What not to do in medical statistics

Neal Alexander[1]

O que não fazer em estatística médica

[1]Infectious Disease Epidemiology Unit. London School of Hygiene and Tropical Medicine. Keppel Street, London WC1E 7HT, United Kingdom. E-mail: neal.alexander@lshtm.ac.uk

Abstract

There have been major efforts to improve the application of statistical methods in medical research, although some errors and misconceptions persist. In this paper I will review some of the topics which most often cause problems: a) comparison of two methods of clinical measurement; b) comparison of baseline values between arms of a randomized trial; c) absence of evidence as opposed to evidence of absence; and d) regression to the mean. I will also revisit a statistical error in one of my own publications. I review some causes of the continuing misuse of statistics, and make some suggestions for modifying the education of statistical and non-statistical medical researchers in order to alleviate this.

Key words *Statistics, Biostatistics*

Resumo

Tem havido grandes esforços na aplicação de métodos estatísticos na pesquisa médica, embora algumas concepções equivocadas ainda persistam. No presente artigo faz-se uma revisão de alguns tópicos que frequentemente causam problemas: a) comparação de dois métodos de medidas clínicas; b) comparação de valores de base entre os braços de um ensaio randomizado; c) ausência de evidência em oposição a evidência de ausência; e d) regressão à média. Uma revisita aos erros estatísticos em uma de minhas próprias publicações também é feita. Foi feita a revisão de algumas causas do uso inadequado da estatística, assim como algumas sugestões são dadas para modificar a formação de pesquisadores médicos estatísticos e não estatísticos.

Palavras-chave *Estatística, Bioestatística*

Introduction

The quality of statistics in medical research has received much attention over the past twenty years. Many books and journal articles have tried to improve statistical understanding and practice, and journal editors have placed greater emphasis on these aspects. An illustrative episode was the enhancement of statistical review at The Lancet following the controversy, and tragic fallout, of a report it had published of poor survival in patients attending the Bristol Cancer Help Centre.[1,2]

In recent years, usage of statistical methods seems to have improved, although errors persist.[3] In this paper I will review some of the more common of these. For each of them I will attempt to explain the nature of the error, and suggest a valid alternative. However, continued improvement in published statistical analyses will require more than continued explication of correct methods. I will suggest that one of the obstacles to improving practice is the poor rapport which often pertains between statistical and non-statistical researchers. Accordingly, I want to present more than a list of errors. So I will include a description of a method which is sometimes said to be erroneous but in fact is valid (although suboptimal). And I will also present an error I made in one of my own publications.

Quantifying agreement between two methods of clinical measurement

As medical technology develops, there is often a need to compare two methods of measuring the same quantity. We may also need to evaluate the agreement between replicates made by the same observer (repeatability) or in different laboratories (reproducibility).[4]

Example: blood pressure measured by arm cuff and finger monitor

As an example of the comparison of two methods of measurement, we will consider a dataset of two methods of measuring blood pressure. Two hundred people had their systolic blood pressure measured once using a standard arm cuff, and once using a finger monitor.[5] Figure 1a shows a common, but incorrect, approach to the analysis of such data. One set of measurements is plotted against the other, and a correlation coefficient is calculated. The corresponding *p* value is often also calculated and, if this is small, the conclusion may be reached that the two methods have 'significant agreement'.[6-8]

Why a correlation coefficient does not measure agreement

We can start by asking ourselves what is the meaning of the *p* value (<0.0001) calculated for Figure 1a. In general, a small *p* value means that the test statistic is larger than would be expected by chance. The exact meaning of 'by chance' is defined by the null hypothesis of the statistical test being done. In the current example, the null hypothesis is that the two methods of measurement are unrelated. Since the two methods are designed to measure the same quantity (systolic blood pressure) it would be remarkable if they were completely unrelated. However, some kind of relation between them does not mean they are interchangeable. For example, if a nutritionist can guess a person's weight to within, say, 5 kg, then their estimates will be correlated with the measurements of a set of scales, but this does not mean that the scales can be dispensed with. Rather, we would need to know the number of kilograms, or mmHg in the blood pressure example, within which the two methods agree. This cannot be inferred from the correlation coefficient or *p* value alone.

One specific problem with the correlation coefficient is that its magnitude depends on the range of the data. Suppose that agreement between two methods is constant over the data range, in the sense that the average difference between them is constant. Suppose we split the data into two halves according to whether they are above or below the median value, and then calculate a correlation coefficient for each half. Each of these two correlation coefficients will be smaller than the correlation coefficient for the complete data set, even though the degree of agreement is (by assumption) the same throughout the data range. This shows that the correlation coefficient does not measure the degree of agreement. Another way to think of this is to imagine a new data point, consistent with the original data but at a much lower or much higher value. Inclusion of this data point will make the correlation coefficient increase in magnitude, and the *p* value decrease, even though the additional data point is reflecting the same relationship as the original data.

In passing, we may note another problem with the approach shown in Figure 1a. A regression line is often included, even though this breaks the symmetry between the two variables. This is because the results of the regression depend on which variable is chosen as the outcome, and which as the

predictor. But there is no reason to prefer one variable over the other for either of those roles. Unlike correlation, regression is not symmetric in terms of its two input variables.

Correct approach: Bland and Altman method

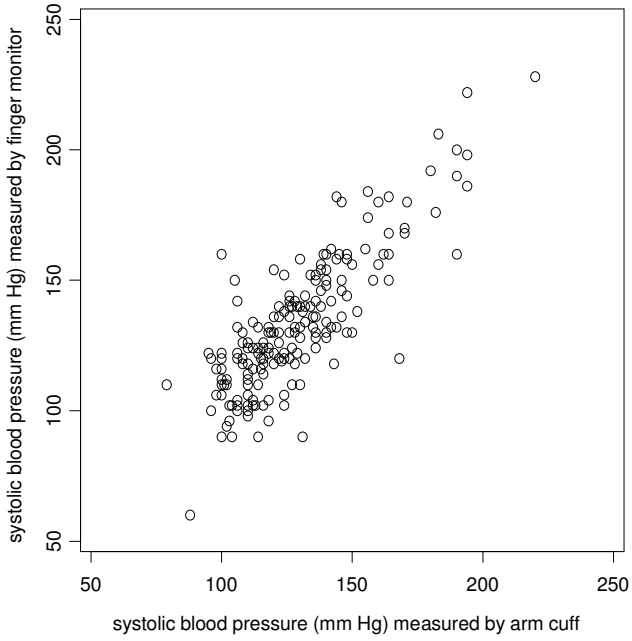
A meaningful assessment of agreement will be expressed in terms of the measurement units (mmHg in our example), rather than a correlation coefficient or *p* value. This is shown in Figure 1b, in which the vertical axis is the difference between the two measurements, and the horizontal axis is their average. This shows whether the size of the between-method difference changes with the magnitude of the quantity being measured. In our example, there is no sign that this is the case: the scatter on the vertical axis does not seem to increase or decrease depending on the value of the horizontal axis.

We can calculate the mean difference as a measure of whether there is a systematic difference between the methods. In Figure 1b, the mean difference is -4.3 mmHg, ie the arm cuff reads, on average, 4.3 mmHg lower than the finger monitor. This is shown as the central of the three dashed horizontal lines. The variation in agreement can be measured by the standard deviation of the differences. In our example this is 14.6 mmHg. If the differences have an approximately Gaussian ('normal') distribution, then 95% of them will lie in the range between the mean and plus or minus 1.96 times the standard deviation. In our data this range is from -32.9 mmHg to +24.3 mmHg. In other words, we can expect that, on 95% of occasions, the arm cuff measure between 32.9 mmHg less, and 24.3 mmHg more, than the finger monitor. The ends of this range are called the limits of agreement. Note that they are in the units of the original measurement (mmHg), which allows the degree of agreement to be judged clinically.

The simple technique described above was

Figure 1a

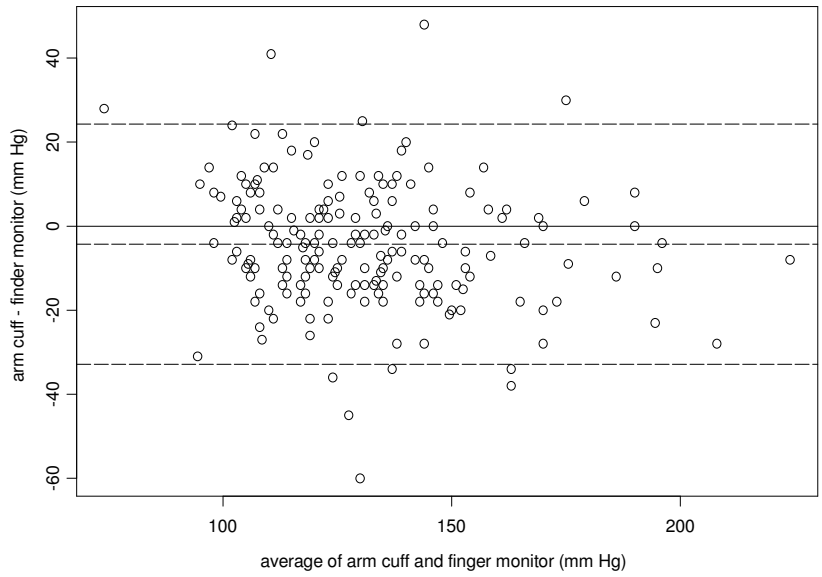
Assessing agreement of arm cuff and finger monitor in measurement of systolic blood pressure.



Finger monitor results plotted against those from arm cuff. This is the basis of the incorrect technique for measuring agreement: calculating a correlation coefficient and p value (in this example 0.83 and <0.0001, respectively).

Figure 1b

Bland and Altman plot of the difference between the two methods (arm cuff minus finger monitor), *versus* their average.



The central horizontal dashed line is the mean difference (-4.3 mmHg). The other horizontal dashed lines are the limits of agreement: 24.3 mmHg and-32.9 mmHg. These limits are equal to the mean difference plus and minus 1.96 times the standard deviation of the differences. The difference between the two methods will lie between these limits on 95% of occasions.

proposed by Bland and Altman,[10] whose papers have become two of the most cited in the medical literature.[9-11]

The choice of axes in Figure 1b ensures a lack of correlation between them. It may be tempting to plot the difference against a single one of the methods, especially if one of them is an accepted standard, but doing so will introduce a spurious correlation.[5] It can be shown mathematically that, if we denote the two methods x_1 and x_2, then x_1-x_2 is uncorrelated with x_1+x_2 but is intrinsically correlated with either x_1 or x_2.

It is often the case that the absolute difference between methods is larger for higher values of the actual measurements. In such cases the plot will show a greater scatter to the right of the horizontal axis, and the limits of agreement will not be applicable over the whole data range. Repeating the Bland and Altman technique on the logarithms of the values may resolve the problem. This yields results in terms of ratios rather than differences. However, if the data are counts, eg of parasites or CD4 cells, then a square root transformation may be effective.[12]

Hypothesis tests of baseline variables

It is common to use statistical hypothesis tests to compare baseline variables between the arms of a trial, and use the results to assess how 'successful' was the randomization. For example, Bassuk *et al.*[13] carried out a trial of antioxidants for prevention of cardiovascular events in 8171 female health professionals. The authors tested each of 28 baseline variables three times, comparing each intervention (vitamin E, vitamin C and beta-carotene) to its respective placebo. Of the 84 hypothesis tests done, 8 p values were less than 5%. The authors say this number was 'low, and no greater than what would be expected by chance'. They concluded "The randomization was successful, as evidenced by similar

distributions of baseline demographic, health, and behavioral characteristics across treatment groups."

However, the use of hypothesis tests to try to measure the success of a randomization is illogical, and may distract attention from a real problem. To see why it is illogical, we should again ask ourselves: what is the null hypothesis? Here, the null hypothesis is that the observed between-arm difference in the baseline variable was due to chance. However, because the trial was randomized, we know that a difference in a baseline variable was due to chance. In other words, the randomization ensures that the null hypothesis is true. The only reason to use a hypothesis test to do a between-arm comparison of a baseline variable is if one doubts that the randomization was done correctly.[14,15] The investigators should not have reason to doubt this!

Some authors use this kind of hypothesis test to identify variables to adjust for in later analysis. For example, Ellis *et al*.[16] did a randomized trial of the effect of an educational booklet on women's willingness to participate in a trial of treatment for breast cancer. They did hypothesis tests of 16 baseline variables. Two variables, anxiety and depression, had *p* values less than 5%, and these were included in subsequent multivariable analysis. However, as Assmann *et al*.[17] point out: 'A significant imbalance will not matter if a factor does not predict outcome, whereas a non-significant imbalance can benefit from covariate adjustment if the factor is a strong predictor'. In other words, hypothesis tests are not a suitable basis to decide the variables for which to adjust.

This does not mean that baseline values should not be reported, only that hypothesis tests of them are 'philosophically unsound, of no practical value and potentially misleading'.[15] The importance of baseline imbalance should be assessed according to the size of difference, and degree of association of the variable with the outcome. Baseline imbalance is unlikely to be a problem except in small trials. However, if there are any variables considered to be strong predictors of the outcome, they can be adjusted for, and this should be specified in advance in the analysis plan.[18] It is not advisable to adjust for many variables in the primary analysis, because that may decrease the precision of the between-arm comparison.[19] In fact, there may be no strong reason for adjusting for any of the baseline variables. In that case, it is often advisable for the primary analysis to be unadjusted.[18]

Absence of evidence is not evidence of absence

If a study fails to find a statistically significant effect ($p > 0.05$), it is tempting to conclude that the intervention does not work. However, this is not necessarily a valid conclusion. A result which is not statistically significant is, in itself, an absence of evidence: this is not the same as evidence of absence.[20] As we saw in previous sections of the current paper, we should think not only about the *p* value but also the magnitude of effect. For example, a study with a very small sample size would be able to detect only a very large effect. In other words, a *p* value larger than 0.05 may be due to insufficient sample size, rather than a small effect. The easiest way to think about the effect size is via confidence intervals (usually 95% confidence intervals). A 95% confidence interval for a parameter means a range which we are 95% confidence contains the true parameter value.

As an example, we can consider the trial of a) behaviour change interventions; b) syndromic management as methods of reducing HIV transmission, done in rural Uganda by Kamali *et al*.[21] The two interventions had incidence rate ratios of HIV (relative to control) of 0.94 and 1.00 respectively, with *p* values of 0.72 and 0.98. The authors concluded that "The interventions we used were insufficient to reduce HIV-1 incidence". However, the confidence intervals for these rate ratios show that the study cannot rule out a useful benefit of either intervention.[22] The 95% confidence interval for the rate ratio for behaviour change interventions was 0.60-1.45, and for syndromic management it was 0.63-1.58. So, for example, the behaviour change interventions may be capable of reducing HIV incidence by as much as 40%. They may also increase the incidence by as much as 45%: either way, we cannot tell from this study.

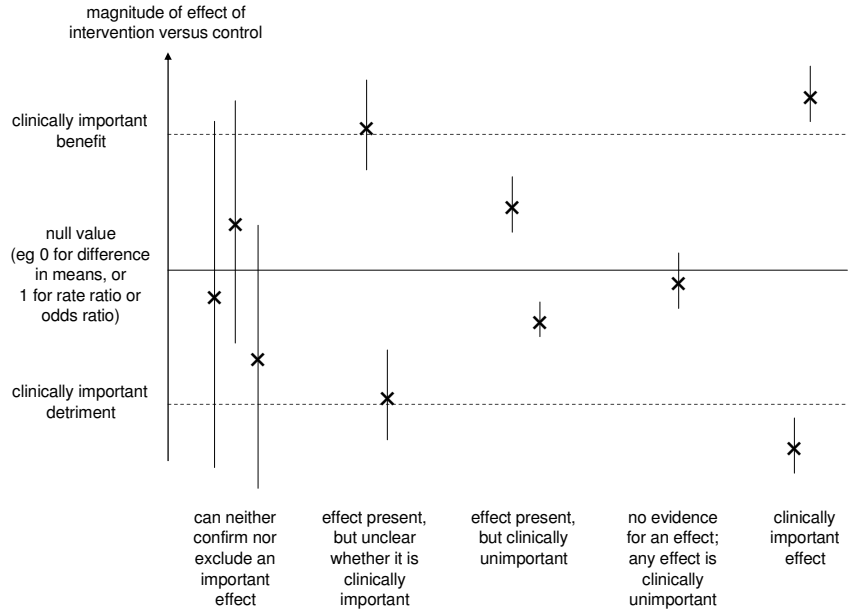
This and other scenarios are shown in Figure 2. Our conclusion from a particular study should depend not only on the *p* value but also on whether the confidence interval includes clinically important effect sizes. We can say that a reduction of HIV incidence by 40% would be clinically important. So, based on its confidence interval, we can place the Kamali *et al*.[21] trial in the leftmost category of Figure 2: 'can neither confirm nor exclude an important effect'.

Regression to the mean and change from baseline

Regression to the mean is a phenomenon which was studied in the early days of medical statistics. Francis Galton analysed the heights of parents and

Figure 2

Using the confidence interval of an intervention effect to reach a conclusion on its clinical importance.



The vertical axis shows the magnitude of effect, with larger benefits towards the top, and larger detriments towards the bottom. The vertical lines show confidence intervals from hypothetical studies. Along the bottom of the figure are conclusions which can be drawn from each group of confidence intervals.

children, and found that unusually tall parents tended to have children who were also tall, but not quite as tall as their parents. Similarly, short parents tended to have children who were not quite as short as they were: in other words, closer to the average. This phenomenon was called regression to the mean.[23] The word "regression" came to be applied not only to this phenomenon, but the statistical technique for quantifying it, which soon proved to be applicable to other problems as well. The two senses of the word can still cause confusion.

One can think of regression to the mean as resulting from of a kind of selection bias. Everitt[24] describes it as 'the phenomenon that a variable that is extreme on its first measurement will tend to be closer to the centre of the distribution for a later measurement'. Problems can arise when one forgets that a criterion for being in the dataset was an initial extreme measurement. A sporting example may be helpful. In British football, there are awards for the best manager (coach) of the month. Some journalists talk about the 'curse of manager of the month'. This means that a coach who receives the award one month tends to do badly the next month. However, one should bear in mind that, by definition, a winner of the award has done exceptionally well, and a performance level which is difficult to attain is even more difficult to maintain. Analysis of points per game shows that coaches whose teams win in one month usually fall back slightly in the subsequent month, but still do very well.[25]

In medical research, failure to account for regression to the mean can lead to several types of problem.[26,27] For example, when planning studies, it is common to seek populations in which the levels of the condition are high. If this varies over time then it is likely that an extremely high level will be followed by ones which are not quite as high. This may mean that power calculations were optimistic. Another example is change from baseline. If patients in a trial are selected on the basis of extreme values of a variable, then that variable is likely to show regression to the mean, compared to their baseline levels. This may lead the researchers to think that a benefit is due to

the intervention offered to the patient.

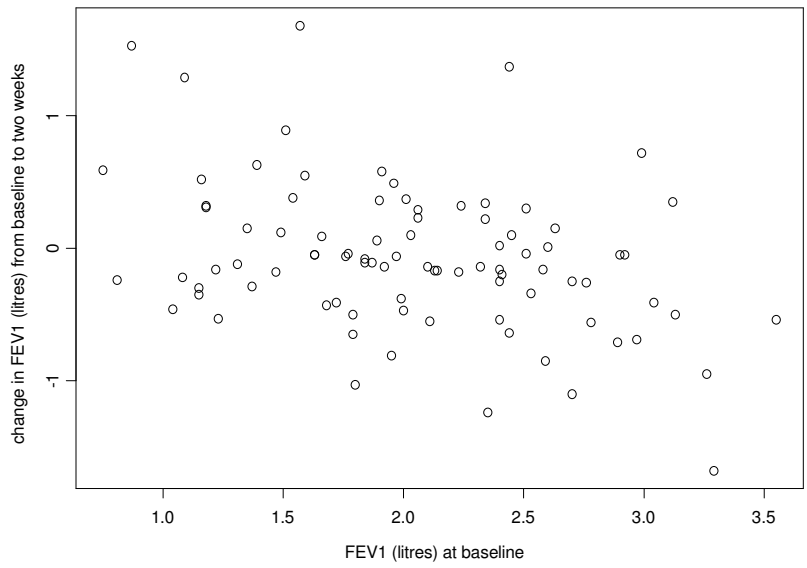
An example is shown in Figure 3, which contains data from the placebo arm of a trial of asthma therapies.[28] Those patients with low forced expiratory volume in one second (FEV_1) at baseline tended to have increased at the two week follow-up, while those with high baseline FEV_1 had decreased. Remember, this is the placebo arm. Imagine that we had done a non-controlled study with change from baseline as an endpoint, and included only those with the lowest values of FEV_1. In such a study, we would have observed an average increase, even if the intervention was as ineffective as placebo. Of course, we know that control arms are to help protect against this kind of pitfall. In the following section we will look in more detail about how to take into account baseline values in a controlled trial.

Comparing change from baseline between arms in a randomized trial

I was once advised by a statistical colleague against using the change from baseline ('change score') as an endpoint in a controlled trial, because of regression to the mean. Nevertheless, it is not invalid to do so. Analysis of change score is unbiased in the sense that, if repeated over multiple trials of the same design, on average it gives the correct answer.[19] However, there is a more powerful alternative, called analysis of covariance or ANCOVA. (Although ANCOVA is just a type of regression analysis, here I will persist in calling it ANCOVA to avoid possible confusion with regression to the mean.) ANCOVA uses the data to assess the degree of correlation between the baseline and final values, and does the adjustment on that basis. By contrast, analysis of change score effectively assumes a correlation of 1, although in practice it is less. This means that, in any particular dataset, regression to the mean will over-adjust for the baseline value. The over-adjustments may be positive or negative, and hence, thanks to randomisation, they cancel each other out if averaged over trials of the same design. However, they make the standard error of the change score analysis higher than that of ANCOVA. So comparing change scores between arms is not invalid but is not the most powerful option.

Figure 3

Example of regression to the mean.



FEV_1 = forced expiratory volume in one second, at baseline, and two weeks after receiving placebo, in a trial of trial of asthma therapies.[28] Those with the smallest values at baseline tend to have increased at two weeks, while those with the largest values tend to have decreased.

A simple unadjusted analysis will have smaller standard error than the change score (but not smaller than ANCOVA) if the correlation between baseline and follow-up measurements of the outcome variable have a correlation less than 0.5. For example, Hsieh *et al*.[29] did a randomised trial of acupressure *versus* physical therapy for low back pain. They analysed each of nine outcomes by ANCOVA and by change from baseline, at two different follow-up times (Table 3 of their paper). For every one of these 18 analyses, ANCOVA has the narrower confidence interval.

As we saw for hypothesis tests of baseline variables, the variables on which to adjust, if any, should be defined in the analysis plan. If we do choose to adjust for baseline measurement of the outcome variable, then the most powerful option is to use ANCOVA (regression analysis), although analysing change from baseline is not invalid.

A mistake I made: weights in meta-regression

Some statistical analyses require some data points to be given more weight than others. For example, a sample survey may be carried out to estimate the prevalence of a certain condition, such as the prevalence of ever having had an HIV test in the United Kingdom.[30] This survey had a greater sampling frequency in Greater London because 'prevalence of risk behaviours was expected to be higher'. This means that, when estimating the national prevalence, the Greater London data must be given less weight, to prevent them contributing overly to the results. In general, the weight of any data point should be proportional to the reciprocal of (1 divided by) its sampling variance.

Another kind of analysis which requires weighting is example is meta-analysis, i.e. the attempt to summarize the results of separate studies in a single analysis. Here, larger studies are given more weight. In 1991 I performed the statistical part of a meta-analysis of the role of maximum cytoreductive surgery for ovarian cancer.[31] I gave each study a weight proportional to its number of patients. This was done on a heuristic basis, linked to the idea that the standard error of a mean or proportion is proportional to the reciprocal of the square root of the sample size ($1/\sqrt{n}$). Since the sampling variance is the square of the standard error, choosing the weights to be proportional to n ensures, for a mean or proportion, that they are inversely proportional to the sampling variance.

There were two problems with this. Firstly, the outcome variable was the log of the median survival time, not a mean or proportion for which the weights proportional to *n* would be justified. I did not attempt to derive the sampling error of our outcome variable, although it turns out I was lucky in this respect (see below).

> "Standard error of logarithm of median survival for meta-regression
>
> If we assume a constant death rate λ in a single study, then the survival times will be drawn from an exponential distribution, which has mean $1/\lambda$ and variance $1/\lambda^2$.[39] To estimate the variance of the log of the sample mean ($\bar{x}$), we can use Taylor series (the 'delta method'): $\text{var}(y(\bar{x})) \approx (dy/d\bar{x})^2 \text{var}(\bar{x})$.[40] We have $y(\bar{x})=\log(\bar{x})$, so $dy(\bar{x})/d\bar{x}=1/\bar{x}$. So $\text{var}(\log(\bar{x})) \approx (1/\bar{x})^2 \text{var}(\bar{x})=(1/\bar{x})^2(1/\lambda)^2/n$, which is evaluated at the expected value of $\bar{x}$, i.e. $1/\lambda$. So $\text{var}(\log(\bar{x})) \approx 1/n$. Finally, we need the variance of the logarithm of the median, not the mean. But, since the median is a constant ($\log_e 2$) times the mean, the variance of its logarithm is the same as the variance of the logarithm of its mean.
>
> Incidentally, we may note that the sampling variance ($1/n$) does not depend on λ. This explains why the logarithmic transformation stabilized the variance in the original analysis (although I did not realise that at the time)."

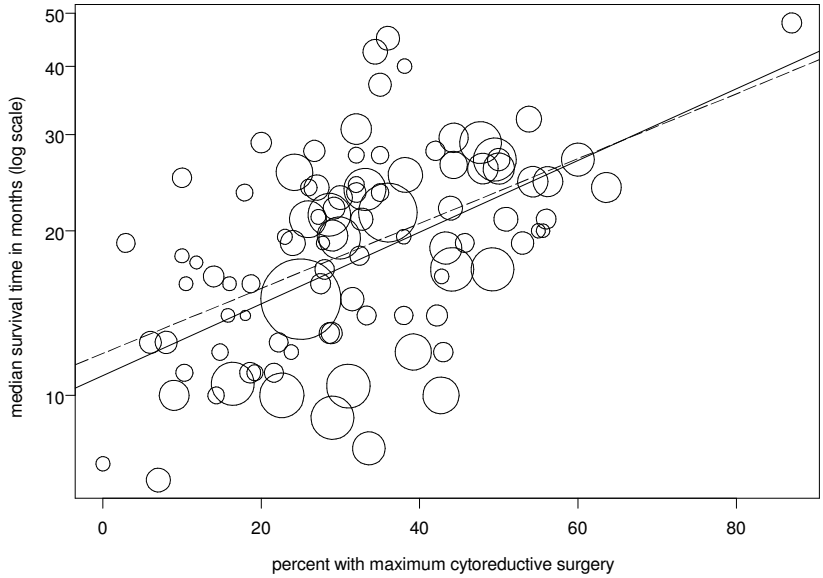
The second problem, which is more general to meta-analysis, is that sampling error, and the predictor variables, are not necessarily the only source of variation between studies. In regression, scatter about the line may be greater than can be explained by small studies having large sampling error. In other words there may be real differences in the underlying effects which are not captured by a regression line and sampling error. Meta-analysis uses a 'random effect' to represent this additional variation. However, my original ovarian cancer meta-analysis did not allow for this possibility.

I repeated the analysis using the 'meta-regression' method described by Knapp and Hartung,[32] implemented in the 'metareg' command in the STATA software. For this, it is necessary to specify the standard error of the dependent variable for each study, i.e. of the logarithm of the median survival time. Following the reasoning in the appendix, we can set this to be $1/\sqrt{n}$. This corresponds to the weights used in the original analysis, used to represent random sampling error. But the new method also allows us to estimate systematic variation between studies.

Figure 4 shows the relation between the outcome

Figure 4

Meta-analysis of Maximum Cytoreductive Surgery (MCS) in ovarian cancer[31]



The horizontal axis is the percent of each study group for whom MCS was achieved. The area of each circle is proportional to the number of patients in the study group. The lines represent the regressions of the logarithm of median survival (vertical axis) on the percent MCS from a) the original naively weighted analysis (solid line) and b) meta-regression32 (dashed line).

variable, median survival time (log-transformed) and the percentage of each study group which experienced Maximal Cytoreductive Surgery (MCS). Each study is represented by one circle, with the area of each being proportional to the sample size, and therefore to the weights of the original analysis. The corresponding regression (with only MCS as a predictor) estimates a 16.3% increase in median survival time for each 10% increase in MCS. Meta-regression indicates that there is indeed variation other than sampling error. This is measured by the parameter τ^2, and the test of the null hypothesis that $\tau^2 = 0$ has a p value less than 0.001. In other words, there is strong evidence that the between-study variation is not only due to sampling error. Nevertheless, the estimate of the relation between survival time and MCS is similar to the original analysis: 14.6% (rather than 16.3%) increase per 10% increase in MCS.

The main conclusion of the original paper was that the relationship between MCS and survival time is confounded by other variables, in particular: a) the type of chemotherapy, as measured by dose intensity and inclusion of platinum; and b) the case mix, as measured by the percent with Stage IV disease. In the original analysis, adjustment for these factors reduced the effect of MCS considerably, from 16.3% to 4.1%. Using the Knapp and Hartung[32] meta-regression method, the adjusted estimate is 7.3% per 10% increase in MCS. There is still evidence of variation in excess of sampling variation (the p value for τ^2 is still less than 0.001). This adjusted estimate for MCS is noticeably larger than the original estimate, although still much less than the unadjusted one.

Discussion

We can see some shared features of these problems in medical statistics. Misconceptions often arise by relying uncritically on p values, and may be

dispelled by thinking about exactly what is meant by the *p* value of a given analysis. Comprehension is also aided by using confidence intervals to interpret statistical analysis in clinical terms.[33] More generally, analyses whose parameters cannot be related back to clinical reality should be viewed with caution at best.

Unfortunately, however, pedagogical articles such as this, and other education resources, cannot by themselves raise the quality of statistics in medical journals to an acceptable level: we must consider factors which are more integral to the process of doing medical research.

Although statistical analysis is sometimes a necessary part of a research publication, it is not always valued by researchers, and the publication process itself can sometimes foster such an attitude. Bland[34] relates that several times he has told colleagues that their data did not need any hypothesis tests, since the results were completely clear, only to be told that they needed *p* values to get the work published. The path of least resistance to publication may still be to run an analysis which can be used to sprinkle the work with symbols intended to impress the reader, such as χ^2 and *p* values, without necessarily being the correct choice, or even without describing the method in the text.[35,36]

However, it would be insufficient to blame such occurrences on laziness of investigators (although that may sometimes play a part). We should also ask ourselves what barriers exist to accessing correct statistical advice. One may be that there is simply no statistician available to consult. I believe that other reasons include personality clashes, and lack of common ground, between statistically and biomedically trained personnel. This may be most evident between statisticians and physicians. Members of these two groups often find it difficult to establish a rapport. Statisticians may often have 'a certain shyness'[37] but they, like physicians, can be low in tact, and high in pride. Statisticians often have extensive training in mathematics, the 'queen of the sciences', with clinicians are often the most prestigious cadre in research institutes. The fact that their training paths have usually been separate since secondary school, until a statistician joins a research institute, accentuates the difficulty in establishing a productive working relationship.

Many biomedical researchers are not particularly numerate. They may even have 'chosen biological science in an attempt to avoid mathematics'.[37] This may help explain why some biomedical researchers prefer to seek the advice of one of their colleagues who has similar training to their own, but who has more of an affinity for numbers. Hence the presence in some departments and institutes of people who are known as statistical troubleshooters even though they do not formally have that responsibility. Such 'gurus' often go a good job, helped of course by their knowledge of the biomedical field which has yielded the data. However, there should not be a need for them to exist parallel to, and separate from, more formally trained statisticians.

Some of these problems could be alleviated by making closer links between the training of biological and statistical disciplines. Statistical education of medical students is sometimes poor, with the ambivalent attitude of students to numerical information often worsened by an overly mathematical way of teaching. In recent years, some efforts have been made to improve this, and I believe that this could beneficially be mirrored in the education of biostatisticians. In particular, many of the above problems could be eased if biostatistical masters degrees were more often earned within biomedical research institutes and included experimental work, perhaps as part of a two year course. Such work should also be part of continuing education, as pioneered by Stephen Evans at the London Hospital, where statisticians attended ward rounds and became familiar with measurement methods.[37] Although there are arguments against it[38] I also think that the normal career path for medical statisticians should include gaining a doctoral degree (in the UK this is not always the case). Being responsible for a research project of such size enhances the capacity for future work, and makes for a career path parallel to that of other academics. This and my other suggestions are intended to foster mutual respect between statistical and non-statistical colleagues. We should try to ensure that choosing a valid statistical method is not a baffling ordeal, but a task which can be done comfortably, even if sometimes time-consuming: less like writing a grant application budget, and more like deciding what to have for dinner.

Acknowledgements

I am grateful to Dr. Carl-Johan Lamm and colleagues at Astra-Zeneca for permission to use data from the asthma clinical trial, to Dr. James Carpenter for facilitating contact with them, and to Dr. Cynthia Braga, Prof. Eulalio Cabral Filho and Dr. Jailson de Barros Correia for the invitation to make the presentation at the Instituto Materno Infantil Prof. Fernando Figueira, IMIP in the city of Recife, State of Pernambuco, Brazil, on which this paper is based. I am also grateful to Prof. Richard Hayes for useful suggestions.

Referências

1. Bagenal FS, Easton DF, Harris E, Chilvers CED, McElwain TJ. Survival of patients with breast cancer attending Bristol Cancer Help Centre. Lancet. 1990; 336: 606-10.
2. Smith R. Charity Commission censures British cancer charities. Br Med J. 1994; 308: 155-6.
3. Welch GE, Gabbe SG. Statistics usage in the American Journal of Obstetrics and Gynecology: has anything changed? Am J Obstet Gynecol. 2002; 186: 584-6.
4. Braun-Munzinger RA, Southgate BA. Repeatability and reproducibility of egg counts of Schistosoma haematobium in urine. Trop Med Parasitol. 1992; 43: 149-54.
5. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet. 1995; 346: 1085-9.
6. Kapeller P, Barber R, Vermeulen RJ, Ader H, Scheltens P, Freidl W, Almkvist O, Moretti M, del Ser T, Vaghfeldt P, Enzinger C, Barkhof F, Inzitari D, Erkinjunti T, Schmidt R, Fazekas F, European Task Force of Age Related White Matter Changes. Visual rating of age-related white matter changes on magnetic resonance imaging: scale comparison, interrater agreement, and correlations with quantitative measurements. Stroke. 2003; 34: 441-5.
7. Gil Z, Abergel A, Spektor S, Khafif A, Fliss DM. Patient, caregiver, and surgeon perceptions of quality of life following anterior skull base surgery. Arch Otolaryngol Head Neck Surg. 2004; 130: 1276-81.
8. Desai MY, Lai S, Barmet C, Weiss RG, Stuber M. Reproducibility of 3D free-breathing magnetic resonance coronary vessel wall imaging. Eur Heart J. 2005; 26: 2320-4.
9. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician. 1983; 32: 307-17.
10. Bland MJ, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1: 307-10.
11. Bland JM, Altman DG. This week's citation classic: comparing methods of clinical measurement. Curr Contents. 1992; CM20: 8.
12. Alexander N, Bethony J, Corrêa-Oliveira R, Rodrigues LC, Hotez P, Brooker S. Repeatability of paired counts. Stat Med. 2007; 26: 3566-77.
13. Bassuk SS, Albert CM, Cook NR, Zaharris E, MacFadyen JG, Danielson E, Van Denburgh M, Buring JE, Manson JE. The women's antioxidant cardiovascular study: design and baseline characteristics of participants. J Womens Health. (Larchmt). 2004; 13: 99-117.
14. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. Lancet. 1990; 335: 149-53.
15. Senn S. Testing for baseline balance in clinical trials. Stat Med.1994; 13: 1715-26.
16. Ellis PM, Butow PN, Tattersall MH. Informing breast cancer patients about clinical trials: a randomized clinical trial of an educational booklet. Ann Oncol. 2002; 13: 1414-23.
17. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000; 355: 1064-9.
18. International Conference on Harmonisation; guidance on statistical principles for clinical trials; availability--FDA. Notice. Fed Regist. 1998; 63: 49583-98.
19. Senn S. Statistical issues in drug development. Chichester: Wiley; 1997.
20. Altman DG, Bland JM. Absence of evidence is not evidence of absence. Br Med J. 1995; 311: 485.
21. Kamali A, Quigley M, Nakiyingi J, Kinsman J, Kengeya-Kayondo J, Gopal R, Ojwiya A, Hughes P, Carpenter LM, Whitworth J. Syndromic management of sexually-transmitted behaviour change interventions on transmission Uganda: a community randomised trial. Lancet. 2003; 361: 645-52.
22. Alderson P. Absence of evidence is not evidence of absence. Br Med J. 2004; 328: 476-7.
23. Kevles DJ. In the name of eugenics. Cambridge: Harvard University Press; 1995.
24. Everitt B. Cambridge dictionary of statistics in the medical sciences. Cambridge: Cambridge University Press; 1995.

25. Pullein K. The manager of the month curse is a fallacy. The Guardian. 2005 Dec 2. Available from: http://football.guardian.co.uk/News_Story/0,1563,1656124,00.html?gusrc=rss. [2007 Apr 20].

26. Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. Br Med J. 2003; 326: 1083-1084.

27. Bland JM, Altman DG. Some examples of regression towards the mean. Br Med J. 1994; 309: 780.

28. Carpenter J, Pocock S, Lamm CJ. Coping with missing data in clinical trials: a model-based approach applied to asthma trials. Stat Med. 2002; 21: 1043-66.

29. Hsieh LL, Kuo CH, Lee LH, Yen AM, Chien KL, Chen TH. Treatment of low back pain by acupressure and physical therapy: randomised controlled trial. Br Med J. 2006;332: 696-700.

30. Burns F, Fenton KA, Morison L, Mercer C, Erens B, Field J, Copas AJ, Wellings K, Johnson AM. Factors associated with HIV testing among black Africans in Britain. Sex Transm Infect. 2005; 81: 494-500.

31. Hunter RW, Alexander NDE, Soutter WP. Meta-analysis of surgery in advanced ovarian carcinoma: is maximum cytoreductive surgery an independent determinant of prognosis? Am J Obstetr Gynecol. 1992; 166: 504-11.

32. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Stat Med. 2003; 22: 2693-710.

33. Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J. (Clin Res Ed) 1986; 292: 746-50.

34. Bland M. An introduction to medical statistics. Oxford: Oxford University Press; 1987.

35. Oliver D, Hall JC. Usage of statistics on the surgical literature and the orphan P' phenomenon. Aust N Z J Surg. 1989; 59: 449-51.

36. Alexander N. Paper critique as an educational method in epidemiology. Med Teacher. 2003; 25: 287-90.

37. Altman DG, Bland JM. Improving doctors' understanding of statistics. J Royal Stat Soc. [Series A] 1991; 154: 223-67.

38. Pocock SJ. Life as an academic medical statistician and how to survive it. Stat Med. 1995; 14: 209-22.

39. Evans M, Hastings N, Peacock B. Statistical distributions. New York: Wiley; 2000.

40. Armitage P, Berry G, Matthews JNS. Statistical methods in medical research. Oxford: Blackwell Scientific Publications; 2001.

Recebido em 13 de abril de 2007
Versão final apresentada em 10 de julho de 2007
Aprovado em 31 de julho de 2007