Aust NZ J Obstet Gynaecol
1991; 31: 4: 346

When Predictions Don't Predict

Andrew L. Speirs[1], FRACOG, FRCOG, Ricardo H. Asch[2], MD and Sherman J. Silber[3], MD
Royal Women's Hospital, Melbourne

EDITORIAL COMMENT: *We accepted this paper to remind readers that the rules of clinical common sense as well as the results of statistical calculations are needed to judge published results of clinical trials. Statistical rules for design of a trial should be followed, and advice of a statistician sought before rather than after a study begins. It seems self-evident that if the raw data does not show a clinically meaningful difference in the results obtained, that statistical gymnastics cannot save the day; for example if a series of patients with prolonged pregnancy is large enough, then a difference in the perinatal mortality rate of 0.2% between conservative treatment and induction of labour will be statistically highly significant, yet the difference may be too small to convince clinicians that one regimen is superior to the other.*

Summary: One of the most widespread abuses of statistical methods involves the misinterpretation of statistical significance after the application of multiple comparisons. This paper highlights the problem by demonstrating 'significance' when clearly none exists.

Papers submitted to medical journals commonly follow a similar theme:

> '15 semen parameters were studied in patients having in vitro fertilization (IVF) and the following factors were found to predict pregnancy, $p < .001$ . . .'

There are vital weaknesses in such an analysis. By looking for relationships between a large number of factors and the occurrence of pregnancy one can fall into the trap of thinking a relationship exists, 'proving' it with $p < .001$, and yet there is no relationship. As an example we could take the letters of the patient's surname and use them to predict IVF success, when of course common sense tells us that there can be no such predictive connection.

MATERIALS AND METHODS

Twenty-eight consecutive azoospermic men undergoing microsurgical epididymal sperm aspiration (MESA) for IVF were divided into 2 groups: pregnancy with a livebirth ($n = 7$, 25%) and no successful pregnancy. The occurrence of various letters of the alphabet in the surname of the patient was studied to identify letters which were important in predicting pregnancy. Such an idea would at face value seem absurd, but we thought it worth studying since the same statistical methodology is commonly used to seriously evaluate factors from a large pool of possibilities which might predict pregnancy.

1. Reproductive Biology Unit, The Royal Women's Hospital, Melbourne, Australia.
2. Department of Obstetrics and Gynaecology, U.C. Irvine, USA.
3. St. Luke's Hospital, St. Louis, USA.

Address for correspondence:
Andrew L. Speirs,
Reproductive Biology Unit,
The Royal Women's Hospital,
Melbourne, Australia, 3053.

RESULTS

To preserve confidentiality the actual letters of the patient's surnames have been rearranged in alphabetical order. These codified names, the important letters

Table 1. Patient Names with Letters in Alphabetical Order, Outcome of MESA Treatment and Results of Analysis for the Presence of the Letters G, Y and N

		Name includes		
Alphabetized surname	Live birth	G	Y	N
aceeghmrrtuz	yes	yes	—	—
aabcehlntu	yes	—	—	yes
abeefgimnu	yes	yes	—	yes
agiinort	yes	yes	—	yes
abeily	yes	—	yes	—
aeenswy	yes	—	yes	yes
dlnu	yes	—	—	yes
aeeklm	—	—	—	—
achrstwz	—	—	—	—
himst	—	—	—	—
ceelmo	—	—	—	—
beeeginrs	—	yes	—	—
ablu	—	—	—	—
ehioorsv	—	—	—	—
cox	—	—	—	—
abelmnss	—	—	—	yes
ehnrtu	—	—	—	yes
abck	—	—	—	—
cdeehilr	—	—	—	—
emors	—	—	—	—
bbgis	—	yes	—	—
ceklu	—	—	—	—
ehhiks	—	—	—	—
aaegilnortt	—	—	—	yes
abdnnor	—	—	—	yes
aahlmpsuu	—	—	—	—
adeilwz	—	—	—	—

Table 2. Summary of 'GYN' Analysis Related to Successful Outcome with MESA Treatment

	n	GYN positive
Livebirth	7	7 (100%)
Not pregnant	21	6 (29%)

p = .003 (Fisher exact test)

studied and the outcomes are shown in table 1. All patients with a successful pregnancy had surnames containing the letters G, Y or N whereas only 6 (29%) of the unsuccessful patients had the letters G, Y or N in their surnames. These results are summarized in table 2. Thus the presence of a G, a Y or an N in the surname of patients undergoing sperm aspiration and IVF was highly predictive of a successful result.

DISCUSSION

Had this been a study of proteins in follicular fluid or the many subtleties of sperm morphology and motion, nondiscerning readers would be trying to assimilate the findings of this paper into their understanding of reproductive physiology. Instead we have chosen to study parameters which clearly could have no usefulness in predicting pregnancy. How then can the presence of the letters G, Y or N in the surname of patients ('GYN positive') so convincingly predict pregnancy? It is *not* that statistical methods can be used to prove almost anything, but rather that gross abuses may not be obvious to the nonstatistician.

The misleading conclusions of this study of patients' names (and of many other papers which claim scientific merit) are brought about by failing to recognize the effect of multiple statistical comparisons, i.e. data dredging (1) or data snooping (2). To understand this effect, recall that p = .01 means that if there is really *no difference between 2 populations* under study (for example the presence or absence of the letter K in the surnames of *all* MESA IVF patients does not predict pregnancy) then there is 1 chance in 100 of finding such different pregnancy rates related to the surname letter K in a *sample* of MESA patients. There is therefore a probability of .99 of finding no difference at the p = .01 level. However, when many variables (such as *all* 26 letters of the alphabet) are studied with a 'lets go hunting' approach, the probability that there will be no p = .01 significance in any of them would be $.99^{26}$. This computes to .77. Therefore in 23% of such studies of 26 variables a p = .01 'significance' will turn up somewhere by chance alone, not only in 1% of studies as suggested by p = .01. Thus when there have been multiple comparisons of many different variables it is difficult for the reader to judge the statistical significance of those reported. By studying groupings of any 3 letters in the surnames of these patients (as well as single and 2 letter groups) we had some 3,000 combinations for analysis and were virtually certain to find p = .003 somewhere. We were pleasantly surprised to find 'significance' with the interesting letters G, Y and N. If these letters had not worked we could have continued to try others and eventually found a combination that was 'predictive' of pregnancy.

The potential for deceit in data analysis is not removed by prospective studies; it is just as easy to study multiple variables prospectively and only report those that are 'statistically significant'. The soundest approach is generally to identify variables of interest in a pilot study, and then to test the hypothesis in a separate (preferably prospective) study. None of the pilot study cases should be included in the definitive study.

A related presentation of the multiple-analysis deception occurs when the cut-off level of some studied parameter is chosen to maximize the statistical significance after viewing the experimental data. If many statistical analyses could have been (or even were) undertaken with nearby cut-off points with little statistical significance, it is deceptive to report only the 'best' result when selected retrospectively. If all the data is presented it may be obvious that there are many cases close to the chosen cut-off. In such cases, repeating the statistical test with the borderline cases transferred to the other group may provide a different perspective.

As data accumulates in a clinical trial a researcher may be tempted to analyze at monthly intervals, ceasing trial as soon as 'statistical significance' is found. Employing the usual methods of statistical testing in such a manner will result in a deceptively (3) significant p value for reasons related to the 'multiple analysis effect' discussed earlier. Techniques are available for appropriate analysis of such trials (4).

Papers which employ tactics described in this paper should not be considered meaningless. Rather the statistical significance reported should be viewed as dubious. It may well be true that the conclusions reported are valid, but confirmation would require further study.

Understanding the effect of multiple comparisons should improve the interpretation of data, the quality of papers submitted for publication and the ability of the reader to discern fact from statistical fiction.

References

1. Armitage P. Statistical Methods in Medical Research. Blackwell Scientific, Oxford, 1977. p 205.
2. Colquhoun D. Lectures on Biostatistics. Clarendon Press, Oxford, 1971. p 207.
3. McPherson K. Statistics: The problem of examining accumulating data more than once. N Engl J Med 1974; 290: 501-502.
4. Demets DL. Practical Aspects in Data Monitoring: A Brief Review. Stat Med 1987; 6: 753-760.