CLAIMS FOR GLANDULAR SUPPLEMENTS STILL UNSUBSTANTIATED

Donald R. Davis, Ph.D., Thomas W. Sager, Ph.D.,
Stephen Senn, Ph.D., Douglas G. Altman, D.Sc., and Clement J. McDonald, M.D.

In 1993 Santoro and Weyhreter claimed that glandular supplements had significant normalizing effects on deviantly high and low serum levels of ALT enzyme, CO_2, and thyroid hormone (1). To support their claim, they offered empirical evidence that groups selected for high pre-treatment serum levels had lower levels after treatment than before. Similarly, groups selected for low pre-treatment serum levels had higher levels after treatment than before. In their experimental design, there were a number of comparisons between pre- and post-treatment values, for which there were no control groups. But the design also included a few comparisons between a treatment group and a paired control group. Both groups in a pair were selected for serum levels that were deviant in the same direction (high or low), for the same substance (ALT, CO_2, or thyroid). Treatment groups were given glandular supplements, diet and exercise. Control groups received only diet and exercise. Experimental results showed that both treatment and control groups became less deviant, with the treatment groups showing more improvement than corresponding control groups.

Unfortunately, the experimental design was compromised in two crucial respects:

1) The treatment groups were significantly more deviant in serum levels than the corresponding control groups.
2) Throughout, the experiment failed to anticipate or control for confounding by the statistical phenomenon known as "regression to the mean" (RTM).

The upshot of these two effects is that improvements in serum levels comparable to those seen by the authors are expected even in the absence of any treatment whatsoever, in all their reported comparisons, both with and without control groups.

Davis noted these problems in 1994 (2). After a number of private communications to the authors from Davis (January 6, 1994; March 14, 1994; April 18, 1994 and June 26, 1997) and from Davis and Sager (October 13, 1998 and November 9, 1998), the authors responded publicly in late 1998, joined by a statistical consultant (3). Using the same data as in the 1993 article (but excluding inappropriate duplicate data for some subjects), they offered amended statistical analyses that purported to reaffirm the effectiveness of glandular treatments. Unfortunately, the amended analyses remain as compromised as the originals.

Confounding by Regression To The Mean

To understand intuitively* how unequal deviancy in the sample groups interacts with RTM to compromise these results, consider the high-ALT case, by way of illustration. In the high-ALT treatment group, initial serum levels vary from 29

*For the technically inclined, we assume, if the null hypothesis of no treatment effect is correct, that the population of pre-treatment measurements has the same distribution as the population of post-treatment measurements, that individuals in the pre and post populations are independent of each other but share a common correlation between the pre and post measurements, and that each individual has an individual-specific true level that is measured with zero-mean error and a common standard deviation at each measurement.

to 633 with a mean of 99. Normal ALT levels are about 23.5, according to the authors' data. It is understandable to test the treatment on individuals with elevated ALT levels. After all, those are individuals whom the treatment is intended to help. However, selection of individuals with high *measured* ALT levels does not necessarily select individuals with ALT levels *truly* that high, because of unavoidable biological fluctuations and laboratory uncertainties. Suppose we apply the term "false high" to an individual with measured ALT above his true, average ALT level, and "false low" to the reverse. Clearly, most false highs are found above the middle of the measured ALT distribution, and their incidence increases with increasing measured ALT level. False lows are prevalent below the middle of the measured ALT distribution, and their incidence declines with increasing measured ALT value. (Indeed, if the data are numerous, the highest measured ALT is almost certainly a false high, and the lowest measured ALT is almost certainly a false low.) Thus, selection of individuals for high measured ALT levels, as in the authors' high-ALT treatment group, nets many false highs, but few offsetting false lows. Upon subsequent remeasurement, individuals with false highs (and lows) revert to values more evenly distributed around their true values, thus lowering the measured mean of the entire high-ALT treatment group. The lowering of the mean of this group will occur whether or not there has been any intervening treatment. It reflects the elimination of the biased high measurement errors with which some individuals in the group were initially endowed by virtue of their method of selection.

By contrast with the high-ALT *treatment* group, the high-ALT *control* group varies initially from 24 to 37, with a mean of 29. Thus, the control group is not nearly as extreme as the treatment group, and therefore contains relatively fewer and smaller false highs, and hence will decline less upon remeasurement. The decline in remeasured mean of a group selected for initial high measurements is one manifestation of a more general phenomenon called regression to the mean (4-9). RTM effects are not small. Studies of RTM in clinical contexts show that RTM alone, without treatment, is capable of producing apparent improvements of the same magnitude as reported by the authors (10-12). Some workers now think that much of what is called the placebo effect is really RTM (11,13-15). This is one reason why randomization of study subjects is so important (not done by the authors).

In summary, the main features of the authors' results (improvement in both treatment and control groups, and more improvement in treatment than control) can be explained by RTM and greater selection bias in the treatment group than in the control. Similar remarks apply to the low-ALT groups and to the CO_2 and thyroid cases.

Authors' 1998 Reply

Therefore, the burden of proof lies upon the authors to show why their results should not be considered entirely the effects of selection bias and RTM. In their 1998 reply (3), the authors seem to address this issue. Their reanalysis of their original data purports to show 13 statistically significant normalizing effects. Nine of these tests are comparisons to baseline, in which post-treatment/pre-treatment differences are assessed *without control groups*. Four of the tests are post-treatment/pre-treatment assessments with control groups, using the experimental design described earlier in this letter. The statistical tests are standard nonparametric tests of differences (paired sign and Mann-Whitney tests). Unfortunately, none of the nine baseline tests corrects for RTM or selection bias. In fact, the statistical theory underlying the paired sign test explicitly assumes that there is no selection bias, that is, that the pre-treatment measurements are random samples. Use of nonparametric procedures instead of parametric t-tests does not avoid these problems. The four Mann-Whitney tests with control groups would have corrected for RTM and would have been valid comparisons if the control and treatment groups had been similar at baseline. However, as noted earlier, the four treatment groups were considerably more deviant

than the corresponding control groups, by average factors of 4.2, 3.2, 1.9, and 14.5. Thus selection bias is much more severe and the RTM effect stronger in the treatment groups than in the control groups. Therefore, the statistical assumptions underlying these four tests are not met. We cite the authors' consultant (16):

"If confounded effects are to be avoided, the control and treatment groups must be similar with respect to any characteristic that could affect the results . . . If data are based on an experiment with fundamental design flaws, nothing can be done to correct the damage resulting from the confounded effects . . . When this sort of confounding is present, no statistical method can repair the damage, and the experiment has no value."

Actually, there are statistical methods designed to help repair the damage of uncontrolled RTM that were brought to the authors' attention (17,18). But the authors do not report using them.

Adjustment for Covariate

The authors' analyses implicitly assume that their subjects' baseline readings are unbiased measurements of serum levels. We have argued above that this assumption is not correct. But even if we adopt their assumption, their "statistically significant" differences become insignificant upon properly adjusting for initial baseline differences. As noted above, the authors' control and treatment groups differ markedly in baseline serum levels. When experimental groups differ according to some covariate, it is necessary to take those covariate differences into account to reduce bias, for example, in most epidemiological studies and sometimes in randomized clinical trials (when randomization fails to create well matched experimental groups). Also, this accounting is regarded as essential for cutoff designs (like the authors') where subjects are allocated to treatment according to measured pre-treatment (baseline) values (19,20).

The statistical methodology is called analysis of covariance and amounts to regressing the post-treatment measurement on a binary indicator of membership in the control or treatment group and the baseline measurement. We applied this methodology to the authors' data for ALT (low and high), CO_2 (low and high) and thyroid (low), after log transformation of the pre- and post-treatment values. When we evaluate changes after treatment as the authors did—without including the baseline values in the model—we find that all 5 post-treatment changes are larger in the treatment groups than in the control groups ($P < 0.001$ to 0.017). But when we include baseline values in the model, in every case the change after treatment is no longer statistically significant ($P > 0.05$ to 0.6). *In other words, in the authors' data, there are no differences in outcome between the test and control groups that are not attributable to the baseline differences between those groups.* There is no need to postulate any effect of the authors' treatment. The results of this covariance analysis are fully consistent with the RTM explanation presented above. We note also that the covariance analysis can be applied even when the baseline measurements are subject to measurement error and selection bias resulting from a cutoff design like the authors' (19,20). The key is to model the covariate relationship correctly.

Time Reversal Perspective

Appreciation for the importance of dealing properly with RTM in the experimental design for clinical interventions can be enhanced by a clever time-reversal analysis, with sometimes amusing results. Suppose it were proper to demonstrate the effectiveness of an intervention by showing that subjects selected for high pre-treatment levels subsequently register lower levels after treatment. Then it must also surely be proper to expect that subjects on treatment and selected for high post-treatment levels would register still higher levels when taken off treatment. Now, few experiments actually take subjects off treatment. Nevertheless,

the effect of taking subjects off treatment can be gauged by reversing the usual time order of analysis, because each subject was on treatment at the time of the second measurement and was off treatment at initial measurement. RTM predicts that, if an intervention is ineffective (or if there is no intervention), not only will high pre-treatment subjects decline when given treatment, but also high post-treatment subjects will decline when taken *off* treatment. We would thus have the amusing "proof" that putting subjects on treatment is effective and—by reversing *pre* and *post* labels—taking them off treatment is also effective! Thus, a check on the reality of a treatment's effectiveness can be obtained by reversing the time order of analysis. If similar results are obtained in time-backward as in time-forward analysis, then RTM and selection bias dominate the results to the exclusion of a real intervention effect.

Practical application of the time-reversal test is not often possible, because one should have relatively complete pre- and post-treatment distributions, and pre-treatment subjects selected for high levels manifestly do not constitute a complete distribution. Nevertheless, when time-reversal has been subjected to a proper empirical test, the results agree with the predictions of RTM (11). Most of the Santoro-Weyhreter data do not lend themselves to time-reversal analysis, because of incomplete distributions. However, in two cases it is possible to approximate a complete distribution by combining their low data with corresponding high data. For example, when the low-ALT control group is combined with the high-ALT control group, it appears that a relatively complete distribution is obtained. When the high-ALT control group (subjects with ALT > 23.5) goes on the diet/exercise control treatment, its mean declines from 28.7 to 27.0. But when the subjects with post diet/exercise ALT levels exceeding 23.5 are "taken off" diet/exercise, their mean ALT also declines from 28.9 after diet/exercise to 26.1 before diet/exercise! Similar observations hold for the combined low-CO_2 and high-CO_2 control groups. These time-reversal analyses strongly suggest significant selection bias and RTM in the control groups. We expect that the authors' more extreme treatment groups have even more selection bias and RTM. Unfortunately, combining corresponding low and high treatment groups results in seriously incomplete distributions, so time-reversal cannot be applied to them (or to the spurious example in the authors' 1998 letter (3)).

Conclusion

In conclusion, we find that the authors' claims for therapeutic effectiveness of glandular supplements remain unsubstantiated. Neither their original nor amended analyses deal appropriately with the selection bias and regression to the mean that compromise their experimental design. The burden of proof still lies upon them to demonstrate why their results should not be considered entirely the effects of selection bias and regression to the mean. The authors have agreed that a "randomized, double-blind, placebo-controlled design is necessary" to evaluate their hypothesis (3). Have they done such a study during the 6 years since this need was first noted (2)?

Donald R. Davis, Ph.D.
Biochemical Institute
The University of Texas
Austin, TX 78712
d.r.davis@mail.utexas.edu

Thomas W. Sager, Ph.D.
Director, Center for Statistical Sciences
Department of Management Science
and Information Systems
The University of Texas
Austin, TX 78712

Stephen Senn, Ph.D.
Department of Statistical Science
University College London
London WC1E 6BT, United Kingdom

Douglas G. Altman, D.Sc.
Director, ICRF/NHS Centre for
Statistics in Medicine
Institute of Health Sciences
Oxford OX3 7LF, United Kingdom

Clement J. McDonald, M.D.
Director, Regenstrief Institute
Indiana University School of Medicine
Indianapolis, Indiana 46202

REFERENCES

1. Santoro RL, Weyhreter AF. Support of human gland/organ function with raw protein concentrate as measured by improvements in serum chemistry values. *J Appl Nutr* 1993; 45:48-60.
2. Davis DR. "Glandular" supplement study disputed [letter]. *J Appl Nutr* 1994; 46:62-4.
3. Santoro RL, Weyhreter AF [letter]. *J Appl Nutr* 1998; 50:48-52. Consultant Susan Shott's full agreement is cited by the authors and the editor.
4. Anderson B. *Methodological errors in medical research.* Oxford: Blackwell Scientific, 1990:98-108.
5. Newell D, Simpson J. Regression to the mean. *Med J Austral* 1990; 153:166-8.
6. Stigler SM. Regression towards the mean, historically considered. *Stat Meth Med Res* 1997; 6:103-14.
7. Bland JM, Altman DG. Some examples of regression towards the mean. *Br Med J* 1994; 309:780.
8. Senn S. Regression to the mean. *Stat Meth Med Res* 1997; 6:99-102.
9. Campbell DT, Kenny DA. *A primer on regression artifacts.* New York: Guilford Press, 1999.
10. Yudkin PL, Stratton IM. How to deal with regression toward the mean in intervention studies. *Lancet* 1996; 347:241-43.
11. McDonald CJ, Mazzuca SA, McCabe GP. How much of the placebo 'effect' is really statistical regression? *Statistics in Med* 1983; 2:417-27.
12. McDonald CJ, McCabe GP. [Letter of correction]. *Statistics in Med* 1989; 8:1301-2.
13. Kienle GS, Kiene H. Placebo effect and placebo concept: a critical methodological and conceptual analysis of reports on the magnitude of the placebo effect. *Alt Ther Health Med* 1996; 2:39-54.
14. Kienle GS, Kiene H. The powerful placebo effect: fact or fiction? *J Clin Epidemiol* 1997; 50:1311-8.
15. Sech SM, Montoya JD, Bernier PA, et al. The so-called "placebo effect" in benign prostatic hyperplasia treatment trials represents partially a conditional regression to the mean induced by censoring. *Urology* 1998; 51:242-50.
16. Shott S. *Statistics for health professionals.* Philadelphia: WB Saunders, 1990: 1-4.
17. Lin HM, Hughes MD. Adjusting for regression toward the mean when variables are normally distributed. *Stat Meth Med Res* 1997; 6:129-146.
18. George V, Johnson WD, Shahane A, Nick TG. Testing for treatment effect in the presence of regression toward the mean. *Biometrics* 1997; 53:49-59.
19. Senn, S. *Statistical Issues in Drug Development.* New York: Wiley, 1997: 81-85, 104-105.
20. Trochim WM, Cappelleri JC. Cutoff assignment strategies for enhancing randomized clinical trials. *Control Clin Trials* 1992; 13:190-212.