



Three things that every medical writer should know about statistics

by Stephen Senn

Introduction

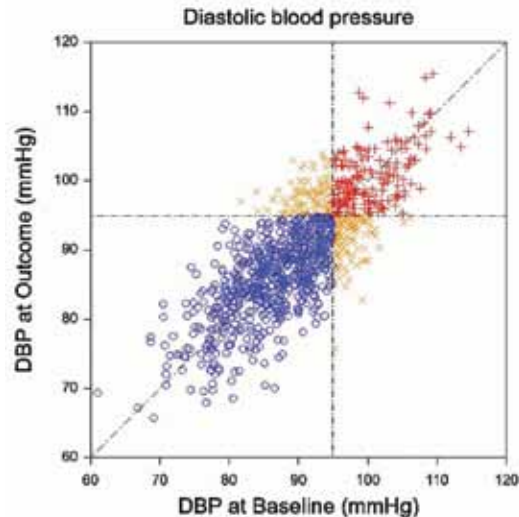
The joke goes that there are three kind of statistician: those who can count and those who can't. Therefore, readers of the *Write Stuff* will forgive me, I hope, if I end up writing about more than three things. It should be obvious, in that case, as to which sort of statistician I am. There are, of course, many more things than three that every medical writer should know about statistics because there are many things about statistics that anybody working in drug development should know and medical writers are in the unenviable position of having to know about everything. However, everybody has to start somewhere and three is a number with a great tradition. The three things I am going to write about are *regression to the mean* [1], the *error of the transposed conditional* [2] and *individual response* [3]. The first is a widespread phenomenon that has a powerful influence on the way that results appear to us, the second is a pernicious fallacy and the third is a sort of Holy Grail-cum-wild goose chase that is responsible for leading many a researcher astray.

Regression to the mean

Regression to the mean is the tendency for members of a population who have been selected because they are extreme to be less extreme when measured again [4, 5]. Because entry into clinical trials is usually only allowed if patients have extreme values (diastolic blood pressure above 95 mmHg, Hamilton depression score greater than or equal to 22, forced expiratory volume in one second less than 75% of predicted etc.), regression to the mean is a phenomenon that is likely to affect many clinical trials. We can expect that patients will appear to improve even if the treatment is ineffective. Regression to the mean is a plausible explanation, for example, for the 'placebo effect' which then becomes, as I hope to explain, a purely statistical rather than psychological phenomenon.

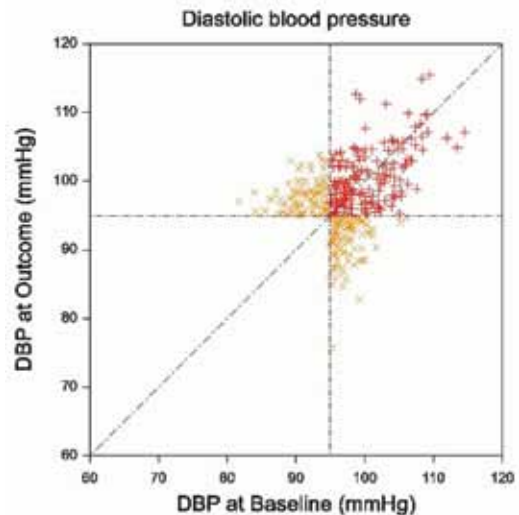
How does it occur? Consider figure 1. This shows a simulated set of results for a group of 1000 individuals who have had their diastolic blood pressure (DBP) measured on two occasions: at 'baseline', X, and at 'outcome', Y. The figure plots Y against X and the simulation has been arranged so that the expected values of X and Y are identically equal to 90 mmHg and that the standard deviations are 8 mmHg with a correlation of 0.79. An arbitrary but common cut off of 95 mmHg is taken as being the boundary for hypertension. Individuals are labelled as being of one of three sorts: hypertensive at both baseline and outcome (labelled with a red +), normotensive at both baseline and outcome (labelled with a blue 0) and hypertensive on one occasion and not the other (labelled with an orange x).

Figure 1 Simulated results at baseline and outcome for diastolic blood pressure (mmHg) for 1000 individuals in a population.



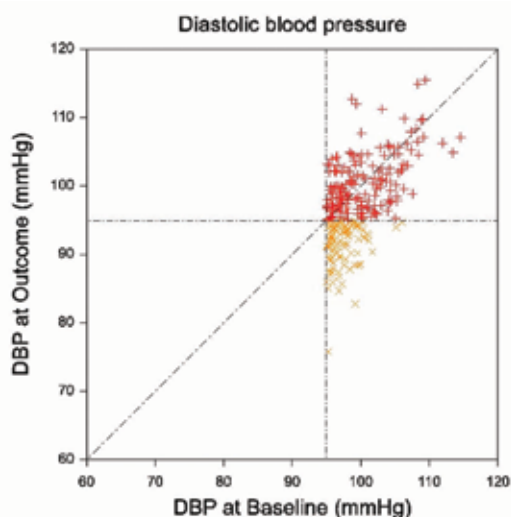
Now consider a plot of a subset of the individuals, namely those who are 'hypertensive' on at least one occasion. This plot is given in figure 2. Just as was the case in figure 1 there is no essential difference as to whether we look at results at baseline or outcome, the mean result on either occasion, although higher than it was before because the 'normotensives' have been removed, will be the same.

Figure 2 Simulated results at baseline and outcome for diastolic blood pressure (mmHg) for 1000 individuals in a population with those who are normotensive on both occasions removed.



However, neither of these plots is what we would observe in a standard clinical trial. Instead, we would observe something like figure 3. Figure 3 has been obtained from figure 2 by removing those patients who were normotensive at baseline but hypertensive at outcome. Why? Because if they were normotensive at baseline they would never be recruited into the trial and hence never followed up. Now we can see that the way that we have chosen subjects has an inherent bias if we measure the effect of treatment as the difference between outcome and baseline. The outcome values are on average lower than the baseline values but this is only because of the way that we have sampled. It says nothing about the effect of treatment.

Figure 3 Simulated results at baseline and outcome for diastolic blood pressure (mmHg) for 1000 individuals in a population with those who are normotensive at baseline removed.



The consequence is that on average patients will appear to improve even if the treatment is ineffective. In fact, patients given placebo can be expected to improve for reasons that are purely statistical. There is no need to invoke psychology, the healing hands of the physician, the white coat effect and so forth. The way that the data are collected suffices.

Does it matter? Not in a controlled clinical trial provided that we only consider, describe and interpret differences between treatment and control groups. Both of these will be subject to the same regression to the mean effect, which is therefore eliminated by comparison. Hence, the joke about a medical statistician. If you ask him, “how’s your wife?” he answers, “compared to what?” Only head to head comparisons have meaning. Alas, many clinical trial reports reveal that trialists have no idea why they have carried out a controlled clinical trial. Pages of ink are wasted describing the response in each group, although this is meaningless. Reports would be sharper and understanding would be improved if these ignorant descriptions were dumped where they belong in the waste paper basket.

What are the lessons for a medical writer? He or she should think comparatively. Controlled clinical trials are about comparisons, or to use some statistical jargon *treatment*

Statistics: What medical writers should know

contrasts, that is to say difference between treatments. Given a choice between a graph that shows the course over time of each treatment together with standard error bars or a plot of the difference between treatments together with confidence interval for that difference, choose the latter and dump the former. If survival is the outcome of interest, it is the log-hazard ratio, a statistic used to model the difference between treatments, that should take pride of place and not the median survival within each group. For a binary outcome, stress the odds ratio rather than the probability for each group.

The error of the transposed conditional

All French are Europeans but not all Europeans are French. I can put this in the language of probabilities. With a probability of 100% someone who is French is European. However, the probability that a randomly chosen European (taking this to mean a citizen of the European Union) is French is only about 13% (since the population of France is about 65 million and that of the European Union about 500 million).

Here is another example. The probability that a randomly chosen woman has breast cancer is, thank goodness, quite low. However the probability that a randomly chosen breast cancer victim is a woman is extremely high. Or how about the *prosecutor’s fallacy*? The probability of the DNA on the scene of the crime matching that of the defendant is one in a million, therefore, claims the prosecution, there are 999,999 chances out of a million that he is guilty. However, in a population of 100 million (which could be the number of adult males in the USA) there must be 100 individuals about whom we could make a similar statement. They can’t all be almost certainly guilty.

Patients given placebo can be expected to improve for reasons that are purely statistical

This is all very obvious and elementary, yet, surprisingly, even experienced trialists find it hard to grasp that the probability of A given B is not the same as the probability of B given A. Consider

that most ubiquitous of statistics, the P-value. A P-value is the probability of seeing a result as extreme or more extreme than that observed if the null hypothesis is true. In other words it says something about the probability of the evidence given the null hypothesis. It is not, therefore, the probability of the hypothesis given the evidence. Yet it is often misinterpreted as being the probability that the null hypothesis is true. This is just an egregious error.

P-values are a concept in frequentist statistics. The frequentist approach to statistics is the approach generally used in drug development. In this approach it is never possible to talk of the probability of a hypothesis being true. The hypothesis is either true or false. The problem is we don’t know which. If one wished to make statements about the truth of a hypothesis one would have to use the Bayesian

Statistics: What medical writers should know

➤ system of inference but, in fact, this is rarely employed in a regulatory context.

What are the lessons for medical writers? Be careful in re-phrasing statistical statements. Here there be tygers! You may find the way that statisticians formulate probabilistic statements clumsy. However, you simplify at your peril. Unfortunately, many prefer a simple lie to a complicated truth but being truthful is what reporting clinical trials is all about.

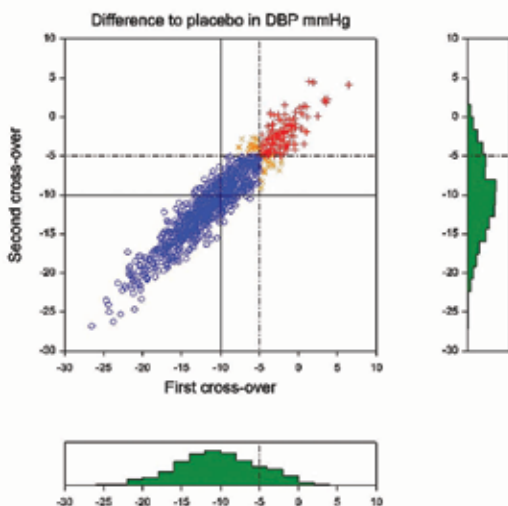
Individual response

Despite a gross of ‘points to consider’ documents to the contrary, individual response is usually not identifiable in a clinical trial. By that I mean that it is usually not possible to say who has and who has not responded to treatment in a clinical trial. I shall make good this claim in due course but cannot resist pointing out that many in drug development, including, if the published record is anything to go by, most of the European regulators, are deeply confused on this issue.

Consider a thought experiment in which we run two placebo-controlled cross-over trials in hypertension in succession. In each cross-over trial we will be able to compare the DBP under treatment and placebo. We will thus be able to construct two estimates for every patient of the effect of treatment compared to placebo using the difference between active treatment and placebo: one for each of the two cross-over trials. What pattern might we expect from such an experiment?

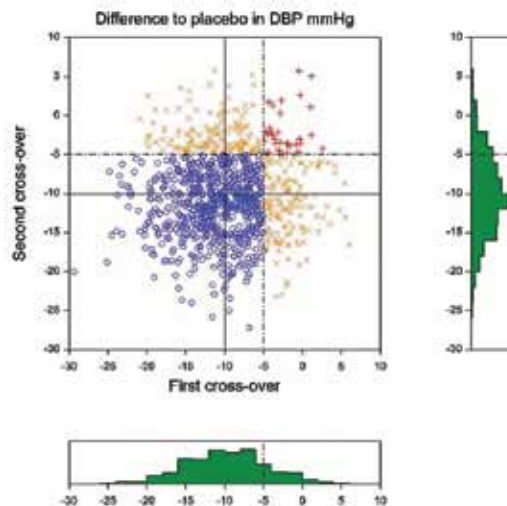
Figures 4 and 5 show two very different patterns. What these figures plot is the ‘response’ patient by patient for the second cross-over trial compared to the first. In fact these are based on a simulation in which I set the difference between treatment and placebo in DBP to be on average 10 mmHg. That is to say I set it to be 10 mmHg lower on average under treatment. Solid lines at -10 mmHg indicate the centres of the distributions. Also shown are dashed lines at -5 mmHg corresponding to an arbitrary definition of response whereby any patient who shows a 5 mmHg improvement compared to placebo is labelled a ‘responder’.

Figure 4 Patients cross-classified by difference in DBP (mmHg) treatment minus placebo for two successive cross-over trials. Case of strong correlation.



What the figures show right away is that this sort of arbitrary label is very silly. A great deal of information is lost by turning a continuous scale into a binary one and the boundary is, of course, arbitrary. What the figures also show, however, is that one has to be very careful as to how one interprets ‘response’.

Figure 5 Patients cross-classified by difference in DBP (mmHg) treatment minus placebo for two successive cross-over trials. Case of no correlation.



In the case of figure 4, there is a strong correlation between the degree of response in the first cross-over and the degree of response in the second. It thus becomes meaningful to think of response as being a feature of the patient. There seem to be some patients who respond well to the anti-hypertensive agent—these are those shown as blue circles. There are also some who respond less well and, using our standard of -5mmHg, these are shown as red crosses. Very few patients show an inconsistent pattern.

However, in figure 5 the situation is quite different. Here ‘response’ in the first cross-over does not appear to predict response in the second. This can be shown by producing the following table (see table 1) of cross-classified results. Of the 826 patients who responded in the first cross-over 678 responded in the second. The proportion, therefore is 678/826=0.82. However of the 174 who did not respond in the first, 140 responded in the second. The proportion, therefore, is 140/174 =0.80. This is almost identical to the previous figure. It thus seems that at any time any patient has about four chances out of five of responding and about one chance out of five of not responding. It thus becomes meaningless to think in terms of responders and non-responders.

Table 1 Patients cross-classified by response in two cross-over trials

Count			
Second cross-over	responder	non-responder	Count
First cross-over			
responder	678	148	826
nonresponder	140	34	174
Count	818	182	1000

Also, included in figures 4 and 5 are some histograms showing the marginal distributions of 'response'. Now suppose that, in fact, we had never run the second cross-over trial. All that we would have to go on would be the marginal figures, summarised by the histograms given on the X axis of figures 4 and 5. The point is, however, that from these marginal histograms it is impossible to predict whether we would see something like figure 4 or figure 5 in carrying out a second cross-over trial.

In other words, a single cross-over trial is inadequate for distinguishing responders from non-responders. What applies to cross-over trials applies *a fortiori* to parallel group trials. It is a fact of drug development that we simply do not run the sort of trial that would allow us to separate responders from non-responders and much of the hype about pharmacogenomics is based on the largely untested hypothesis that patients respond very differently to treatment from each other [6].

What are the lessons for medical writers? The less ink you waste in any clinical trial discussing individual 'response' the better. What most clinical trials deliver are averages. That's all. A placebo-controlled parallel group trial in hypertension will tell you what the average difference compared to placebo in DBP is due to taking treatment. That's it. Anything else is unscientific speculation.

Categorising patients as responders and non-responders by comparing their values to baseline does not control for regression to the mean, does not establish that patients can be so-classified in any meaningful way, is in any case arbitrary and is inefficient. (An analysis of responders rather than in terms of a continuous measurement such as DBP will lead to a large increase in the sample size needed.)

In conclusion

I finish by giving a list of recommendations

1. Be aware of regression to the mean. The way that data are collected in clinical trials means that comparison to baseline is inherently misleading.
2. Think comparatively. Randomised clinical trials are valuable because they provide concurrent control and thus eliminate many biases. This elimination is only achieved, however, by actually taking the last step and making the head to head comparison. The difference between treatment and control is what the clinical trial is about. The rest is unimportant.
3. Be careful with probability. Be on your guard for oversimplifications. Do not confuse the probability of A given B with that of B given A.
4. Watch out for the weasel word *response*. Use it sparingly. It takes hundreds and sometimes thousands of patients to show that a treatment works at all. It is usually impossible to tell in any individual case whether a patient has or has not benefitted.

It becomes meaningless to think of responders and non-responders. We do not run trials that would allow us to separate responders from non-responders.

Statistics: What medical writers should know

Finally, I should like to recommend a book I rather like. In fact, I wrote it myself. It's called *Dicing with Death* [7]. If you find that this article challenged your intuition and if you are not afraid of being challenged, you might find it interesting.

Stephen Senn

Department of Statistics,
University of Glasgow,
Glasgow, UK
stephen@stats.gla.ac.uk

References:

1. Senn, SJ. Editorial: regression to the mean, *Statistical Methods in Medical Research* 1997; 6: 99-102.
2. Senn, S. Transposed conditionals, shrinkage, and direct and indirect unbiasedness, *Epidemiology* 2008; 19: 652-654.
3. Senn, SJ. Individual response to treatment: is it a valid assumption?, *BMJ* 2004; 329: 966-968.
4. Bland, JM, Altman, DG. Regression towards the mean, *BMJ* 1994; 308: 1499.
5. Bland, JM, Altman, DG. Some examples of regression towards the mean, *BMJ* 1994; 309: 780.
6. Senn, SJ. Individual Therapy: New Dawn or False Dawn, *Drug Information Journal* 2001; 35: 1479-1494.
7. Senn, SJ, *Dicing with Death*, Cambridge University Press: Cambridge, 2003.

The sequel: JAMA and conflicts of interest

A box in the last issue of *TWS* (18(2):143) reported that *The Journal of the American Medical Association (JAMA)* had changed its policies on investigations of conflicts of interest to require whistleblowers to wait until *JAMA* had completed its investigations (however long these might take) before going public on their complaint. This policy and the events surrounding Jonathan Leo's complaint and *JAMA*'s reaction resulted in heavy criticism of *JAMA*. The editors had published the new policies in an editorial on its website [1]. The next event was that the editorial disappeared from the website and all biomedical databases. This raised questions in the medical editors' community about the ethics of erasing the publication record. Udo Schuklenk pointed out on the WAME listserv that according to information he had received from Wiley-Blackwell anything published with a DOI number online must not be changed in any print version or on-line without a notice of retraction or erratum because an on-line paper with a doi number is treated just like a print article [2]. *JAMA* did not publish any such notice or erratum. A new milder version of the editorial was published in the 8th July issue of *JAMA* [3]. The American Medical Association's board of trustees are reported to have examined concerns raised over how *JAMA*'s editors had handled the issue. Rebecca Patchin, who is the chairwoman of the association, stated, "We anticipate *JAMA*'s procedures for resolving undisclosed conflicts of interest by journal authors will be improved" [4].

1. www.udo-schuklenk.org/files/jamamarch.pdf
2. Schuklenk U. *JAMA* vs Leo the bizarre story continues. WAME listserv discussion, July 08, 2009
3. DeAngelis CD, Fontanarosa PB. Resolving Unreported Conflicts of Interest *JAMA* 2009;302(2):198-199.
4. *BMJ* 2009;339:b2926