Follow-Up in Study Design: How Much Is Enough? By Dr. Arthur C. Croft

The road of scientific investigation is littered with "new discoveries" and "breakthroughs" that failed to pan out in subsequent studies. More often than not this is the result of simple regression to the mean, small sample size, biased sampling, subgroup analysis or some other form of data dredging.

Regression to the Mean

Regression to the mean frequently explains initial successes to new therapies. We see this phenomenon virtually everywhere in the natural world. Tall parents often have shorter children; shorter parents have taller ones. Severe winters are followed by milder winters, etc. Enrolling **a group of migraineurs** in a treatment trial is likely to capitalize on the natural downswing in frequency or severity of at least some of the enrollees placed in the treatment group. Conversely, some of the migraineurs in the control group will likewise be cycling into a natural upswing of frequency and severity. These natural variations could be misinterpreted as treatment effects using statistical approaches.

Small Sample Size

Small sample size is common because it is convenient. As a rule of thumb in research, the administrative and practical cost to follow one subject for one year is approximately $1,000. Studies with sample sizes under 50 are, not surprisingly, quite common. Statistical methods do allow us to account for sample size, but they do not account for study design problems such as bias, confounding and a long list of what epidemiologists refer to as fallacies – some of which are particularly subtle and hard to spot.

The risk of falsely rejecting the null hypothesis – also known as a type I error or false positive – is a risk in biased samples. The opposite, a type II error or false negative, is also possible when bias is present. Bias is often pretty obvious, but can also be quite subtle.

For example, in the fictional migraine therapy example, if the subjects were all selected from patients attending a large university headache center, can we assume that they are fully representative of most migraineurs? It could be that these people have more severe headaches, higher frequency of headaches, and longer duration of headache suffering than the average migraineur. Conversely, suppose they were self-selected and self-diagnosed by simply completing a "Do You Suffer From Migraine Headaches?" questionnaire in the waiting room of their GP.

Subgroup Sampling

A subtle statistical manipulation is seen in subgroup sampling. Researchers naturally want to see their hard work published and know journal bias will keep their report out of the best journals if they fail to report some kind of significant findings. A failure to achieve them will usually mean publication in a lesser journal, if at all, and it could take years to find a home. Very often, a little creative secondary analysis nets the sought-after p value of less than 0.05. Using the migraine example again, suppose the difference in headache severity and frequency between the control group and the active treatment group did not reach significance at the end of the study; say p= 0.09. But then the authors noted that about half of the group as a whole were nondrinkers. So they ran the analysis again, this time comparing the results separately for drinkers and nondrinkers, and found a significant difference.

In the strictest sense, this would not be considered good science. At the least, this practice should be made transparent to study readers. This is done with an initial statement of the hypothesis being tested so the reader understands that the study evolved into something else. The reader can draw their own opinions as to the appropriateness of that. Readers will otherwise have no way of knowing whether the authors' initial hypothesis was about **alcohol consumption**.

Data Dredging

In the fictional headache example, the authors failed to reject the original null hypothesis (i.e., that there was no difference in the frequency and severity of headaches among migraineurs treated with the "treatment" versus the "controls" getting the placebo, so they explored around for a new one mid-study to salvage the original study. A less charitable description would be data dredging. (Interestingly, the software vendors who sell statistical software now have very sophisticated programs using neural networks on other very robust algorithms that look for relationships and associations in data. They call this data mining.) One should always ask when reading such a study whether this subgroup analysis would ever have been explored had the original analysis produced a p value of under 0.05.

Tracking Subjects Over Time

I mentioned that it was expensive to follow subjects over time. One way to reduce costs is to follow subjects for only a very brief period. But, without a reasonably long follow-up, the reader has no way to appreciate the true value of the therapy. Even very expensive therapies might be worthwhile if they produced lasting effects.

In a study comparing medical, chiropractic, and acupuncture therapy in the treatment of chronic spinal pain, chiropractic was shown to edge out medicine and acupuncture, with the exception of cervical spine pain, which went to acupuncture.[1] But the follow-up of these patients several months later was even more telling. Those treated by chiropractic showed the largest long-term benefit in all categories.[2] Chiropractic was not only more effective in reducing pain and other parameters measured, but also had the most robust long-term effects.

In the whiplash literature, there have been a few very long-term follow-up studies. The longest such study to date has been recently reported in the *Journal of Bone & Joint Surgery* (British) by Rooker, et al.[3] They've followed this group for many years and have reported the outcome in previous studies. I am acquainted with two of the authors, Martin Gargan and Gordon Bannister, and their names are well-known to aficionados of this particular literature. In this recent study, the authors reported:

"We have reviewed 22 patients at a mean of 30 years (28 to 31) after a whiplash injury. A complete recovery had been made in ten (45.5%) while one continued to describe severe symptoms. Persistent disability was associated with psychological distress but both improved in the period between 15 and 30 years after injury. After 30 years, ten patients (45.5%) were more disabled by knee than by neck pain."

On first blush, it sounds like the results might contradict the earlier (i.e., 10-year) results. But I think it is fair to say that this study has actually outlived its usefulness. When 30 years have passed since a single minor injury episode, the relentless onslaught of age, subsequent injuries or microtrauma, along with various other common health issues, will gradually have sufficiently muddied the clinical picture to such an extent that making realistic associations between the very long-distant injury and current health state becomes tenuous at best. Complicating matters, of the original study group, only 36 percent remain, and again, there was no control cohort.

The authors reported that the disability level had actually improved in some of the subjects. What could account for these improvements? Possibilities include retirement and a less physical lifestyle, gradual accommodation to their disability, newer, more effective medications, access to previously unavailable treatments (e.g., ESI, RF) or surgery. Notwithstanding these trivial cavils, I will give my two British colleagues credit for mounting the longest-running follow-up study of a whiplash group in history. In research, long-term follow-up is definitely a good thing, but there are logical and logistic limits to all good things.

References

1. Giles LGF, Muller R. Chronic spine pain: a randomized clinical trial comparing medication, acupuncture, and spinal manipulation. *Spine,* 2003;28(14):1490-1502.
2. Muller R, Giles LG. Long-term follow-up of a randomized clinical trial assessing the efficacy of medication, acupuncture, and spinal manipulation for chronic mechanical spinal pain syndromes. *J Manipulative Physiol Ther,* 2005;28(1):3-11.
3. Rooker J, Bannister M, Amirfeyz R, Squires B, Gargan M, Bannister G. Whiplash injury: 30-year follow-up of a single series. *J Bone Joint Surg Br,* Jun 2010;92(6):853-855.