The Duke University scandal – what can be done?

It took courage, persistence, and dogged research to persuade journals and indifferent academia that a major piece of cancer research was desperately flawed – even though it was guiding treatment in clinical trials. **Darrel Ince** tells the story, and finds that the Duke problem was not a one-off.

If you look at the current mortality rates of American cancer sufferers you will see an appreciable decline. One of the contributory factors was a major drop in lung and bronchus cancers between 1990 and 2010. This particular decline is partly due to research carried out in Britain during the early 1950s by Richard Doll and Austin Bradford-Hill. Doll and Bradford-Hill published two articles in the *British Medical Journal*[1, 2] which are now regarded as classics.

I am looking at both these works as I write this article. They are written in an almost Victorian font and layout that resembles that used in an anti-alcohol tract from a temperance society or in a prospectus for a crooked company that claimed to be able to bottle moonbeams. However, they are two remarkable publications: first, for their effect (they led to a decline in smoking habits that was a major factor in the decline in mortality); second, for the fact that they are simple and can be read by the non-medical, non-statistical specialist. They reflect a much less complicated world than the one we inhabit today. All Bradford-Hill and Doll did was look at mortality and smoking and use very simple statistical methods, normally found in second-year undergraduate courses, to establish a link between smoking and lung cancer and preclude other factors such as industrial pollution.

If we move forward to 2006 we see another world. A world where the human genome has been totally sequenced, where scientific research is inconceivable without a computer, and where huge amounts of data are being generated by computer-controlled equipment.

In 2006 a large amount of data was created by researchers at Duke University in the United States. They were engaged in research into an area of medical research known as personalised medicine. The main aim of their work was to establish whether a patient's genetic make-up can be used to identify therapeutic regimes that would provide better responses.

The Duke researchers used a collection of drug sensitivity data and the results from devices known as micro-arrays to predict the response of cancer sufferers to various chemotherapy treatments. The micro-arrays identified bio-markers which were correlated with response in easily available data from cell lines; these markers were then examined in patient samples in order to predict the best chemotherapy. In a November 2006 paper in *Nature Medicine*[3], the researchers claimed success.

This was a remarkable breakthrough. The paper that initially announced their results[3] was named as one of the top publications of 2006 by *Discover* magazine, and received broad publicity. For cancer patients it provided hope. In the past an initial chemotherapeutic regime might fail because of the patient's insensitivity to the regime and other regimes would need to be tried. For the cancer research and personalised medicine communities it was a much needed proof of concept. For Duke University it represented a revenue stream that might have generated hundreds of millions of dollars a year. For the junior researchers who authored the paper it would mean tenured posts. For the senior researchers it was a major step towards further rewards such as million-dollar personal research grants and commercial contracts with pharmaceutical companies.

The original research article was followed by others in gold-standard journals. The researcher responsible for most of the work, Anil Potti, was hailed by Duke as a rising star: a promotional video was made and he was regarded as an ambassador for his university. The only problem was that the reported research was flawed. The story of how this was discovered is one that has major import for universities, medical researchers, statisticians and academic journals.

The three people who can be credited with the discovery that the Duke research was built on mud are Keith Baggerly, Kevin Coombes and Paul Goldberg. Baggerly and Coombes are biostatisticians at the M.D. Anderson Cancer Center, located in Houston, Texas. Goldberg is the editor of *The Cancer Letter*, a publication targeted at cancer researchers, health care professionals and staff working in the pharmaceutical industry.

Clinical staff at M.D. Anderson were so excited by the work reported by the Duke University researchers that they asked Baggerly and Coombes to investigate. They asked the Duke researchers for their data and computer programs and set to work. Almost immediately they encountered difficulties. As an example, one of the many problems they encountered was that on November 26th, 2006 (to quote Baggerly's notes to an Institute of Medicine inquiry in 2011): "We checked the drug sensitivity data for the cell lines we inferred, and found that some 'sensitive' lines were more resistant than some 'resistant' lines and vice versa." Effectively, results opposite to those reported were exhibited by the Duke researchers' data. This was just one problem of many: Baggerly and Coombes discovered data they did not understand, mislabelled data and descriptions of idiosyncratic steps that they could not reproduce. When they applied their best understanding of the reported techniques to correctly labelled data, they obtained results no better than chance.

Baggerly and Coombes continually corresponded with the principal researchers at Duke and kept them informed of their worries. Throughout the two groups' interactions the Duke researchers claimed that their research worked, even though Baggerly and Coombes provided evidence that it had major problems.

As a researcher, when you find problems with reported research the first thing to do is to contact the original researchers with your worries. Baggerly and Coombes did this, but felt that they were not being listened to. As a consequence they issued a series of short communications to the journals that published the Duke work. The response of the journals was mixed. Some communications were published. One journal sent one of the communications to a single statistical referee who pronounced that the issue was one of statistical argument rather than anything more serious; another journal published a communication, then refused a second without documenting a reason; a third journal published a communication, and then refused a second on the grounds that their house rules forbid more than one communication on a particular piece of research.

When a journal publishes criticism of a paper, the original authors are generally allowed space to respond. What was revealing about one of the responses of the Duke researchers was their view

of reproducibility. The example below is from a 2008 issue of the *Journal of Clinical Oncology*.

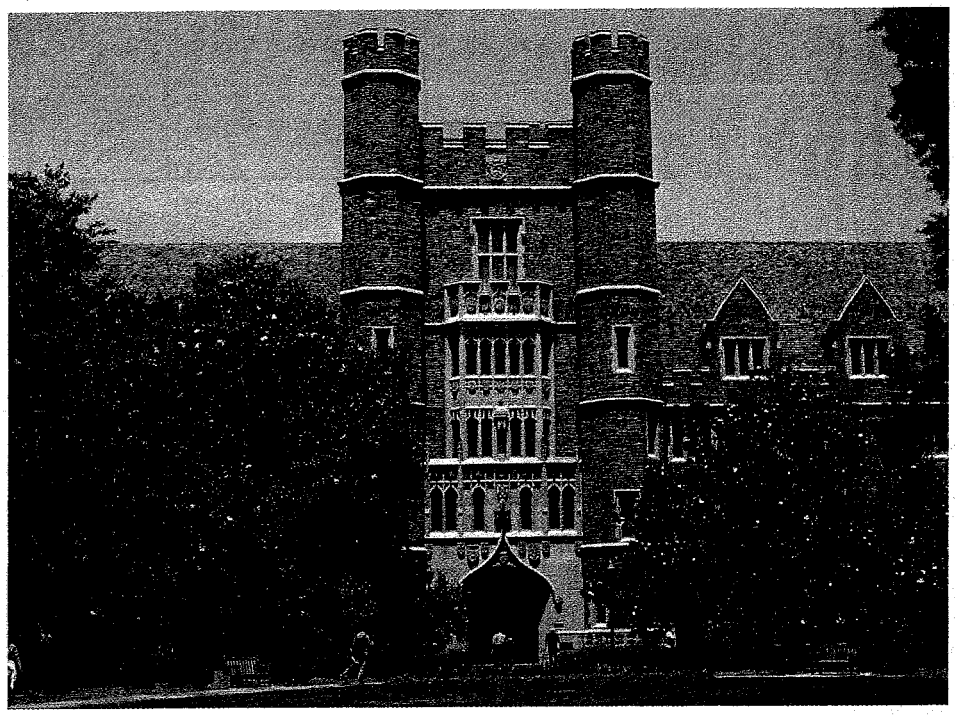
> A focus on these errors as presented by Baggerly et al is misleading since it suggests they are a contributing factor in the supposed lack of reproducibility, which is not the case. Most importantly, the claim that they cannot reproduce the results of the study, when in fact they did not even try to do so, is an egregious flaw in their commentary. To reproduce means to repeat, using the same methods of analysis as reported. It does not mean to attempt to achieve the same goal of the study but with different methods.

Baggerly and Coombes had tried a completely independent audit of the Duke research; this is standard statistical practice and something that the authors of the rebuttal had not grasped.

By 2008 there was a breakdown in the interaction between Baggerly and Coombes and the Duke group. In 2009 Baggerly and Coombes were shocked to discover that clinical trials (eventually involving 109 cancer patients) had been started back in 2007 based on research that they regarded as flawed, and were still actively recruiting patients. By this time they had grown tired of trying to convince the academic medical journals that there were major problems and so developed a paper for the *Annals of Applied Statistics*[4]. The paper was refereed and published very quickly. In it the authors detailed many of the problems that they had encountered and expressed concern that patients in the clinical trials may have been put at unnecessary risk.

As a result of the *Annals* paper, Duke University suspended the clinical trials and their Institutional Research Board (equivalent to a research ethics committee in a British university) commenced an inquiry in which two external statistical reviewers were asked to examine the data and the computer programs used by the Duke researchers. The inquiry, with a few caveats, passed the work as being valid. Unfortunately, at that stage the inquiry report was not published by Duke University. The clinical trials were then restarted. Given the criticisms that had been made by the M.D. Anderson researchers, the result was very surprising. What was also surprising about the review was that an important document written by Baggerly and Coombes, providing new extensive criticism of the Duke work, was sent to the management at Duke University but was not shown to the reviewers.

Fortunately, a number of events brought the whole issue to a head. First, staff at the National Cancer Institute in the US were unable to reproduce some of the Duke results; earlier some European researchers with whom the Duke staff had collaborated had similar problems. Second,



Entrance to the Medical Centre, Duke University. Medical research has changed since the days of mock-Gothic. Photo: Wikimedia Commons

Paul Goldberg, the editor of *The Cancer Letter*, obtained a redacted copy of the Duke report from the National Cancer Institute which raised issues of methodology. Third, *The Cancer Letter* discovered that Anil Potti, a key researcher in the Duke team, had falsified some aspects of his curriculum vitae. Fourth, 33 senior bio-statisticians, epidemiologists and clinicians wrote to the senior management of Duke University asking for the clinical trials to be suspended based on the fact that each signatory was concerned that "despite written statements from the external experts, who uniformly stated they were not given sufficient information to confirm the validity of the models, the trials have been re-initiated". Potti was put on administrative leave; he eventually resigned after admitting to problems with the data.

Quickly the reason for the positive outcome from the Duke review emerged. The data that was provided to the investigators had been subject to modification carried out in what the Duke management referred to as a 'non-random' way. The end result was that four journal papers (to date) were retracted: they became non-publications. Duke University started a research inquiry, and the Institute of Medicine in the US commenced a general inquiry into the level of evidence that should be required before "omics" based signatures are used to guide treatment in clinical trials.

So, over a period of four years, two statistical researchers attempted to convince both Duke University and a number of prestigious journals that there were serious defects in the research carried out at the university. Because Baggerly and Coombes were only provided with partial data (data that kept changing) they had to carry out a forensic process that took them between 150 to 200 days to complete. Clearly there were major problems which all the participants need to examine.

It was clear that there was a lot of sloppiness in data curation and software storage. A historical document provided to the Institute of Medicine review panel by the Duke management detailed a way forward (for both their university and others):

> Sustained statistical collaboration is critical to assure proper management of these complex datasets for translation to clinical utility as illustrated by the efforts of Dr. Barry to re-evaluate prior work without clear primary sources for the data, and records of the precise use of statistical methodologies and programs. The fundamental methods of managing data and validating statistical algorithms are not something basic scientists are generally familiar with, thus statisticians need to take an active role in participating in basic science research, both in terms of teaching research methods and in improving the design of studies.

Dr Barry is a statistician at Duke University who was tasked in the later stages of the scandal with examining the raw research data from the team. As the fragment above details, he encountered the same problems that Baggerly and Coombes discovered.

One message is clear: if you are going to do some sophisticated statistics then use a trained statistician. Doll and Bradford Hill worked at a time when scatter plots and histograms were the order of the day. Today's researchers deal with huge quantities of computer-generated data that can be highly stochastic, error-prone, multidimensional, and incapable of being intuitively interpreted. This is not a new message. Statisticians have been examining research output and expressing major worries about methods for a long time. For example, Doug Altman noted in 2000[5] that there was a shockingly high number of poor statistical analyses in a number of important medical journals.

His prescient message[5,6] was that in 2000 (when computers played a smaller part in the *generation* of research data): the misuse of statistics was very important; a general climate of sloppiness was bad for science; statistics was much more subjective (and difficult) than was usually acknowledged; major improvements in the quality of research published in medical journals were unlikely in the present research climate; and too much research was done primarily to benefit the careers of researchers.

An allied issue is that of reproducibility. The philosopher Karl Popper laid down ground rules for this. He wrote that a scientific theory only has any validity if it can be refuted. As long as it is not refuted it stays in the room and gets stronger as attempts at refutation fall by the wayside. Because Baggerly and Coombes were only provided with partial materials, they had to carry out a hugely time-consuming audit of the Duke research: they spent an inordinate amount of time trying to figure it out. Reproducibility was, perhaps, less of an issue in the 1980s and 1990s (Disclosure: I estimate that the research detailed in about 12 of my 120 publications over the last 35 years cannot be reproduced mainly because I no longer have the data). However, in this millennium it is vital.

There are a number of lessons that universities, journals and researchers need to learn. First, the universities. In his testimony to the Institute of Medicine inquiry, Baggerly stated that, with respect to the review that Duke carried out:

- the Duke reviewers did not verify the provenance of the data,
- the Duke report was not published,
- the Duke data were not released, and
- members of the Duke administration and Institutional Research Board withheld information (some of our reports) from the reviewers.

Consequently, said Baggerly, the review was neither complete nor transparent, but it was nonetheless used as the basis for restarting clinical trials.

When a university investigates a potential serious problem with research it should adopt a quasi-legal approach and ensure that it is as transparent as possible. This is the first lesson.

There is another lesson for universities. They should regard statistics in the same way as they regard support for computing. This means that they should set up specialist units that have the same status as the units that provide advice to staff on computing and programming issues and software problems. Some already do; many others need to do so.

The journals need to look to the future. Clearly their response to the communications from Baggerly and Coombes was patchy. Journals should treat disputes about published papers very seriously. One model is to employ an ombudsman. *The Lancet* has adopted this sort of model to deal mainly with complaints involving procedural issues such as the journal taking a long time to consider a submitted article. This role should be extended to disputes about published work with the full refereeing process being employed, even down to the running of code with experimental data.

Universities should support statistics as they support computing: specialist units should give advice on each

Journals should also adopt the rule that you do not get published unless data and software are lodged in a public repository independent of the researchers' institution. This independent lodging is important in that one of the problems that Baggerly and Coombes encountered was that the data stored at Duke was a moving target. Increasingly journals are adopting this rule with respect to data and some, including the prestigious journal *Science*, are also requiring that program code be made available on request. However, there is still a long way to go.

A very revealing survey of reproducibility was published in 2009, around the time that Baggerly and Coombes went public with their *Annals of Applied Statistics* paper, by John Ioannidis and colleagues[7]. They examined the reproducibility of research using micro-array technology (the very technology employed by the Duke researchers) reported in 18 articles in the journal *Nature Genetics*. They discovered only two could be reproduced in full, six could be reproduced in part or with some discrepancies and that ten could not be reproduced at all.

The publication of this article is to the huge credit of *Nature Genetics*.

There are also lessons for the research funders such as the National Science Foundation and the Medical Research Council. First, they should make it a condition of funding that all data and code are lodged with them and made publicly available on their websites. Failure to do this should result in funding streams being cut off from researchers who do not adhere.

There is also a second agenda that research funders should address. There is a lack of tools for packaging up data and programs. There are some exceptions, for example *Sweave* is a program that packages text written in the document processing language LaTeX with code written using the programming language R. However, a scientist would need quite a menu of technical skills to use it. The research funders, as a matter of urgency, should initiate research projects which have as their end-point the development of tools for reproducibility. They should also fund studies similar to that of Ioannidis[7] to examine how serious the problem of reproducibility is.

As I edited a final draft of this article I came across a staggering set of statistics in a research paper in the *Journal of Medical Ethics*[8]. In a terrific but depressing piece of research, Grant Steen showed that between 2000 and 2010 around 80 000 patients (a conservative estimate) had undergone clinical trials based on research that was incorrect and for which papers were retracted. I thought that the Duke problem was something of a one-off; Steen's article has disabused me of this notion.

My thanks to Keith Baggerly for help in writing this article.

References

1. Doll, R. and Bradford-Hill, A. (1950) Smoking and carcinoma of the lung. *British Medical Journal*, **2**, 739–748.
2. Doll, R. and Bradford-Hill, A. (1954) The mortality of doctors in relation to their smoking habits. *British Medical Journal*, **6**, 1451–1455.
3. Potti, A., Dressman, H. K. *et al.* (2006) Genomic signatures to guide the use of therapeutics. *Nature Medicine*, **12**, 1294–1300.
4. Baggerly, K. A. and Coombes, K. R. (2009) Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. *Annals of Applied Statistics*, **3**, 1309–1334.
5. Altman, D. G. (2000) Statistics in medical journals: Some recent trends. *Statistics in Medicine*, **19**, 3275–3289.
6. Altman, D. G. (1994) The scandal of poor medical research, *British Medical Journal*, **308**, 283–284.
7. Ioannidis, J. P. A., Allison, D. B. *et al.* (2009) Repeatability of published microarray gene expression analyses. *Nature Genetics*, **41**, 149–205.
8. Steen, R. G. (2011) Retractions in the medical literature: How many patients are put at risk by flawed research. *Journal of Medical Ethics*. doi:10.1136/jme.2011.043133.

Darrel Ince is Professor of Computing at the Open University. He is currently writing a book on the Duke scandal, entitled *The Cracks in Science*.