Statistical Methods for Pharmaceutical Research and Early Development

Lyon, France, September 27-29, 2010.

Improving statistical quality in published research: the clinical experience

Martin Bland

Professor of Health Statistics
University of York

http://martinbland.co.uk

A bit about myself

1969-1972: agricultural chemical industry, ICI.

Then clinical and epidemiological research.

1972-1976: St. Thomas's Hospital Medical School, London,

1976-2003: St. George's Hospital Medical School, London,

2003-now: Health Sciences, University of York.

What I intend to talk about

A very personal account of how clinical research has changed in the past 38 years.

- Key factors in this change,
- What I did to push things along,
- The situation in non-clinical biomedical research,
- Suggestions for the future.

Then and Now

The *Lancet* and the *British Medical Journal* from September 1972

Research reports which used individual subject data, excluding case reports and animal studies.

The *Lancet*: 31 reports, median sample size was 33 (quartiles 12 and 85).

The *British Medical Journal*: 30 reports, median sample size 37 (quartiles 12 and 158).

Bland JM. (2009) The tyranny of power: is there a better way to calculate sample size? *British Medical Journal* **339**: b3985.

Then and Now

The *Lancet* and the *British Medical Journal* from July 2010

Research reports which used individual subject data, excluding case reports and animal studies.

The *Lancet*: 16 reports, median sample size was 1,626 (quartiles 527 and 14,774).

The *British Medical Journal*: 15 reports, median sample size 10,170 (quartiles 234 and 48,649).

The sample size for studies in these journals has increased hugely.

Then and Now

Methods of statistical inference employed (including studies not on individual subjects)

September 1972, in the Abstracts of the papers:

The *Lancet*: in 39 papers, five mentioned P values or significance.

The *BMJ*: in 32 papers, four mentioned P values or significance.

Then and Now

Methods of statistical inference employed (including studies not on individual subjects)

September 1972, in the “Results” section of the papers:

The *Lancet*: 19 of 39 papers quoted the results of significance tests, either as P values or test statistics, and one gave confidence intervals in graphical form (Pollack *et al.* 1972).

The *BMJ*: 22 of 32 papers gave the results of significance tests, none at all presented confidence intervals.

Pollack M, Nieman RE, Reinhard.JA, Charache P, Jett MP, Hardy PH. (1972) Factors influencing colonisation and antibiotic-resistance patterns of gram-negative bacteria in hospital patients. *Lancet* **2**: 668-1.

Then and Now

In 1972, very little description of statistical methods appeared in “Methods” sections of the papers.

Three BMJ papers gave a reference for their statistical methods.

One of these merely noted that “Statistical analyses were performed using methods described by Snedecor (1956)” (Bottiger and Carlson 1972) .

A standard statistical textbook, already superseded by the 1967 edition.

Bottiger LE, Carlson LA . Relation between serum-cholesterol and triglyceride concentration and hemoglobin values in non-anemic healthy persons. *British Medical Journal* 1972; **3**: 731-3.

Then and Now

Snedecor (1956) was also cited by Ellis (1972).

Bishop *et al.*, (1972) quoted Dixon and Massey (1951) a book then more than twenty years old.

Bishop MC, Woods CG, Oliver DO, Ledingham JGG, Smith R, Tibbutt DA. (1972) Effects of haemodialysis on bone in chronic renal failure. *British Medical Journal* **3**: 664-667.

Ellis FR, Keaney NP, Harriman DGF, Sumner DW, Kyei-Mensah K, Tyrrell JH, Hargreaves JB, Parikh RK, Mulrooney PL (1972) Screening for malignant hyperpyrexia. *British Medical Journal* **3**: 559-561.

Then and Now

The *Lancet* and the *British Medical Journal* from July 2010

In both journals, all papers included statistical inference in the abstract.

The *Lancet*: 15 of the 16 papers had confidence intervals and 8 had P values.

The BMJ: 13 of the 15 had confidence intervals, 7 had P values.

So we have much greater sample sizes and much greater prominence for statistics in the papers.

We also have a clear change of emphasis, from significance testing to estimation.

What Happened?

Several initiatives might have contributed to this change.

They are not independent things, but different aspects of the same drive.

Often it is hard to say exactly when these movements began, because a lot of people were involved in them.

What Happened?

Evidence-based medicine

Treatment decisions should be based on objective evidence rather than the evidence of experience and authority.

Such evidence was going to include statistics.

Use of this term began in the 1990s, but the ideas were around long before.

A doctor-led movement (e.g. Dave Sackett and Gordon Guyatt at McMaster University).

Statisticians, as people whose business was the evaluation of evidence, were enthusiastic cheerleaders.

What Happened?

Systematic review

Collect together all the trials which had been carried out of a therapy and try to form a conclusion about effectiveness.

Iain Chalmers led a huge project to assemble all the trials ever done in obstetrics (Chalmers *et al.*, 1989).

The Cochrane Collaboration aims to do the same for all of medicine.

Chalmers I, Enkin M, Keirse MJNC. (eds) *Effective Cure in Pregnancy and Childbirth*, Oxford University Press, Oxford, 1989.

What Happened?

Systematic review

A doctor-led initiative.

Statisticians were enthusiastic supporters, developing methods of data synthesis to combine the results of these trials where possible.

Richard Peto: expert opinion on three approaches to the treatment of myocardial infarction, as expressed in leading articles in the *New England Journal of Medicine* and the *Lancet*. Contrasted with the exactly opposite conclusions which he had drawn from a systematic review of all published randomised trials in these areas.

What Happened?

Large simple trials

Alternative solution to the problem of inadequate sample sizes.

Richard Peto (Peto and Yusuf 1981) led the call for large, simple trials, the first being ISIS-1 (ISIS-1 Collaborative Group, 1986).

Peto R, Yusuf S. (1981) Need for large (but simple) trials. *Thrombosis and Haemostasis* **46**: 325-325.

ISIS-1 (First International Study of Infarct Survival) Collaborative Group. (1986) Randomized trial of intravenous atenolol among 16,027 cases of suspected acute myocardial infarction. ISIS-l. *Lancet* ii: 57-66.

What Happened?

Large simple trials

This was spectacularly successful (Peto *et al*. 1995).

Probably explains the great increase in sample size reported from 1972 to the present.

No clinical researcher with aspirations to be in the top flight can now be happy unless a trial with a four-figure sample size is in progress.

Peto R, Collins R, Gray R. (1995) Large-scale randomized evidence: large, simple trials and overviews of trials. *Journal of Clinical Epidemiology* **48**: 23-40.

What Happened?

Confidence intervals not P values

A very statistically-led movement was to present inference using confidence intervals rather than significance tests.

Gardner and Altman (1986) was a very important paper in this, which led to the *British Medical Journal* including this in its instructions for authors.

Other journals, such as the *Lancet*, followed suit.

Gardner MJ and Altman DG. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. *British Medical Journal 292*: 746-50.

What Happened?

Quality assessments in journals

There is a long history of articles criticising the quality of statistics in medical journals, but these mostly come from the mid-sixties onwards (Altman,1991).

Altman (1981) was an important article calling for improvement.

These articles began to sting to journal editors into action and led to instructions to authors about statistical aspects of presentation of results.

Altman DG. (1981) Statistics and ethics in medical-research. 8. Improving the quality of statistics in medical journals. *British Medical Journal* **282**: 44-47.

Altman DG. (1991) Statistics in medical journals - developments in the 1980s. *Statistics in Medicine* **10**: 1897-1913.

What Happened?

Statistical referees

Following reviews of statistics, journals began to introduce statistical referees.

The systematic use of a panel of statisticians to referee all research papers before they appeared in the journal.

The main difficulty is finding enough statisticians.

What Happened?

The CONSORT statement

First published in 1996 (Begg *et al.*, 1996).

Guidelines for reporting trials, encouraging researchers to provide information about methods of determining sample size, allocation to treatments, blinding, statistical analysis, etc.

Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF, Simel D, Stroup DF. (1996) Improving the quality of reporting of randomized controlled trials - The CONSORT statement. *JAMA-Journal of the American Medical Association* **276**, 637-639.

What Happened?

The CONSORT statement

Since been updated (Moher *et al.*, 2001) and produced several variations and imitators.

It has now been adopted by many journals as part of their instructions to authors.

Moher D, Schulz KF, Altman DG, CONSORT Group. (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* **357**: 1191-4.

What Happened?

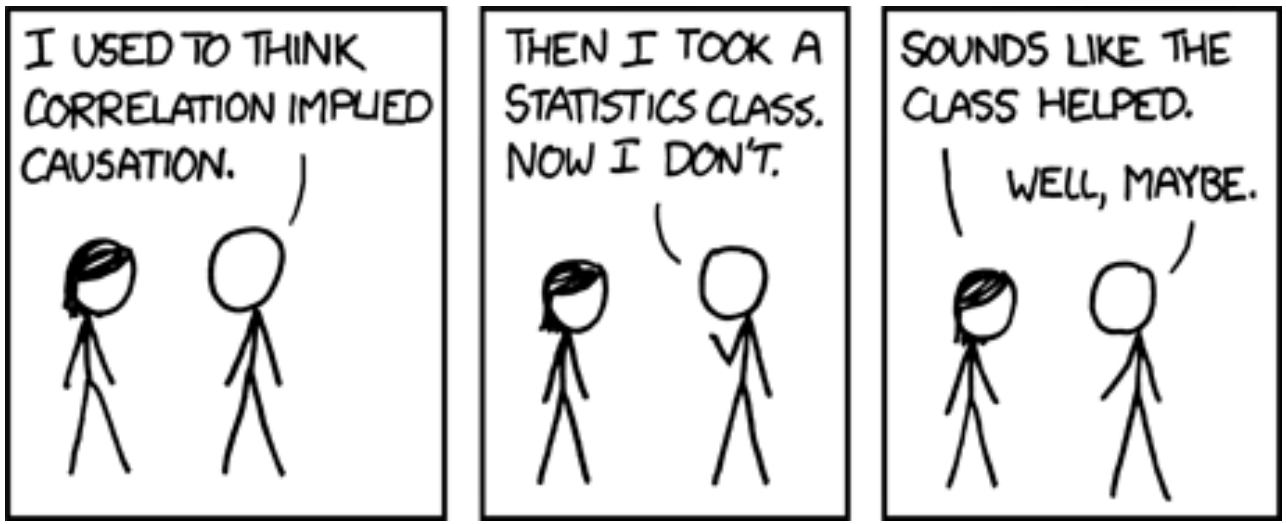
Post Hoc Ergo Propter Hoc?

“After this therefore because of this” — a logical fallacy.

What Happened?

Post Hoc Ergo Propter Hoc?

“After this therefore because of this” — a logical fallacy.

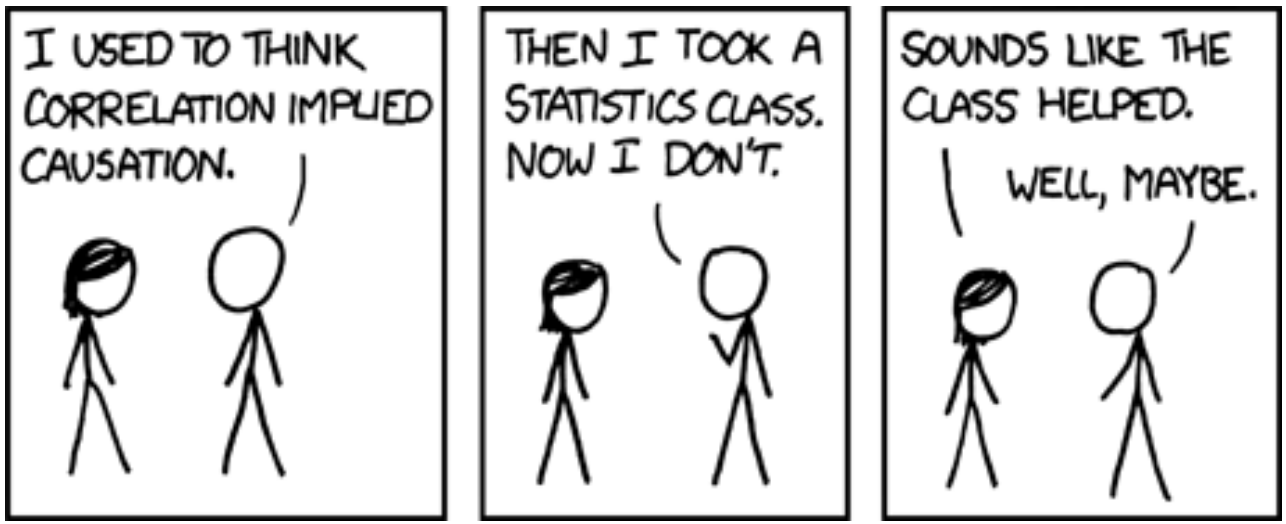


http://xkcd.com/552/

What Happened?

Post Hoc Ergo Propter Hoc?

“After this therefore because of this” — a logical fallacy.



http://xkcd.com/552/

We cannot know which, if any, of these forces is responsible for improvements in the statistical quality of the elite clinical literature.

My rôle in the campaign

(What did you do in the war, Daddy?)

My rôle in the campaign

Confidence intervals

Meeting of teachers of statistics in medical schools.

Core curriculum for medical statistics.

Reported our conclusions about t tests and chi-squared tests.

My rôle in the campaign

Confidence intervals

Meeting of teachers of statistics in medical schools.

Core curriculum for medical statistics.

Reported our conclusions about t tests and chi-squared tests.

Demolished by David Clayton, who said that what he wanted students to learn was how to make estimates about the world and put confidence intervals around them.

I saw that he was right.

My rôle in the campaign

Confidence intervals

I redesigned my courses to put estimation first.

From then on, in analyses carried out for researchers I stressed confidence intervals.

When my text book *An Introduction to Medical Statistics* (Bland 1987) appeared, the chapter introducing confidence intervals came before that introducing significance tests, and their superiority was emphasised.

Bland M. (1987) *An Introduction to Medical Statistics*. Oxford University Press, Oxford.

My rôle in the campaign

Statistical mistakes in journals

I wrote letters to journals when I saw blatant mistakes in statistical analysis.

Sometimes the letters were published (e.g. Bland and Altman 1977) .

Occasionally these mistakes were accepted by the authors, more often not.

The letters made the point to discourage future authors from copying flawed methods and interpretations.

Bland JM, Altman DG. (1977) Enteric disease in San Francisco. *Lancet* 2, 306.

My rôle in the campaign

Statistics Notes in the *BMJ*

Doug Altman and I wrote Statistics Notes in the *British Medical Journal*.

These began in 1994 (Bland and Altman 1994) and continued sporadically ever since.

55 published, with six other occasional authors.

Mean of 115 citations by July 2010, a total of 6,337.

Bland JM, Altman DG. (1994) Statistics Notes. Correlation, regression and repeated data. *British Medical Journal*, 308, 896.

My rôle in the campaign

Review committees

Several grant funding bodies and an ethics committee.

On these I stressed the importance of correct statistical design and analysis.

My rôle in the campaign

Review committees

Medical Research Council project board for health services and public health research.

First meeting: a bid for a cluster-randomised trial.

The sample size calculations and proposed statistical analysis they did not take any account of the clustering.

The trial would be underpowered, any P values would be too liberal and confidence intervals too narrow.

My rôle in the campaign

Review committees

The trial would be underpowered, any P values would be too liberal and confidence intervals too narrow.

I explained this to the board and the proposal was rejected.

At the next meeting, the same thing happened again.

Meeting followed meeting.

Then a change occurred.

My rôle in the campaign

Review committees

Cluster randomised trials started being described as such and coming with estimates of intra-cluster correlation coefficients and proposals for multilevel modelling.

I wondered what change had taken place in the world, without me knowing.

My rôle in the campaign

Review committees

Cluster randomised trials started being described as such and coming with estimates of intra-cluster correlation coefficients and proposals for multilevel modelling.

I wondered what change had taken place in the world, without me knowing.

The MRC secretariat were warning applicants that cluster randomised trials which ignored the clustering would not get past the board.

They should find a statistician who understood these things.

My rôle in the campaign

Review committees

Cluster randomised trials started being described as such and coming with estimates of intra-cluster correlation coefficients and proposals for multilevel modelling.

I wondered what change had taken place in the world, without me knowing.

The MRC secretariat were warning applicants that cluster randomised trials which ignored the clustering would not get past the board.

They should find a statistician who understood these things.

My most effective piece of statistical education.

Is it all over?

Clinical research is not statistically flawless.

Things are much better in the major journals.

In the specialist clinical journals things can go on much as before.

Is it all over?

An example: Boots “anti-aging” cream trial

(Watson *et al.* 2009)

Trial received wide media publicity as the first “anti-aging” cream proven to work in a randomised controlled clinical trial.

60 volunteers were randomised in groups of 30 to either the “anti-aging” product or the vehicle without the active ingredient for six months, followed by the “anti-aging” product for a further six months.

Watson REB, Ogden S, Cotterell LF, Bowden JJ, Bastrilles JY, Long SP, Griffiths CEM. A cosmetic ‘anti-ageing’ product improves photoaged skin: a double-blind, randomized controlled trial. *British Journal of Dermatology* 2009: DOI 10.1111/j.1365-2133.2009.09216.x

Is it all over?

An example: Boots “anti-aging” cream trial

Reported that after six months 43% of participants receiving the “anti-aging” cream had improved appearance of wrinkles, compared to 22% of those receiving the placebo.

This was what was picked up by the media.

The authors report four outcome measures: fine lines and wrinkles, dyspigmentation, overall clinical grade of photoageing, and tactile roughness, each measured on a scale of 0 to 8 at baseline, 1, 6, and 12 months.

Is it all over?

An example: Boots “anti-aging” cream trial

No mention that any of the 4 measures was a prespecified primary outcome.

We might surmise that a significant difference in any variable would be taken to indicate evidence of a treatment effect.

The trial was entirely analysed in terms of P values, so prudence should lead us to adjust for multiple testing.

Bonferroni correction: multiply P values by 4.

If we were to include the 6 and 12 months results in the same analysis, we would multiply by 8.

Is it all over?

An example: Boots “anti-aging” cream trial

For wrinkles at six months, the authors gave the results of significance tests comparing the score with baseline for each group separately, reporting the active treatment group to have a significant difference and the vehicle group not.

This is a classic statistical mistake.

The difference within a group being not significant does mot imply that there was no difference or tell us much about the size of any difference that might exist.

We should compare the two groups directly.

Is it all over?

An example: Boots “anti-aging” cream trial

The paper includes some data for the improvement in each group, 43% for the active group and 22% for controls, as picked up by the media.

No P value is given, but in the discussion the authors acknowledge that this difference was not significant.

Is it all over?

An example: Boots “anti-aging” cream trial

The *British Journal of Dermatology* published my letter (Bland 2009b).

A different version subsequently appeared in *Significance* (Bland 2009c).

This happened, of course, only because the publicity generated by Boots brought the paper to my attention.

Bland JM. (2009b) Evidence for an ‘anti-ageing’ product may not be so clear as it appears. *British Journal of Dermatology* **161**, pp1207–1208.

Bland M. (2009c) Keep young and beautiful: evidence for an "anti-aging" product? *Significance* **6**, 182-183.

Non-clinical biomedical research

Often remarked that laboratory research is the next area for statisticians to become involved.

Research scientists do their own statistics and often do them badly.

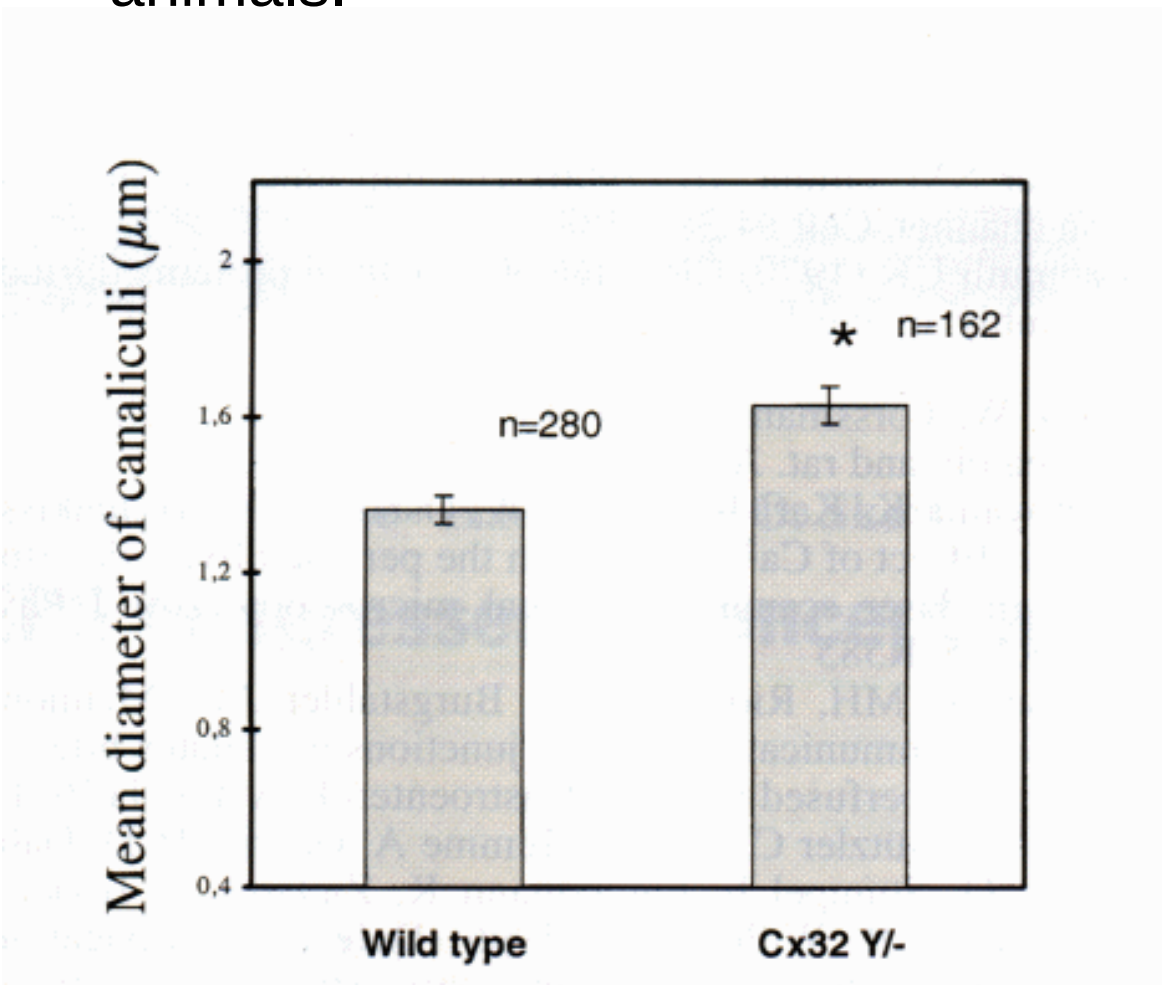
An example: Temme *et al.* (2001).

Compared two genetic strains of mice, wild-type and connexin32-deficient.

Temme A, Stumpel F, Rieber GSEP, Willecke KJK, Ott T. (2001) Dilated bile canaliculi and attenuated decrease of nerve-dependent bile secretion in connexin32-deficient mouse liver. *Eur J Physiol* **442**, 961-966. .

Non-clinical biomedical research

Temme *et al.* (2001) measured the diameters of bile canaliculi in the livers of wild-type and C02-deficient animals.



Non-clinical biomedical research

Temme *et al.* (2001) measured the diameters of bile canaliculi in the livers of wild-type and C02-deficient animals.

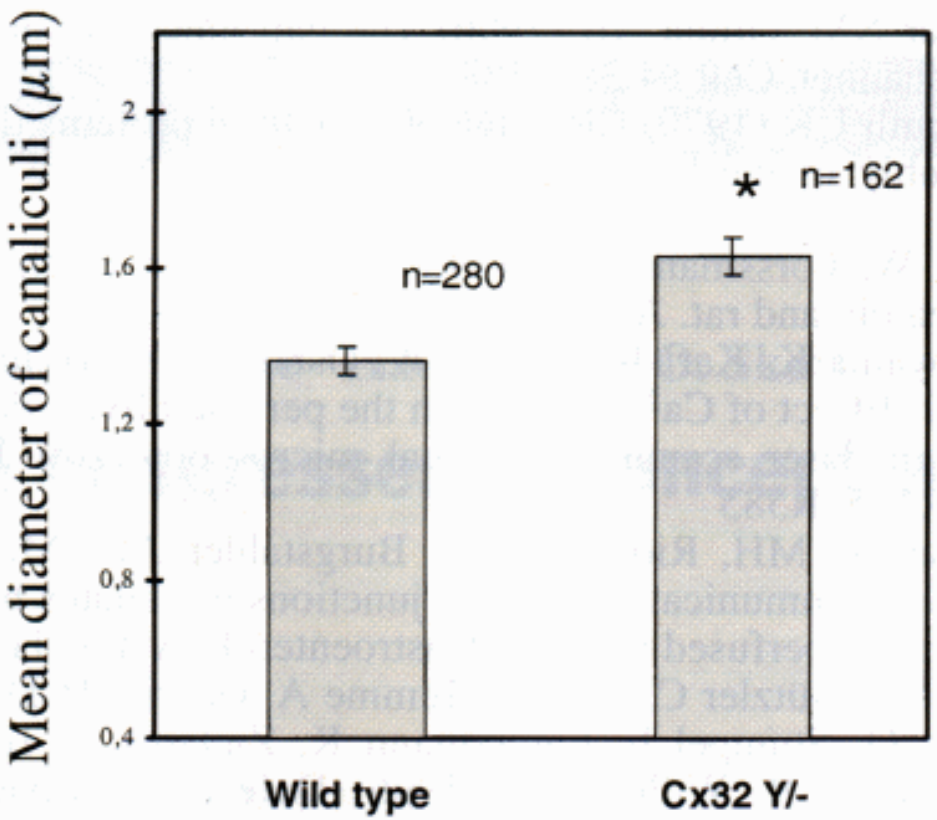


Fig. 3. Morphometric analysis of the diameter of bile canaliculi in wild-type and C02-deficient liver. Means±SEM from three livers. *P<0.005

Non-clinical biomedical research

Temme *et al.* (2001) measured the diameters of bile canaliculi in the livers of wild-type and C02-deficient animals.

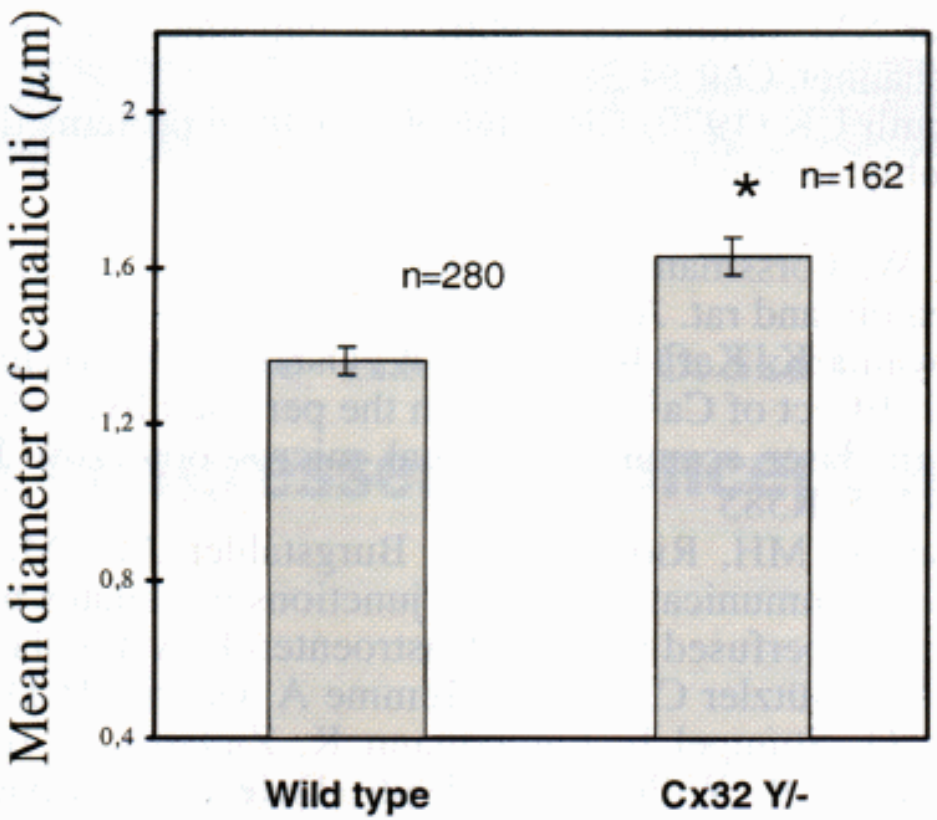


Fig. 3. Morphometric analysis of the diameter of bile canaliculi in wild-type and C02-deficient liver. Means±SEM from three livers. *P<0.005

I think there is a fairly obvious problem with the units of analysis here.

Non-clinical biomedical research

Kilkenny *et al.* (2009)

Review of reporting, experimental design and statistical analysis in published biomedical research using laboratory animals.

Analysed 271 publications.

Reported that in only 59% the hypothesis or objective of the study and the number and characteristics of the animals used were reported.

Kilkenny, C., Parsons, N., Kadyszewski, E., Festing, MFW., Cuthill, IC., Fry, D., Hutton, J., Altman, DG. (2009) Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS ONE 4(11)*, e7824.

Non-clinical biomedical research

Kilkenny *et al.* (2009)

Most of the papers surveyed did not use randomisation (87%) or blinding (86%).

Only 70% of the publications that used statistical methods described their methods and presented the results with a measure of error or variability.

What next?

Our best allies are journal editors.

Once they are convinced that there is a serious problem, they usually want to do something about it.

Reviews of statistics used in particular journals are a good starting point.

Quite easy to do, best done more by than one statistician independently.

They give a statistical publication.

What next?

Our best allies are journal editors.

Once they are convinced that there is a serious problem, they usually want to do something about it.

Reviews of statistics used in particular journals are a good starting point.

Quite easy to do, best done more by than one statistician independently.

They give a statistical publication.

Jeremy Miles reviewed two psychological journals and found two instances of “P<0.0”.

Miles JNV, Hempel S. (2005) The presentation of statistics in clinical and health psychology research. In: *Proceedings of the British Psychological Society*, **13**, 185.

What next?

Case studies of examples where wrong conclusions have been drawn as a result of statistical mistakes provide very powerful evidence, if you can find them.

When you do see mistakes in published research, write a letter to the journal.

Harry them!

What next?

Finally, be positive.

We want to help.

Try offering statistics articles.

I think a few on the benefits of randomisation and blinding would be good starting point.

Statistical Methods for Pharmaceutical Research and Early Development

Lyon, France, September 27-29, 2010.

Improving statistical quality in published research: the clinical experience

Martin Bland

Professor of Health Statistics
University of York

http://martinbland.co.uk