Comments on the Papers Presented in the Session on Causal Inference with Problematic Control Groups

Prof. Joseph L. Gastwirth
Department of Statistics
George Washington University
Washington, DC 20052

After reviewing some general background concerning omitted variables, and Cornfield's inequality, sensitivity analysis and matched data (Greenhouse, 1982; Gastwirth, 1988; Peters 1941; Belson, 1956;Rosenbaum, 2002 and Cochran and Rubin, 1973; available from the author) we suggested that a control or comparison group is *problematic* if one or more of the following occur:
a) There is an insufficient overlap in the distributions of the major covariates in the two groups,

b) If there is a major covariate that is *known* but information about it in the group is unavailable to us,

c) Errors occur in classifying the response that are related to treatment assignment, i.e. they are not "random" with respect to group status and the major covariates, or

d) Some members of the control group received some degree of treatment from another source. This is more likely to occur in case-control studies, e.g., in studying the risk of exposure to second hand smoke one can find highly exposed individuals (long-time spouses of 2+-pack a day smokers, say) but where do you find individuals who ***never*** were exposed to second hand smoke?

The paper by Prof. Marcus and her colleagues concerns a bias in the classification of the response of the subjects by the rater due to information inadvertently given them by the subject, e.g., told about a side effect, which would indicate the particular drug the patient was given. Essentially they carry out a sensitivity analysis to determine the proportion of ratings that could be due to a bias resulting from the raters learning whether a patient received an active treatment. They conclude that the amount of this type of rater bias is too small to change their finding that therapy is helpful. Their result is very much in the spirit of Cornfield's and its generalizations.

Two questions about their analysis did occur to me. In the draft I received in Table 2 the difference between the two response percentages (placebo and treated) are reported but the p-value of the test of significance was not. While I don't believe statistical significance at the .05 level is a magical talisman, it would assist the reader to see how the p-value is affected as the number of possibly biased (due to unblinding) raters' increases. Secondly, the authors combined all active treatment groups into a single one to compare with the placebo. It would seem to me that there should be a partial ordering (at least) of the expected responses since individuals receiving a drug and CBT should do better than subjects receiving only one of the two types of treatments. If this is scientifically sound, a trend test should be carried out as it is a directed test and is more powerful (Agresti, 2002) than a simple 2x2 analysis.

The paper by Rubin and Stuart (2005) proposes an approach to draw controls from two potential control groups, each of which is problematic by itself. The reason they are problematic is shown clearly in their Figure 1. The covariate distribution of each group does not have a sufficient overlap with the treated group to obtain a reliable inference. The usual use of multiple control groups is to account for possible covariates that one may not be able to obtain accurate information on. An important set of studies relying on multiple controls that predates most of the cited references established the Reye syndrome—aspirin association (see Gastwirth, 1988 p.917-8). The Ohio study (1982) that compared cases to control group one: children who were classmates and were sick at the same time as the case and a second of other children of the same age who were sick the same week and lived in the area showed an association. When the FDA wanted to issue a warning label, the industry claimed that parents whose children came down with Reye's syndrome and had heard of the association but were uncertain of the medication they gave their child might answer aspirin. The industry suggested two other control

groups consisting of parents who were under stress; children in a nearby hospitals. The pilot study showed an even higher relative risk from aspirin use in those two control groups; the public was warned and the incidence of the disease dropped sharply.

The authors refer to a companion paper that I have not seen, which appears to improve on the adjustment achieved by simply assuming a constant difference in the effect on response due to the specific control group the chosen control comes from (the estimate of their parameter δ). I look forward to reading it.

I believe their approach is quite promising; however, I would have appreciated some comparisons with other methods. For example:

a) one could use a Peters-Belson regression model fit in the controls from group 2 in the region -2 to +1.5 of Figure 1 and a similar approach using the controls from group 1 in the region 1.5 to 3.5 or 4 of the X-space (see their Fig.1).

b)Alternatively, one might explore whether one can combine matching and regression as in logistic regression with matched case control studies (Gail et al. 1982; Breslow and Day, 1980) so that one uses fewer covariates in the matching process and adjusts for the effect of the others at the analysis stage. Of course, if one simply does not have many controls with similar values of an important covariate, regression methods would not work due to the problem of extrapolation.

c) One could use the "nearest" matches to each treatment group to conduct a Wilcoxon test for each member of the treatment group for whom "reasonably close" matches can be found and combine these in the manner proposed by Bhattacharya and Zhou (1997).

My impression is that their technique could also be used with more than two control groups and I look forward to seeing more uses of their method along with a comparison of the results it yields with other approaches.

The paper presented by Prof. Petkova concerns the problem of classifying patients into categories reflecting their response to treatment using repeated measurements of their health status. Thus, it less concerned with issues created by not having a well-matched control group, but deals with non-compliance. This may lead to misclassification of a subject's status as a control or treatment group member.

What struck me most about the data was the seemingly high level of missing data. Their analysis omits 89 of 139 or 64.03% of the subjects assigned to placebo arm of the study and 65 of the 127 or 51.18% of the participants assigned to treatment (phenelzine). This difference of about 13% in participation rates is *statistically significant* (both Fisher's conditional and the unconditional tests in STATXACT yield a p-value around .03). In many areas one does not see such a large fraction of missing data. For example, in EEO cases missing values might affect about 25% of the records and in government economic and health surveys the response rates are 75% or higher except perhaps on questions concerning specific types of income. Thus, a sensitivity analysis seems to be essential but it should be one that incorporates prior subject matter knowledge. Why do people miss over half of their appointments? Are they likely to have improved so much they don't feel the need to see their doctor? Did the drug or placebo have so little effect that they don't think they should bother any more? Did the drug have a serious side effect?

This suggests to that similar future studies should focus at least as much on improving the subject participation rate, perhaps by including a system for contacting a patient who misses an appointment and re-scheduling them as soon as possible. If the patient misses that appointment, one should try to find out the reason why. Just as in the strengthened Cornfield inequality where the required prevalence of the OV in the smokers more than doubled when we incorporated information about the plausible relative risk of the OV, including knowledge of the likely range of response patterns of the subjects who miss so many appointments might enable the authors to obtain sounder conclusions about the efficacy of the treatment as well as distinguishing among the various types of placebo and treated responders

Appendix

To appreciate how important it is to *carefully* classify the responses, I would like to compare my analysis (Gastwirth, 1988, p. 625-31) of an early matched pair study of similarly qualified "testers" in the *Youritan* equal housing case with

that in a well-regarded text on Statistical Proof of Discrimination (Baldus and Cole, 1980). The opinion reported the treatment for paired testers who visited an apartment complex on the same day to control for who was vacant apartment.

The analysis in Baldus and Cole (1980, p. 236) focused only on whether each member of the pair was told an apartment was available and points out that 6 of 14 blacks (43%) were treated favorably versus 11of 14 (78%) whites. They analyzed the matched data obtaining a p-value of .125 and question the judge's reliance on the difference of 35% in the two percentages as the difference was not significant.

The opinion reported that even when both members of the pair were told an apartment was (or was not) available often the black member was given less encouragement either by being told a background credit check was needed or there would be a longer wait for an available apartment. Incorporating this information into the classification of the treatment received by the testers all the concordant pairs disappear and one has a table with entries:

		White	
		Y	N
	Y	0	13
Black			
	N	1	0

Now the probability of observing only 0 or 1 heads in 14 tosses of a fair coin = .000916 so the two-sided p-value is .00183, clearly < .05. Thus, the judge appears to have been correct in giving the statistics substantial weight! This example illustrates that properly classifying the response or differential response can be crucial.

References

Agresti, A. (2002), *Categorical Data Analysis* (2d ed.), New York: John Wiley.

Baldus, D. A. and Cole, J.W.L. (1980), *Statistical Proof of Discrimination*, Colorado Springs, CO: Shepards/McGraw Hill.

Belson, W.A. (1956). A technique for studying the effects of a television broadcast, *Applied Statistics*, **5**, 195-202.

Bhattacharya, P.K. and Zhou, P.L. (1997). Semiparametric inference in a partial linear model. *Annals of Statistics*, **25**, 244-262.

Breslow, N. and Day, N. E. (1980). *The Analysis of Case Control Studies*. Lyon, France: IARC.

Cochran, W.G. and Rubin, D.B. (1973), Controlling bias in observational studies: A review, *Sankhya, Series A,* **35**, 417-446.

Gail, M.H., Lubin, J.H. and Rubinstein, L.V. (1981). Likelihood calculations for matched case control studies and survival studies with tied dead times. *Biometrika*, **68**, 703-707.

Gastwirth, J.L. (1988), *Statistical Reasoning in Law and Public Policy*, San Diego: Academic Press.

Gastwirth, J.L. and Greenhouse, S.W. (1995), Biostatistical Methods in the Legal Setting, *Statistics in Medicine*, **14**, 1641-1657.

Gray, M.W. (1993). Can statistics tell us what we do not want to hear? The case of complex salary structures. *Statistical Science*, **8**, 144-158.

Greenhouse, S.W. (1982). Cornfield's contributions to epidemiology. *Biometrics*, **38S**, 33-46.

Peters, C.C. (1941), A method of matching groups for experiments with no loss of populations, *Journal of Educational Research*, **34**, 606-612.

Rosenbaum, P.R. (2002), *Observational Studies* (2d ed.), New York: Springer.

Rubin, D.B. and Stuart, E.A. (2005), Matching with multiple control groups and adjusting for group differences, *2005 Proceedings of the American Statistical Association* , Section on Health Policy Statistics [CD-ROM]. Alexandria, VA: American Statistical Association.

Scott, E.L. (1979). Linear models and the law: Uses and misuses in affirmative action. *Proceedings of the Social Science Section of the ASA*.

Yu, B. and Gastwirth, J.L. (2003). The 'reverse' Cornfield Inequality and its use in the analysis of epidemiologic data. *Statistics in Medicine*, **22,** 3383-3401