by using some of the e-health research projects in which I have had some direct involvement. All are incomplete, in the sense that the application of statistical methods is not sufficient to deliver a useful solution; informatics is also needed, to deliver the required input data, to translate the results of the analysis into a user accessible form and to deliver results in realtime.

3.1. Realtime spatial surveillance of gastroentric illness

Diarrhoeal disease affects approximately 25% of the UK population annually (Tam *et al*., 2012).

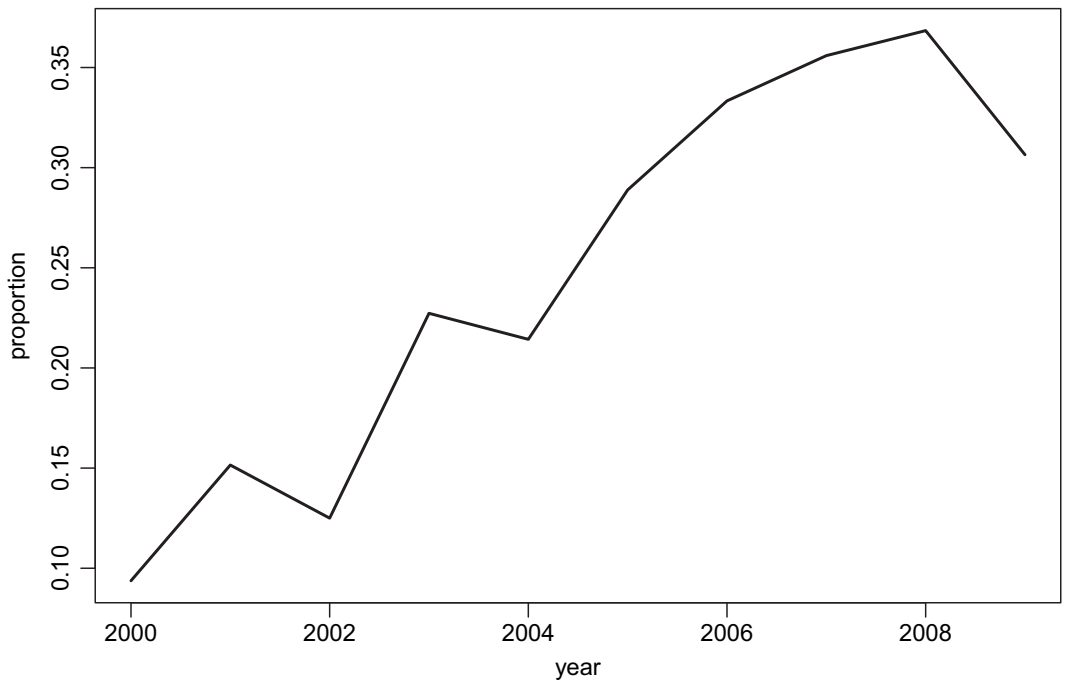


Fig. 1. Proportion of published papers in volumes 1–10 of the journal *Biostatistics* that deal explicitly with genetics or 'omics'

Traditional surveillance methods are effective in detecting point source outbreaks of diarrhoea and vomiting that are characterized by large numbers of incident cases within a tight geographical area over a very short period of time. However, they lack sensitivity to less dramatic fluctuations in incidence that are more typical of low level and/or intermittent contamination of the food supply.

More than 15 years ago, my research group in Lancaster received an approach from Dr Peter Hawtin in the Public Health Laboratory Service (now Public Health England) in Southampton. Peter had spotted the potential for spatial statistical methods to contribute to improved health surveillance systems, and in particular to enable earlier detection of anomalous incidence patterns of gastroenteric illness. The existing system required symptomatic patients to provide a faecal sample, which could then be analysed in the laboratory for the presence of specific pathogens. A system of this kind achieves high specificity but has low sensitivity and is slow. Pathogen identification can take several days, or longer when reference laboratories are busy. Delays of more than a week between first presentation and confirmation of a suspected case are not untypical (Diggle *et al*., 2003).

To address this, we used data from the then new telephone triage service, NHS Direct, to monitor spatiotemporal variation in the rate of calls to NHS Direct for which the caller's primary symptom was vomiting and/or diarrhoea. For each call, we were given the caller's residential postcode and an indication of their recent travel history. After removing data from callers who might be assumed to have become infected while travelling, we fitted a log-Gaussian Cox process model to the data. We modelled the stochastic intensity of the process at location x and time t as

$$\Lambda(x,t) = \alpha(x)\,\beta(t)\exp\{S(x,t)\}. \tag{1}$$

In model (1) $S(x,t)$ is a stationary Gaussian process such that $E[\exp\{S(x,t)\}]=1$ for all (x,t), and $\alpha(x)$ and $\beta(t)$ are deterministic functions that we estimated by using non-parametric kernel desnsity estimation and log-linear regression modelling respectively (Brix and Diggle, 2001; Diggle *et al*., 2005). The rationale for this factorization of $\Lambda(x,t)$ was as follows. We expected to see spatial variation due to a combination of the uneven distribution of the population and differential usage of the NHS Direct service by different sociodemographic groups. We also expected to see temporal variation due to the well-known seasonal pattern of food-borne disease incidence, together with day-of-week effects arising from different patterns of behaviour and the relative inaccessibility of other forms of healthcare at weekends. But, at least on short timescales, we did not expect these effects to interact. We therefore modelled the residual spatiotemporal variation about this expected pattern as a stochastic process, $R(x,t)=\exp\{S(x,t)\}$.

We then used the fitted model to construct *probability exceedance maps*, i.e. maps of quantities $p_c(x,t)=P\{R(x,t)>c|\mathcal{H}_t\}$, where $\mathcal{H}_t$ denotes the locations and dates of all calls up to and including day t. We used the term 'anomaly' to refer to locations and times at which $p_c(x,t)>0.95$, for a value of c that a public health professional would consider to be sufficiently large to be a cause for concern, and which might therefore initiate some form of local investigation.

These maps, updated daily, were used to provide early warnings of spatially and temporally localized anomalies that could be followed up for evidence of a common cause. We developed a prototype implementation in which the receipt of each day's data from the county of Hampshire triggered the overnight running of an MCMC algorithm to evaluate $p_c(x,t)$ over a fine grid for selected values of c. The output from the MCMC run was then used to update a Web interface displaying the corresponding maps. Fig. 2 shows a snapshot for March 8th, 2002.

The project ultimately failed to complete the translation from research to practice, primarily through lack of resources. It has recently been revived by a Health Innovation Challenge Award to Professor Sarah O'Brien (University of Liverpool), who is leading a multidisciplinary team within which we plan to incorporate an updated version of the statistical model as one component of an integrated, rapid response surveillance system.

3.2. National Health Service prescribing patterns

Since December 2011, comprehensive data on National Health Service prescribing throughout England has been made freely available, by general practice and calendar month. Rowlingson *et al*. (2013) combined these data with (also freely available) data from the 'General practitioner quality and Outcomes framework' and Ordnance Survey CodePoint open data, and used these data sets to construct maps of the countrywide variation in prescribing rates for particular conditions. Using a simple kernel smoothing method to identify extreme local variations they found, among other things, striking variations in prescribing rates for methylphenidate (Ritalin), which is the the recommended medication for treatment of attention deficit hyperactivity disorder (see National Institute for Health and Clinical Excellence (2009)).

Fig. 3 illustrates one such example. It shows unsmoothed prescribing rates for September 2011 in and around the metropolitan county of Merseyside. The average spend per child was 60.4 p in the Wirral (between the Mersey and Dee estuaries) and 7.4 p in Liverpool (north of the Mersey). The map makes it clear that the discrepancy between the two figures is not the result of a small number of overprescribing (or, conceivably, underprescribing) practices in a particular area. Whatever the explanation, it is difficult to reconcile differences of this magnitude with any notion of equity of health service delivery nationwide. Barry Rowlingson (Lancaster University) has since supervised a team of two Bachelor of Science graduate interns from Lancaster University's Computing Science department, Joshua Crick and Matthew McComish, to build a system

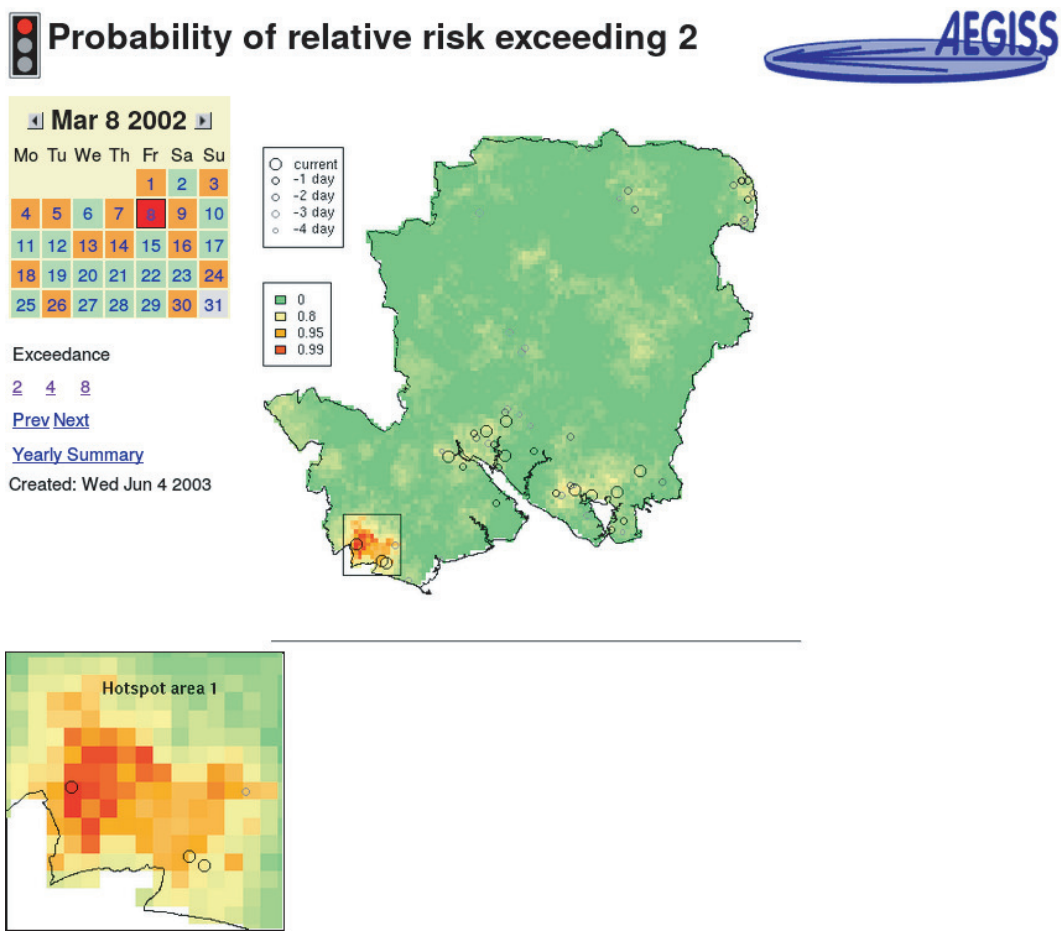


Fig. 2. Screen shot of the AEGISS probability exceedance map for March 8th, 2002: the small circles on the map identify the residential locations of callers over a 5-day moving window; the colour scale of the map has been chosen to highlight only areas with a high predictive probability of exceeding the threshold $c = 2$, corresponding to a doubling of intensity relative to expectation for that time and place; buttons allow the user to toggle through different dates and values of c, or to return to a summary page (design and Web implementation by Barry Rowlingson, CHICAS, Lancaster University Medical School)

for collecting the published data into a single database, substantially streamlining the process of extracting useful information about prescribing rates. This has confirmed that the discrepancy in prescribing rates between the Wirral and Liverpool has been sustained over at least a 13-month period. See `http://chicas.lancaster-university.uk/news/ritalin-march2015.html`.

3.3. Monitoring long-term progression to end stage kidney failure

Kidney failure can occur for many reasons, but in most cases its clinical manifestation is the end result of a process of progressive deterioration in kidney function that can remain asymptomatic, and therefore undetected, for many years. Although most cases of impending kidney failure are irreversible early detection, followed by ameliorative treatment including aggressive control of blood pressure, can slow its rate of progression. This benefits both the patient and the health

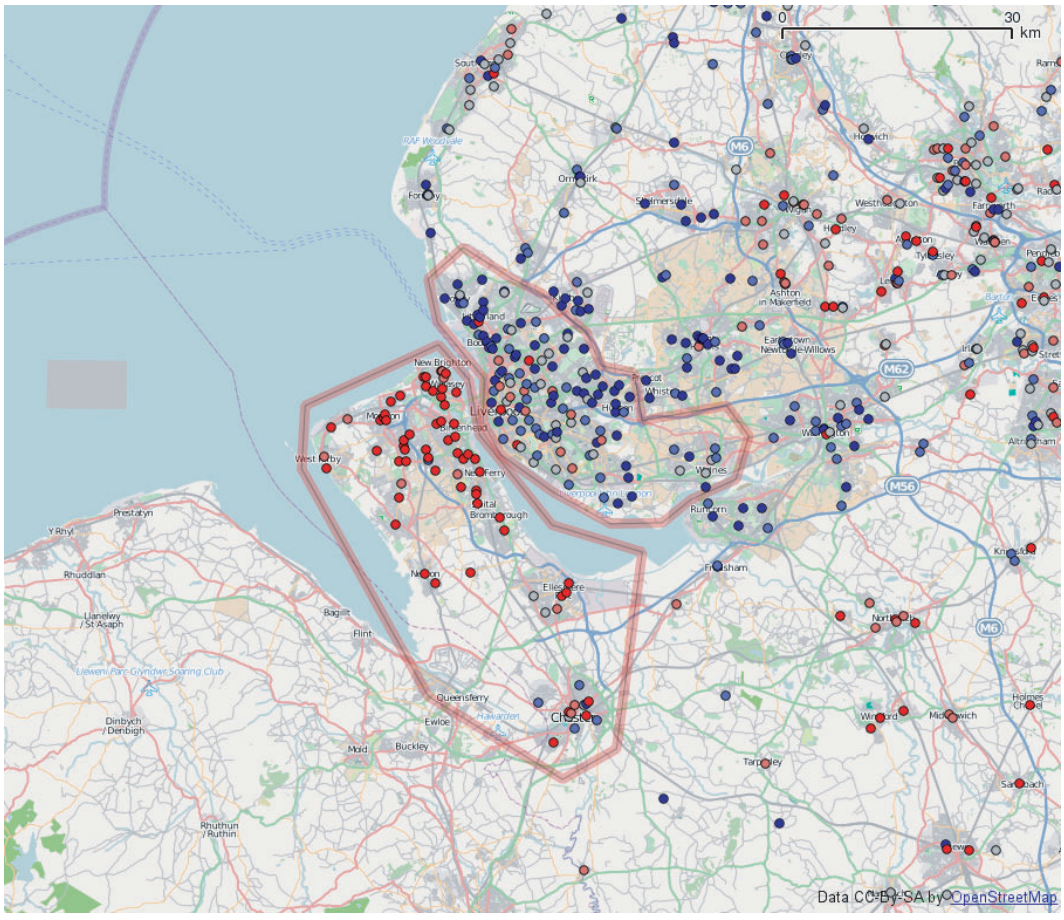


Fig. 3. Methylphenidate prescribing in Merseyside, September 2011: general practitioner practice locations are colour coded according to their level of prescribing, by quintiles of average cost per child from bright blue (lowest), through dull blue, grey and dull red to bright red (highest)

service by delaying the need for invasive and expensive renal replacement therapy, i.e. dialysis or transplantation. Progression is described by the rate of change in a blood biomarker, serum creatinine, which in clinical practice is often converted to an *estimated glomerular filtration rate* eGFR (Levey *et al.*, 1999); on a log-scale, eGFR is essentially equivalent to serum creatinine level adjusted for age, sex and ethnicity. Clinical guidelines in the UK advise that if a person in primary care is losing at least 5% of kidney function per year they should be considered for referral to specialist secondary care.

The Salford integrated record system, which was pioneered in 2003, integrates information from both primary and secondary care throughout the city of Salford. It includes an anonymized research data repository that can be accessed for specific research projects subject to the usual ethical safeguards. In collaboration with clinicians at the Royal Salford Hospital, we have been able to analyse repeated measurement data on serum creatinine levels for 22930 patients considered to be at risk of end stage renal failure (Diggle *et al.*, 2015).

A useful general model for repeated measurement sequences, here of log-transformed eGFR, is

$$Y_{ij} = X_i(t_{ij})\alpha + U_i + S_i(t_{ij}) + Z_{ij}. \quad (2)$$

In model (2), Y_{ij} denotes the log-transformed eGFR-response for subject $i = 1, \ldots, m$ at time t_{ij}, $j = 1, \ldots, n_i$ and $X_i(t_{ij})$ denotes a set of explanatory variables with corresponding regression parameters α to be estimated. The U_i are independent $N(0, \omega^2)$ random variables, the $S_i(t)$ are independent copies of a zero-mean, continuous time stochastic process and the Z_{ij} are mutually independent $N(0, \tau^2)$ random variables representing measurement error.

Diggle *et al.* (2015) modelled $S_i(t)$ as the integral of a continuous-time random walk,

$$S_i(t) = \int_0^t B_i(v)\mathrm{d}v, \quad (3)$$

where $B_i(v)$, the rate of change at time v, is Brownian motion. They then used the fitted model to compute the conditional distribution of $B_i(t)$ given all information on patient i available at time t. Fig. 4 shows the result for one patient. In my opinion, the most useful of the various quantities plotted in Fig. 4 is the predictive probability that $B_i(t) < -0.05$. As with the spatial surveillance application described in Section 3.1, the thinking behind this is that, to assist clinical decision making, it is more useful to report the probability that a clinically agreed criterion has been met than, for example, to give clinicians interval estimates of $S_i(t)$ or $B_i(t)$.

3.4. African programme for onchocerciasis control

The potential for electronic systems to improve health services is not confined to developed countries. The nearly complete penetration of mobile phones into even the economically poorest African countries presents many opportunities to improve the delivery of healthcare, especially to remote areas.

Onchocerciasis is a severe public health problem in wet tropical regions, but especially so in sub-Saharan Africa. The disease is caused by the filarial worm *Onchocerca volvulus* and is transmitted through the bite of an infected *Simulium* blackfly. Its most severe manifestation is clear from its common name: river blindness. The African programme for onchocerciasis control is a multinational programme co-ordinated by the World Health Organization to reduce the prevalence of onchocerciasis (Remme, 1995). The programme involves the mass administration of an antifilarial medication, ivermectin (Mectizan), in affected areas. By the end of 2012, the programme had administered prophylactic medication to more than 100 million people in communities at risk of onchocerciasis infection across 24 participating countries (`http://www.who.int/apoc/cdti/achievements/en/`).

Loa loa filariasis, or loaiasis, is another filarial infection, in this case transmitted by the bite of a *Chrysops* fly. Loaiasis generates a high disease burden in large parts of sub-Saharan Africa but is considered to be a less serious public health problem than onchocerciasis because its patients usually do not suffer permanent consequences.

Implementation of the programme has been hampered in some areas by the recognition that individuals who are heavily infected with *Loa loa* parasites are at risk of experiencing a severe, occasionally fatal reaction to ivermectin (Boussinesq *et al.*, 1998, 2003). Boussinesq *et al.* (2001) have given empirical evidence that highly infected individuals are most likely to be found in areas of high prevalence of loaiasis. This has led to a recommendation that monitoring procedures during mass administration of ivermectin should be strengthened in areas where the prevalence of loaiasis is greater than 20%, which in turn has resulted in considerable effort being devoted to mapping the prevalence of loaiasis Africa wide; see, for example, Thomson *et al.* (2004), Diggle *et al.* (2007), and Zoure *et al.* (2011). The resulting maps are useful for large-scale operational decision making in regions where prevalence varies smoothly, but less so for identifying specific

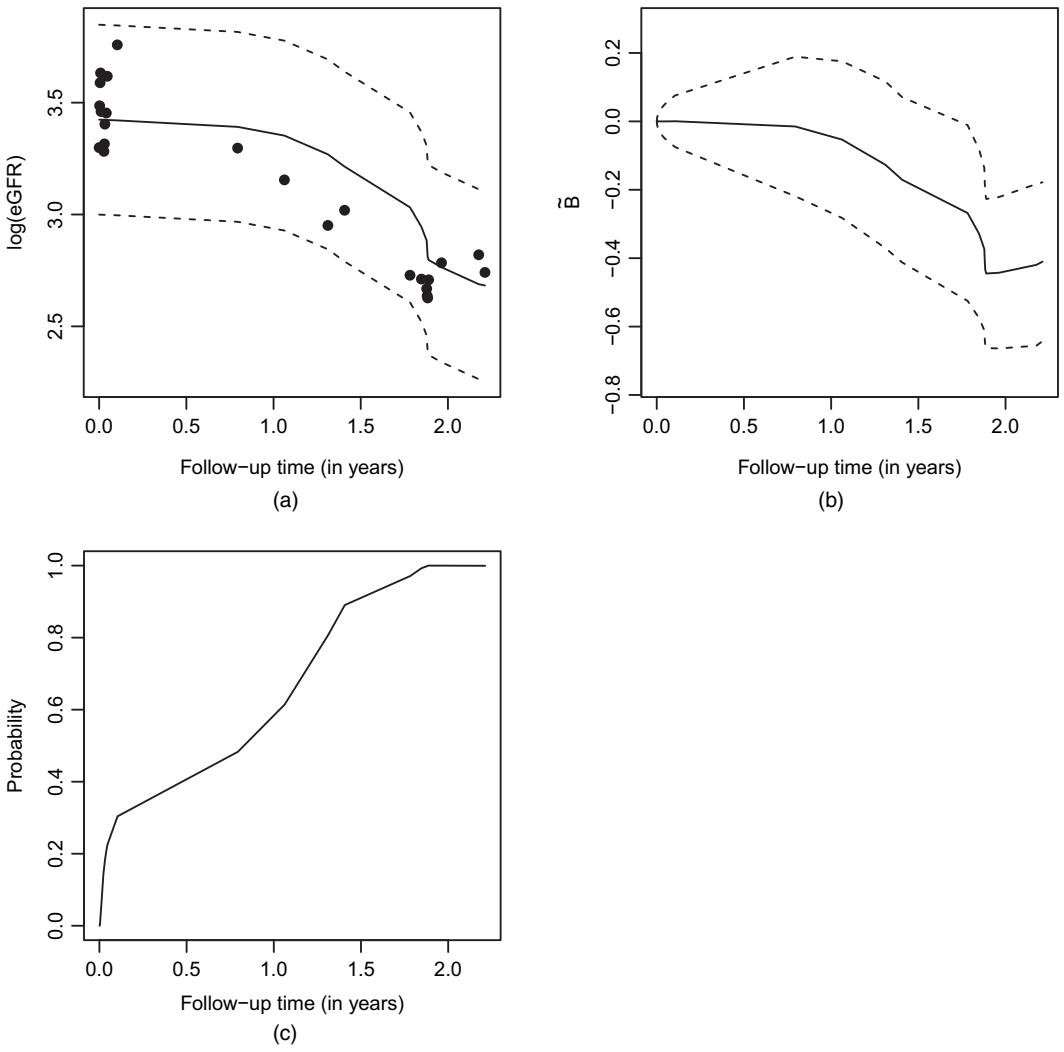
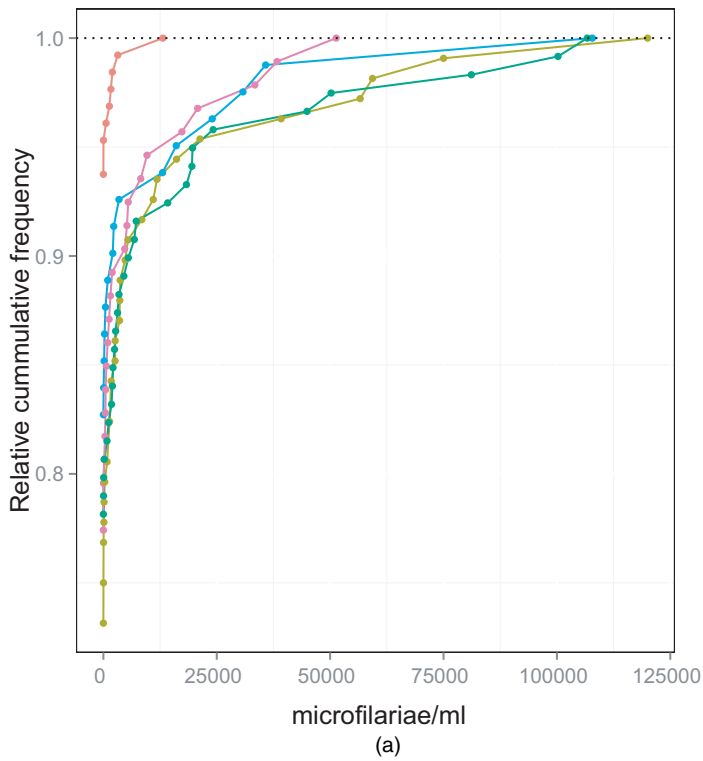


Fig. 4. Analysis of repeated measurements of eGFR for one patient: (a) log-transformed eGFR-measurements (•) with predictive mean (———) and 2.5% and 97.5% predictive quantiles (- - - - - -) of the predictive distribution for the underlying error-free log-transformed eGFR calculated at each time t conditionally on data available at time t; (b) predictive mean (———), 2.5% and 97.5% quantiles (- - - - - -) of the corresponding predictive distribution for the first derivative of log-transformed eGFR; (c) predictive probability that the first derivative of log-transformed eGFR is less than −0.05

communities that are likely to contain high risk individuals. Furthermore, it is quicker, and therefore cheaper, to estimate community level loaiasis prevalence than to screen individuals for levels of *Loa loa* infection. This raises the following statistical problem: given only an estimate of community-level prevalence, what can be said about the likely number of highly infected individuals in the community? The definition of 'highly infected' is currently under debate. Unsurprisingly, there appears to be no sharp threshold below which individuals are at zero risk of experiencing a serious adverse reaction; current discusson within the programme is around levels of infection between 8000 and 30000 parasites per millilitre of blood.

We have analysed data from individual level parasite counts across 223 rural communities in



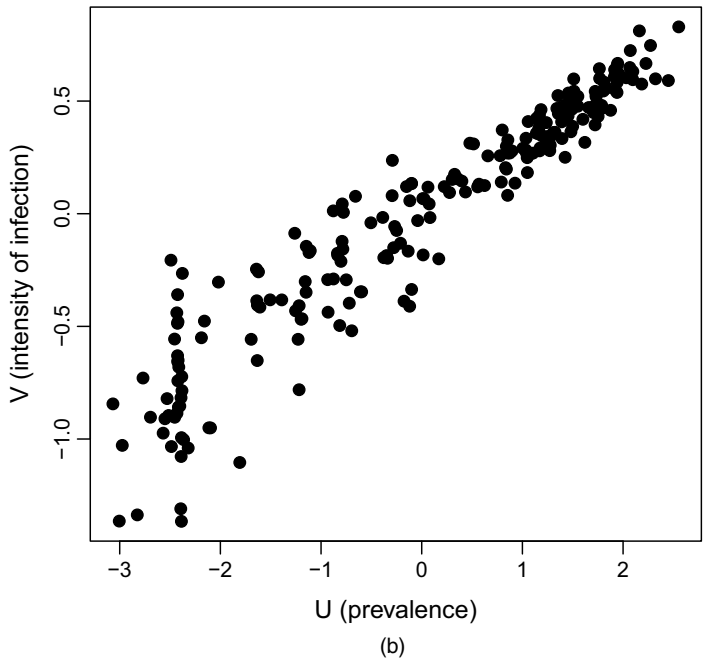


Fig. 5. (a) Emprical distributions of *Loa loa* parasite infection levels for five African villages and (b) point predictors (conditional expectations) of random effects U and V for 223 villages in the statistical model defined by equations (4)–(6); see Section 3.4 for detailed explanation

Cameroon, Congo and the Democratic Republic of Congo to establish that the distribution of positive *Loa loa* parasite infection levels (parasites per millilitre of blood) in any single community is well described by a Weibull distribution. Hence, denoting by Y the parasite infection level for a randomly sampled individual,

$$P(Y \leqslant y) = G(y) = \begin{cases} 1-\rho & y = 0, \\ 1-\rho+\rho\{1-\exp(-y/\lambda)^{\kappa}\} & y > 0. \end{cases} \tag{4}$$

Fitting model (4) separately to data from each village, we found that a common value $\kappa = 0.5$ gave a reasonably good fit, but that values of ρ and λ showed wide variation between villages. Also, the village level covariates that were available to us could explain only a very small proportion of this variation. We therefore adopted a bivariate random-effects model, setting

$$\log\{\rho/(1-\rho)\} = \alpha + U \tag{5}$$

and

$$\log(\lambda) = \beta + V, \tag{6}$$

where (U, V) follow a zero-mean, bivariate normal distribution. Our aim was to infer the probability that a randomly sampled individual will be heavily infected given the number, X say, of *Loa-loa*-positive individuals in a random sample of size n. Expressed more formally, our target for prediction is $T = \rho(U)\{1 - G(c; V)\}$, where c is the threshold that is used to define 'highly infected.' We found a moderately strong positive correlation between U and V (95% likelihood-based confidence interval 0.534–0.864; Fig. 5). As a result we could make usefully precise predictions of T by computing the predictive distribution of T given X and n. Furthermore, because the operationally useful values of c are in the upper tail of the distribution for most villages, these predictions are generally more precise than empirical estimates based on the binomial sampling distribution of the observed number of highly infected individuals in each sample.

The potential connection with e-health is that, in this context, our current laptop implementation is impractical for routine use in the field. However, the computations required to compute quantiles of the predictive distribution can be conducted offline for the relevant range of values of n and all $X \leqslant n$. The results could then be incorporated in a mobile phone implementation that would require only a set of look-up tables for a specified set of quantiles.

4. Statistics in context

4.1. Statistical mathematics and statistical science

About 30 years ago, in a letter to the Royal Statistical Society's former newsletter *News and Notes*, the late John Nelder proposed that 'mathematical statistics' should really be called 'statistical mathematics', a term that he also used in his Presidential address (Nelder, 1986). A similar sentiment appears in a later Presidential address by Professor David Hand, who remarked that

> 'failure to drive home this fundamental distinction between mathematics and statistics when teaching the pool from which the next generation of statisticians will be drawn is a lost opportunity' (Hand, 2009).

Nelder's suggestion appears not to have caught on, which I think is a pity because it provides a counterpoint to another useful term, namely statistical science. My definitions of these two terms would be as follows:

(a) *statistical mathematics* is that part of mathematics that provides the theoretical underpinning of statistical practice;
(b) *statistical science* is the intellectual engagement of statisticians with subject matter experts to advance our understanding of nature in its broadest sense.

Statistical mathematics and statistical science are equally important but very different activities. They require correspondingly different skills and rely on different kinds of motivation. Excellent statistical science can often be conducted by the imaginative application of mathematically simple tools, as in the analysis of a well-designed randomized clinical trial.

The above notwithstanding, just as today's statistics needs yesterday's mathematics, tomorrow's statistics may need today's mathematics. Statistics is not only a branch of mathematics, but it is undoubtedly a mathematical science. The various bodies that represent mathematics, statistics and operational research in the UK therefore need to work together, and to speak with one voice when making the case for the fundamental importance of the mathematical sciences to the future health and wealth of UK society.

4.2. Organizational models

In the UK, most academic statistics groups now sit within departments or schools of mathematics or mathematical sciences. This process has been driven to a considerable extent by successive research assessment exercises (now the research excellence framework, culminating in the subsuming of the statistics discipline within a single unit of assessment: mathematical sciences. In my opinion, this is unexceptionable in so far as it relates to statistical mathematics, but it risks fragmentation of the wider statistics discipline. Put simply, the research excellence framework results give a very incomplete picture of the strength of statistics in UK academia, either overall or in its geographical distribution, because many academic statisticians who work primarily at the interface with substantive areas, e.g. the biomedical or social sciences, have their work evaluated in other units of assessment. Does this matter? I think that it does, because high level policy decisions on funding academic research rely on high level summary information. If statistics research is evaluated solely within the mathematical sciences, much excellent statistical work is ignored. In this context, it is worth remembering that many of the breakthroughs in statistical research have their origins in other disciplines. The foundations of modern statistical design and inference were laid by Fisher, working in an agricultural research station. Some of the most important statistical developments in the mid-20th century, such as the design and analysis of randomized clinical trials (Armitage, 2003) or survival analysis (Cox, 1972), were inspired by the needs of medical research. Arguably the first example of the now ubiquitous framework of hierarchically structured stochastic models came from engineering (Kalman, 1960).

The co-location of statisticians with mathematicians makes sense from a teaching perspective; perhaps less so from a research perspective. So where should statisticians sit in a research organization?

In my own career, I experienced the best of both worlds when I spent 5 years in Australia working with CSIRO's Division of Mathematics and Statistics. Many of my colleagues operated from two offices: one co-located with other statisticians; one co-located with scientists in another discipline—in my case, ecologists in the Division of Wildlife Research. This gave, in effect, a physical manifestation of my dichotomy between statistical mathematics and statistical science. The result was a symbiotic relationship in which statisticians brought to our weekly meetings challenging problems from many different disciplines and took back to those disciplines solutions informed by a very wide range of statistical expertise. The CSIRO's statisticians

published regularly in scientific journals, as well as in applied and theoretical statistics and probability journals. A good number of them began their careers in the CSIRO as consulting statisticians with a Bachelor's or Master's level qualification, only later studying for a doctorate and becoming research scientists in their own right.

Whatever its scientific merits, the CSIRO model that I experienced was eventually perceived to be a luxury, and it did not survive a series of restructuring exercises beginning in the late 1980s. But I still regard it as the ideal organizational model for statistical research and training, and its essence should be eminently achievable if we can successfully promote statistical mathematics and statistical science as distinct, but kindred and equally valuable, activities. In my experience, university structures and devolved budgets often inhibit rather than promote this vision, leading (as with the aforementioned research excellence framework) to a fragmented organization in which multiple statistics groups communicate with each other, if at all, much less frequently than they should. I would like to see every research-led university in the UK create a statistics institute. Each statistician on the university's staff would have a dual appointment, to the institute and to an appropriate second discipline, be it mathematics, computer science or any one of the natural, biomedical or social sciences. Deep involvement of statisticians within the burgeoning number of data science institutes might be a more effective tactic to achieve the same goal, at least in the short term.

4.3. *We are what we teach*

My undergraduate course in the late 1960s and early 1970s taught statistics as a series of independent compartments. The aforementioned path breaking work of Nelder and Wedderburn (1972) broke down the divisions between the various analysis-focused compartments, but only within the limiting framework of independently replicated data. Later methodological research involving Monte Carlo methods of inference for hierarchically structured models achieved a similar unification of approaches to the analysis of dependent data by making likelihood-based inference feasible for almost arbitrarily complex models, albeit irrespective of the capacity for empirical validation of their assumptions. One consequence of this is that it is now rare for a statistician to describe themself by their particular methodological specialization. Overall, our discipline has evolved in two superficially different directions: an ever-increasing armoury of specific tools (remember those 6637 R packages); and the progressive replacement of *ad hoc* tests and estimators by principled, likelihood-based methods of inference.

This, coupled with the penetration of statistical method into so many substantive areas of investigation, should cause us to question our approach to teaching. I shall focus my comments largely on degree level teaching to students with aspirations to become professional statisticians. From this perspective, I cannot overemphasize the need for a solid mathematical foundation. I would like to see less statistics in undergraduate mathematics degrees, counterbalanced by a radical expansion of postgraduate statistics teaching.

Given a solid mathematical foundation, my suggested list of topics for Master of Science degree in statistics is

(a) design,
(b) probability and stochastic processes,
(c) likelihood-based inference,
(d) computation, including numerical methods and programming,
(e) communication, including scientific writing for both technical and lay audiences, and
(f) scientific method, and the foundations of at least one substantive area of application.

Note the absence from this list of any courses on specific statistical methods. The many topics on which I have never taken a lecture course include generalized linear models, survival analysis, longitudinal data analysis, non-parametric smoothing and spatial statistics, all of which I use routinely, and I hope competently. Furnished with a good understanding of probability, stochastic processes and likelihood-based inference, students can learn about specific methods by encountering them in project work. Projects could be stage managed to ensure that students do encounter a range of methodological challenges, but not to such an extent that they lose their open-ended character. I view this as a form of problem-based learning: an approach that is widely used in medical schools (Wood, 2008).

In contrast, an understanding of design, which seems to me fundamental to good statistical practice, is too often regarded as a specialist subject and as a consequence has disappeared entirely from many otherwise respectable statistics degree syllabuses.

My nomination of computation, including programming, presumes a first-degree qualification in mathematics; this is not to deny that computer science graduates can and should be attracted into statistics, in which case mathematical methods might be a suitable alternative for this slot. I should also emphasize that I envisage a course in programming to go much further than the ability to write simple `R` scripts to access packages.

Perhaps most importantly, I find it increasingly untenable that, for example, we expect a degree in biology to include a course in biostatistics, but we teach degrees in biostatistics that include no biology. If it were agreed that we should teach biology to biostatistics students, this could be delivered in different ways. One possibility would be a formal lecture course. Another would be to pair a statistics student with a student in discipline X, and to have the two of them co-author a single dissertation. This second option would raise all sorts of practical questions, but if these could be overcome the result might be that both students would be better prepared for subsequent careers in science.

5. Conclusions

Recognizing the danger of special pleading for one's own subject, I would claim that the unique strength of the statistics discipline is the extent of its relevance to the whole of the natural and social sciences. Where statisticians are organizationally separated from scientists, typically by forming a subsection of a mathematics department, they are in danger of missing the point. This comment is not anti mathematics. As I have tried to emphasize earlier, mathematical underpinning is as essential to the statistics discipline (and therefore to the training of statisticians) as it is to physics, to engineering and to modern biology. But, as a research community, we need to be clear when we are being statistical mathematicians and when we are being statistical scientists. Too many papers in statistics journals still include 'illustrative examples' that add nothing of value to the original methodology contained earlier in the paper; see, for example, Preece (1986).

Some comments in relation to the Society's journals follow.

(a) Our journals need to be more concerned with the dissemination of new insights, rather than with the archiving of immutable facts. The turnaround time from submission of a paper to a decision on its publication should be reduced from months to weeks.
(b) Published papers should be short, including a clear message of how they advance understanding, and intelligible to a general scientific audience. Detailed technical material can and should be published electronically.
(c) We should expect, rather than merely encourage, minimum standards of reproducibility of findings, including the routine deposition of code and data.

Wider issues on which the Society might wish to take a lead include the following list.

(a) In many scientific areas, most obviously the health sciences, concern about preserving the confidentiality of information on human subjects needs to be balanced against the public benefit of insightful statistical analysis (and sometimes critical reanalysis) of disaggregated data. This is especially so in the area that is loosely defined as health informatics. The Society is already active in this area, as exemplified by its data manifesto (`www.rss.org.uk/data-manifesto`). With a new government in place following the 2015 general election, we should continue to press for a more nuanced debate on the balance between personal privacy and public benefit.

(b) The emergence of subdisciplines whose title includes 'informatics' is very welcome, not least because it can put statistical thinking at the heart of cutting-edge science, a case in point being modern biology. An attendant risk is that these developments can inhibit the dissemination of new statistical methods across disciplinary boundaries if statisticians do not publish their results in general statistics journals. In some areas of informatics, there is also a tendency to overemphasize algorithms at the expense of inference and an accompanying assessment of uncertainty. The Society needs to engage with all the emerging informatics subdisciplines (medical, health, environmental,...) in the same mutually supportive way that it does with the mathematical sciences through its membership of the Council for the Mathematical Sciences.

(c) The social implications of the data explosion are arguably greater in developing than in developed countries. In particular, the deep penetration of the Internet and mobile phone technology can lead to radical improvements in the ability of poor communities to access information, education and healthcare. Our International Development Working Group is exploring ways in which the Society can contribute, with an initial focus on national statistical information systems. There is also a need, and the opportunity, for more of our members to be involved in scientifc capacity building initatives.

Acknowledgements

The work on *Loa loa* prevalence and intensity modelling is an on-going collaboration with Alison Galvani and Martial Ndeffo (Yale University, USA), Daniela Schlueter (Lancaster University, UK), Innocent Takougang (Yaoundé University, Cameroon), Tony Ukety (World Health Organization, Geneva, Switzerland), Samuel Wanji (University of Buea, Cameroon) and other contributors to the African programme for onchocerciasis control.

My views on the role relationship between statistics and informatics have been influenced by interactions with colleagues and friends whom I have already mentioned: in chronological order Denis Evans, Barry Rowlingson and Iain Buchan.

My views on the relationship between statistics and science have been influenced by too many people to mention individually, including the co-authors of most of my published work. However, I would like particularly to mention some formative experiences: spending 6 months in 1978 with the Swedish College of Forestry at the invitation of the late Bertil Matérn, where I first took part in fieldwork; visiting the CSIRO's Division of Mathematics and Statistics in 1980, where Nick Fisher, Ron Sandland, Murray Cameron and others exemplified for me the dual role of statistical mathematician and statistical scientist; collaborating with Scott Zeger and others at the Johns Hopkins University School of Public Health on various occasions since 1987, which led to an increasing focus of my work on problems in the biomedical and health sciences; most recently, encouraged by Madeleine Thomson (Columbia University), the privilege of being

given the opportunity to contribute in a small way to the improvement of public health in some of the world's poorest countries.

References

Armitage, P. (2003) Fisher, Bradford Hill, and randomization. *Int. J. Epidem.*, **32**, 925–928.

Baggerly, K. A. and Coombes, K. R. (2009) Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann. Appl. Statist.*, **3**, 1309–1334.

Baggerly, K. A. and Coombes, K. R. (2011) What information should be required to support clinical omics publications? *Clin. Chem.*, **57**, 688–690.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988) *The New S Language*. Pacific Grove: Wadsworth.

Bent, D. H. and Hull, C. H. (1970) *Statistical Package for the Social Sciences*. New York: McGraw-Hill.

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc.* B, **55**, 25–37.

Boussinesq, M., Gardon, J., Gardon-Wendel, N. and Chippaux, J. (2003) Clinical picture, epidemiology and outcome of Loa-associated serious adverse events related to mass ivermectin treatment of onchocerciasis in Cameroon. *Fil. J.*, **2**, 1–13.

Boussinesq, M., Gardon, J., Gardon-Wendel, N., Kamgno, J., Ngoumou, P., and Chippaux, J. P. (1998) Three probable cases of Loa loa encephalopathy following Ivermectin treatment for Onchocerciasis. *Am. J. Trop. Med. Hyg.*, **58**, 461–469.

Boussinesq, M., Gardon, J., Kamgno, J., Pion, S. D. S., Gardon-Wendel, N. and Chippaux, J. P. (2001) Relationships between the prevalence and intensity of Loa loa infection in the Central Province of Cameroon. *Ann. Trop. Med. Parasit.*, **95**, 495–507.

Brix, A. and Diggle, P. J. (2001) Spatiotemporal prediction for log-Gaussian Cox processes. *J. R. Statist. Soc.* B, **63**, 823–841.

Coombes, K., Wang, J. and Baggerly, K. (2007) Microarrays: retracing steps. *Nat. Med.*, **13**, 1276–1277.

Copeland, B. J. (2006) *Colossus: the Secrets of Bletchley Park's Codebreaking Computers*. Oxford: Oxford University Press.

Cox, D. R. (1972) Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B, **34**, 187–220.

Diggle, P. J., Knorr-Held, L., Rowlingson, B., Su, T., Hawtin, P. and Bryant, T. (2003) Towards on-line spatial surveillance. In *Monitoring the Health of Populations: Statistical Methods for Public Health Surveillance.*, (eds R. Brookmeyer and D. Stroup). Oxford: Oxford University Press.

Diggle, P., Rowlingson, B. and Su, T. (2005) Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, **16**, 423–434.

Diggle, P. J., Sousa, I. and Asar, O. (2015) Real-time monitoring of progression towards renal failure in primary care patients. *Biostatistics*, **16**, to be published.

Diggle, P. J., Thomson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, H., Boussinesq, M. and Molyneux, D. H. (2007) Spatial modelling and prediction of Loa loa risk: decision making under uncertainty. *Ann. Trop. Med. Parasit.*, **101**, 499–509.

Diggle, P. J., Zeger, S. L. and Peng, R. D. (2009) Reproducible research and Biostatistics. *Biostatistics*, **10**, 405–408.

Evans, D. A. (1973) Computers in the teaching of statistics. *J. R. Statist. Soc.* A, **136**, 153–190.

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

Fomel, S. and Claerbout, J. F. (2009) Reproducible research. *Computnl Sci. Engng*, **11**,

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *J. Am. Statist. Ass.*, **85**, 398–409.

Gentleman, R. and Lang, D. T. (2007) Statistical analyses and reproducible research. *J. Computnl Graph. Statist.*, **16**, 1–23.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993) Modelling complexity: applications of Gibbs sampling in medicine. *J. R. Statist. Soc.* B, **55**, 39–52.

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994) A language and program for complex Bayesian modelling. *Statistician*, **43**, 169–178.

Hand, D. J. (2009) Modern statistics: the myth and the magic. *J. R. Statist. Soc.* A, **172**, 287–306.

Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Computnl Graph. Statist.*, **5**, 299–314.

Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E. and van Noort, V. (2009) Repeatability of published microarray gene expression analyses. *Nat. Genet.*, **41**, 149–155.

Kalman, R. E. (1960) A new approach to linear filtering and prediction problems. *J. Basic Engng*, **82**, 35–45.

Keiding, N. (2010) Reproducible research and the substantive context. *Biostatistics*, **11**, 376–378.

Laine, C., Goodman, S. N., Griswold, M. E. and Sox, H. C. (2007) Reproducible research: moving toward research the public can really trust. *Ann. Intern. Med.*, **146**, 450–453.

Levey, A. S., Bosch, J. P., Lewis, J. B., Greene, T., Rogers, N. and Roth, D. (1999) A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Ann. Intern. Med.*, **130**, 461–470.
Lindley, D. V. (2000) The philosophy of statistics (with comments). *Statistician*, **49**, 293–337.
Lindley, D. V. (2006) *Understanding Uncertainty*. Hoboken: Wiley.
Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009) The BUGS project: evolution, critique and future directions. *Statist. Med.*, **28**, 3049–3067.
National Institute for Health and Clinical Excellence (2009) Attention deficit hyperactivity disorder diagnosis and management of ADHD in children, young people and adults. National Institute for Health and Clinical Excellence, London. (Available from `http://www.nice.org.uk/CG072`.)
Nelder, J. A. (1986) Statistics, science and technology. *J. R. Statist. Soc.* A, **149**, 109–121.
Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–84.
Payne, R. W. (2009) GenStat. *Computnl Statist.*, **1**, 255–258.
Peng, R. D. (2011) Reproducible research in computational science. *Science*, **334**, 1226–1227.
Peng, R. D., Dominici, F. and Zeger, S. L. (2006) Reproducible epidemiologic research. *Am. J. Epidem.*, **163**, 783–789.
Preece, D. A. (1986) Illustrative examples: illustrative of what? *Statistician*, **35**, 33–44.
Remme , J. H. (1995) The African programme for onchocerciasis control: preparing to launch. *Parasit. Today*, **11**, 403–406.
Ripley, B. D. (1979) Algorithm AS137. Simulating spatial patterns: dependent sample from a multivariate density. *Appl. Statist.*, **28**, 109–112.
Ross, G. (2012) Fisher and the millionaire the statistician and the calculator. *Significance*, **9**, no. 4, 46–48.
Rowlingson, B., Lawson, E., Taylor, B. and Diggle, P. J. (2013) Mapping English GP prescribing data: a tool for monitoring health-service inequalities. *BMJ Open*, **3**, article e001363.
Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.
Tam, C. C., Rodrigues, L. C., Viviani, L., Dodds, J. P., Evans, M. R., Hunter, P. R., Gray, J. J., Letley, L. H., Rait, G., Tompkins, D. S. and O'Brien, S. J. (2012) Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut*, **61**, 69–77.
Thomson, M. C., Obsomer, V., Kamgno, J., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Remme, J. H., Molyneux, D. H. and Boussinesq, M. (2004) Mapping the distribution of *Loa loa* in Cameroon in support of the African Programme for Onchocerciasis Control. *Fil. J.*, **3**,
Wood, D. F. (2008) Problem based learning. *Br. Med. J.*, **326**, 328–330.
Zoure, H., Wanji, S., Noma, M., Amazigo, U., Diggle, P. J., Tekle, A. and Remme, J. H. (2011) The geographic distribution of Loa loa in Africa: results of large-scale implementation of the Rapid Assessment Procedure for Loiasis (RAPLOA). *PLOS Neglctd Trop. Dis.*, **5**, no. 6, article e1210.