CONTROL IN CLINICAL TRIALS

Stephen Senn
Department of Statistics, University of Glasgow, United Kingdom
stephen@stats.gla.ac.uk

Amongst the many types of medical scientific investigation that are possible, the randomized double-blind controlled clinical trial has a very high reputation. Without control there can be no randomization and without randomization no convincing blinding, It seems, therefore, that control is the key feature of such trials. Yet the way in which such trials are analyzed, including the way in which they are presented, shows that many trialists do not understand the value of what they have done. I illustrate the problem with various examples. One possible reason that trialists may underestimate the value of concurrent control is that they do not understand what a powerful source of bias regression to the mean constitutes. I consider how physicians can be taught to understand this difficult phenomenon. I also present a striking example of differences between populations, which can be used to teach care in using control information appropriately.

BACKGROUND

The purpose of a randomized controlled clinical trial is to compare treatments. The standard statistics that are used to do this, whether a difference in mean response, a hazard ratio or a log-odds ratio (as the case may be), reflect the idea of comparing on a suitable scale, the investigational treatment to some comparator studied concurrently. The fact that commonly, in an investigation of a new treatment, half the patients are allocated to some other treatment, whose effects are already well known, only makes sense if one fears that differences in patient populations from trial to trial may be important, and perhaps large enough to introduce serious bias if historical controls are used. Hence, the use of concurrent controls and the elimination of bias through comparison.

Sample size programs in common use reflect this philosophy. You are required to input the ratio of patients on one arm *to* another, the common standard deviation, the type I and type II error rates and clinically relevant *difference* and the software will then deliver the numbers of patients that you need to study on each arm.

So much for the theory. In practice, however, any observer of the literature on clinical trials will note a curious phenomenon: in numerous articles, many paragraphs and most illustrations will be devoted to showing not the difference *between* groups but the mean values *within* groups. Thus, it is very common to show traces over time of the mean response in each group together with standard error bars. The standard error bars are commonly calculated by dividing the standard deviation by the square root of, *n*, the number of observations. Thus the error of not comparing the treatments is compounded by calculating a statistic that only applies when simple random sampling occurs, which is never the case in clinical trials.

One can speculate as to the reasons for this unfortunate situation. One, no doubt, is habit. Journals habitually publish articles that concentrate on results within groups rather than differences between them so that young researchers in copying what has been published perpetuate the practice. Another may be that researchers pay lip service to the possibility of bias but do not understand that it can be important in practice. In this paper I discuss two possible sources of bias, regression to the mean and variation between populations, of which researchers ought to be aware, and present some possible material which may be used to teach understanding of this phenomenon. I start with regression to the mean.

REGRESSION TO THE MEAN

Regression to the mean is the tendency of items that have been selected on the basis of an extreme measured value to have a result, which, whilst higher than average, is lower than it was on selection. Most statisticians understand that clinical trials are inherently subject to regression to the mean, since entry to them is typically on the basis of some extreme measured value. It has been often been discussed, for example, that the so-called placebo effect may in many cases be due to

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

the purely statistical phenomenon of regression to the mean(McDonald, Mazzuca et al., 1983; Senn, 1988; Hrobjartsson & Gotzsche, 2001; Hrobjartsson & Gotzsche, 2004).

An investigation of regression to the mean can involve deep matters and difficult algebra. Fortunately, however, a simple graphical approach exists that make it possible to illustrate the phenomenon extremely effectively(Senn, 2009). Figure 1 shows simulated data for diastolic blood pressure (DBP) readings (mmHg) from a population of 1000 individuals measured on two occasions. The X axis show their readings when first measured ('baseline') and the Y axis their readings when measured again ('outcome'). What the plot shows is that for this population there is no real difference between the two occasions. The mean DBP reading is about 90mmHg on both occasions and the correlation between measurements is about 0.77.

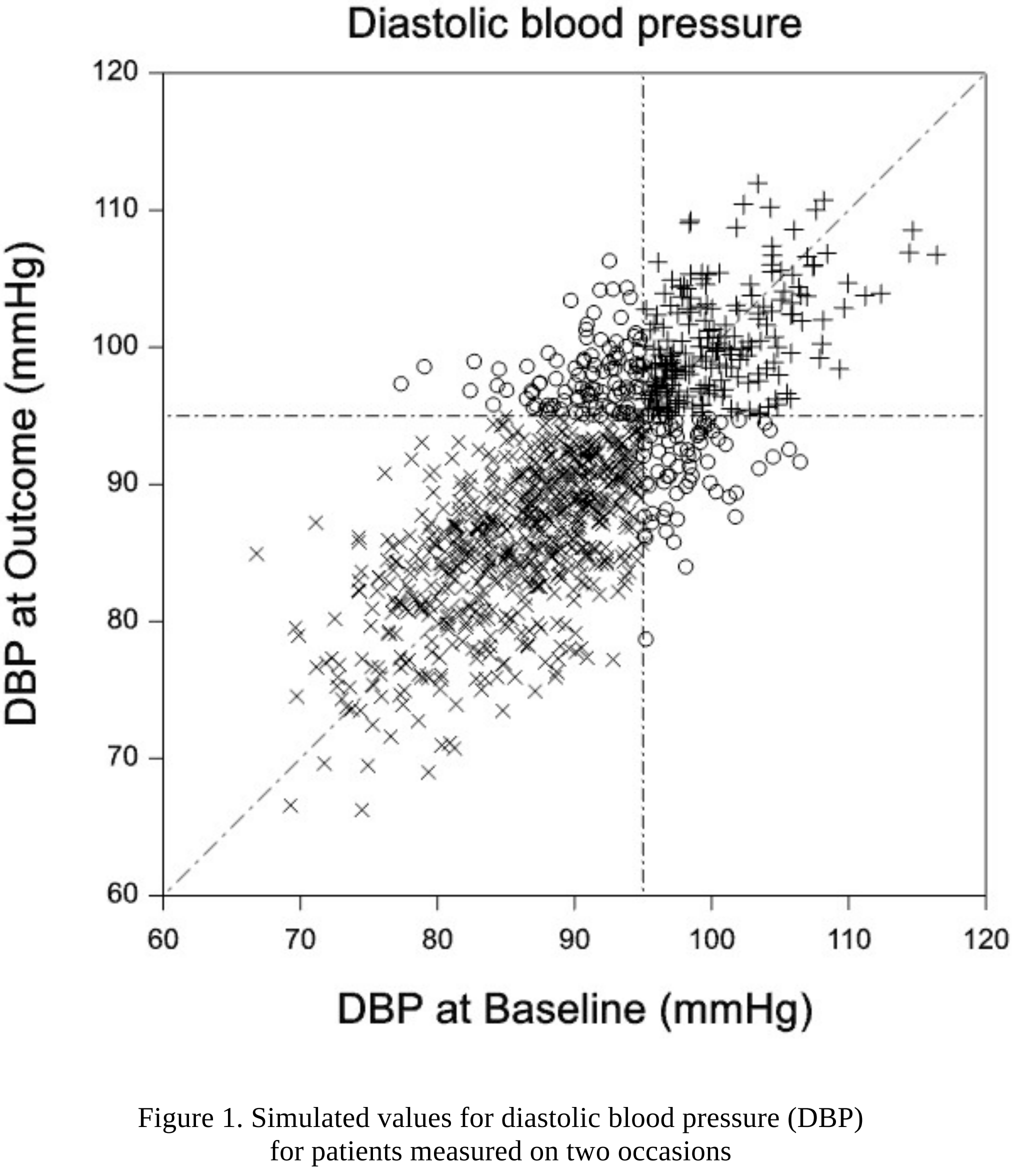

Figure 1. Simulated values for diastolic blood pressure (DBP)
for patients measured on two occasions
(Patients whose DBP was consistently less than 95mmHg are marked x, those whose DBP was consistently greater than or equal to 95mmHg are marked + and those whose DBP was greater than or equal to 95mmHg on one occasion but below on another are marked O.)

Figure 2 shows the same scatter plot but with all those subjects whose DBP was less than 95mmHg at baseline removed. This might be the case for a clinical trial in which patients who were suspected to be hypertensive were screened for possible entry into a clinical trial and only those with a DBP of at least 95mmHg were selected. It can now be seen that the mere act of sampling in this way induces a difference in the shape of the plot depending on whether it is looked at in the X or the Y dimension. The baseline values are all greater than 95mmHg by definition because of the selection process. However, there is no requirement by definition for all values at outcome to be greater than 95mmHg and since the correlation is less than one, some are, indeed, lower than this. The net result is that whereas the mean value at baseline is about 100mmHg, the mean value at outcome is just about 98mmHg.

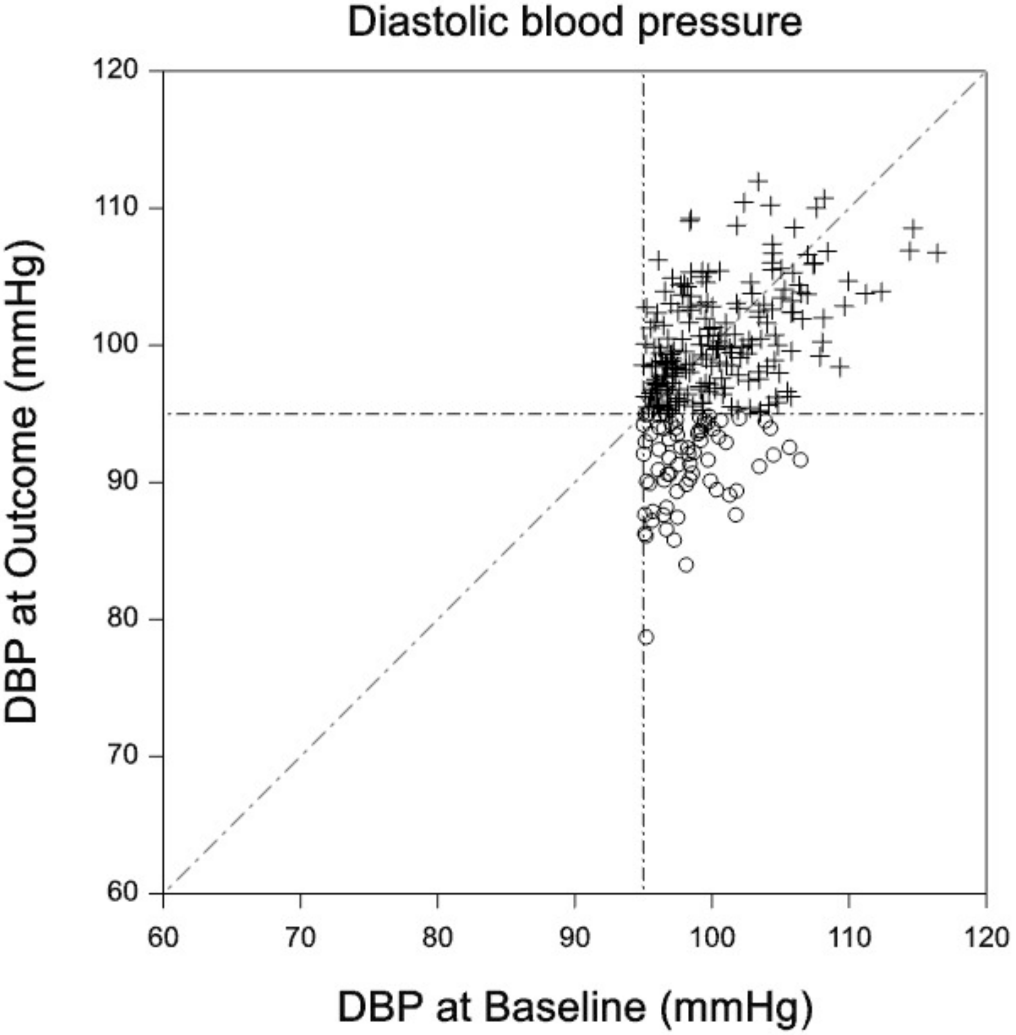

Figure 2. Diastolic blood pressure (DBP) at baseline and outcome for patients selected because their baseline value was greater or equal to 95mmHg

A GOOD PRACTICAL EXAMPLE FOR TEACHING CONTROL

The TARGET study (Farkouh, Kirshner et al., 2004; Schnitzer, Burmester et al., 2004) was a trial in osteoarthritis involving a comparison of lumiracoxib 400mg once daily, ibuprofen 800mg three times daily and naproxen 500mg twice daily. Studies comparing different treatments can often involve complicated blinding schemes in which each patient receiving a given treatment will also have to be given placebos to all of the others in order to avoid revealing which treatment is the genuine one. If different schedules are involved, as was the case here (once, three times and twice daily therapy), this can be particularly complicated. To simplify this, the trial was run as two substudies: one comparing lumiracoxib and ibuprofen and the other comparing lumiracoxib and naproxen (Hawkey, Farkouh et al., 2004). Because, apart from these differences, the same protocol was employed and because the way in which patients were allocated to treatment was at random within substudies but not between substudies, the TARGET trial is an ideal guinea pig to exemplify the value of randomization and concurrent control.

Table 1, which is taken from (Senn 2008) shows the distribution of patients at baseline in the TARGET study by various binary demographic characteristics. It can be seen that there is excellent balance between treatment arms within substudies but very poor balance between substudies.

Table 1. Distribution of selected demographic characteristics in the TARGET study. (Based on Farkouh, Kirshner et al., 2004))

	Sub-study 1		**Sub-study 2**	
Demographic characteristic	**Lumiracoxib** n = 4376	**Ibuprofen** n = 4397	**Lumiracoxib** n = 4741	**Naproxen** n = 4730
Use of low-dose aspirin	975 (22%)	966 (22%)	1195 (25%)	1193 (25%)
History of vascular disease	393 (9%)	340 (8%)	588 (12%)	559 (12%)
Cerebrovascular disease	69 (1.6%)	65 (1.5%)	108 (2.3%)	107 (2.3%)
Dyslipidaemias	1030 (24%)	1025 (23%)	799(17%)	809(17%)

Table 2 shows an analysis of the baseline demographic dichotomies using logistic regression. Since baseline measures are taken before treatment, in a randomized study we only expect to find chance differences between treatment arms. However, randomization did not occur between substudies and a logistic regression model for each of the four dichotomies shows a significant reduction in deviance (on one degree of freedom) if substudy is fitted compared to the null model in which no factors are fitted. If treatment, which has two degrees of freedom, is added to the model including substudy, the reduction in deviance is *not significant*. In other words, in the TARGET study there are, indeed, only chance differences between treatment arms *provided that we include substudy in the model*. However, if we fail to include substudy in the model, then the reduction in deviance fitting the treatment effects compared to the null model is significant for all demographic characteristics. This has clear implications for the way in which outcome variables for the trial should be analyzed (Senn, 2008). Here we should have substudy in the model first with treatment as a subsequently added factor.

Table 2. Results of carrying out significance tests on the baseline demographic variables

Deviances for the four demographic variables

Effect	Aspirin	Vascular History	Cerebrovascular	Dyslipidaemias
substudy	23.57	70.14	13.538	117.98
treatment-given-substudy	0.13	5.23	0.144	0.17
treatment	13.40	47.41	7.745	54.72

Approximate chi square probabilities for the four demographic variables

Effect	Aspirin	Vascular History	Cerebrovascular	Dyslipidaemias
substudy	0.0000	0.00000	0.0002	0.0000
treatment-given-substudy	0.9365	0.07332	0.9304	0.9194
treatment	0.0012	0.00000	0.0208	0.0000

CONCLUSION

The essence of the classical randomized clinical trial is the use of concurrent control. The value of this is often taken by default. It is assumed to be so obvious that it does not require discussion. However, in the way in which they analyze and present clinical trials, trialists frequently undermine the value of concurrent control. A graphical lesson in the dangers of regression to the mean and a practical example of the value of concurrent 'control' may provide the means of imparting some valuable lessons.

REFERENCES

Farkouh, M. E., Kirshner, H. et al. (2004). Comparison of lumiracoxib with naproxen and ibuprofen in the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET), cardiovascular outcomes: randomised controlled trial. *Lancet, 364*(9435), 675-84.

Hawkey, C. J., Farkouh, M. et al. (2004). Therapeutic arthritis research and gastrointestinal event trial of lumiracoxib - study design and patient demographics. *Aliment Pharmacol Ther, 20*(1), 51-63.

Hrobjartsson, A., & Gotzsche, P. C. (2001). Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. *New England Journal of Medicine, 344*(21), 1594-602.

Hrobjartsson, A., & Gotzsche, P. C. (2004). Is the placebo powerless? Update of a systematic review with 52 new randomized trials comparing placebo with no treatment. *Journal of Internal Medicine 256*(2), 91-100.

McDonald, C. J., Mazzuca, S. A., et al. (1983). How much of the placebo 'effect' is really statistical regression? *Statistics in Medicine, 2*(4), 417-27.

Schnitzer, T. J., Burmester, G. R., et al. (2004). Comparison of lumiracoxib with naproxen and ibuprofen in the Therapeutic Arthritis Research and Gastrointestinal Event Trial (TARGET), reduction in ulcer complications: randomised controlled trial. *Lancet, 364*(9435), 665-74.

Senn, S. (2008). Lessons from TGN1412 and TARGET: implications for observational studies and meta-analysis. *Pharm Stat, 7*, 294-301.

Senn, S. J. (1988). How much of the placebo 'effect' is really statistical regression? [letter]. *Statistics in Medicine, 7*(11), 1203.

Senn, S. J. (2009). Three things every medical writer should know about statistics. *The write stuff, 18*(3), 159-162.