Drug Development and Industrial Pharmacy, 23(8), 851-853 (1997)

COMMUNICATION

Flawed Statistical Analyses and Representations of Fact

Sanford Bolton

Visiting Professor, University of Arizona 5495 N. Via Velazquez, Tucson, Arizona 85750

Introduction of generic drug products represent an economic burden, no less a threat, to established brand products. This fact was no doubt largely influential in establishing strict FDA guidelines and bioequivalence studies to prove equivalence. These guidelines and regulations have been honed during the past 20 years as a cooperative effort between the FDA and both brand name and generic drug manufacturers. Nevertheless, the fiercely competitive drug market has provoked some major drug companies to take extreme measures to retain their market share. This is evidenced by recent efforts of the manufacturers of premarin and coumadin to convince the FDA and local regulatory bodies to expand the regulatory requirements for approval of generic products. In the latter and more recent case (warfarin), the manufacturer argues that evidence exists suggesting that the switch to an alternative product results in lack of reproducible clinical effects and adverse events. Their sole evidence to date (after 40 years of marketing coumadin) is a study published in 1988 (1), which unfortunately is replete with shortcomings in statistical assumptions, analysis and study design. This brief commentary is meant to address some common statistical and design problems that are exemplified in the aforementioned study. It is unfortunate that such flawed studies appear in the scientific literature as fact, indeed, "proof," which is then used as evidence to promote drug products or to thwart competition.

The paper in question describes a retrospective study consisting of patients seen in Boston City Hospital in 1980 (eight years prior to the publication). Personnel in the anticoagulation clinic observed an increase in poorly controlled patients. They also observed an increase in patient visits and prothrombin testing. This led to the discovery that panwarfarin (Abbott Labs) had been substituted for the previously prescribed coumadin for some patients. These patients were switched back to coumadin, and a retrospective study was performed measuring patient characteristics, numbers of visits, changes in prothrombin times and dosing, etc. The resulting conclusion was that substitution of panwarfarin for coumadin resulted in adverse patient events and undue cost due to overly variable prothrombin times.

Although this study is an example of the usual problems and biases that are often evident in retrospective studies, I will focus only on some specific statistical misunderstandings (misuses) that appear in the analysis of the data which were used to "prove" and document the problems that resulted from substitution of coumadin with panwarfarin (now discontinued); a brand product, not a generic drug. The data consist of an evaluation of prothrombin time values and patient visits during the period of time when some patients were switched from coumadin to panwarfin. Included in the study were 15 patients who were switched to panwarfarin (switch group) and 40 patients who remained on coumadin

(coumadin group). Statistically significant differences were observed for three of a number of measurements reported in the study. One of these three tests was computed incorrectly, and the other two tests used incorrect procedures. The incorrect calculation was a chi-square test which showed 10 of 15 patients in the switch group and 15 of 40 in the coumadin group having a change of dosage during the study period. Even without a continuity correction (which seems appropriate here) and considering the fact that the two suspect patients were included in the switch group, the significance level is greater than 5% (2).

The problems with the other cases of significance are more subtle. Consider the significance level of 0.001 for comparison of the proportion of visits in which prothrombin time was in range for the two groups: 29 in range in 74 visits (39%) for the switch group, and 121 in range in 177 visits (68%) for the coumadin group. The level of 0.001 is surprising, considering the fact that the proportion of *patients* who were in range for the two groups was not statistically significant ($p = 0.07$). What is the problem? The reason for this apparent anomaly is that it is incorrect to pool all of the data from different individuals to compute the significance test. From a statistical point of view, the flaw is that the observations are not all independent. This can be seen easily using an extreme possible outcome based on the data. Suppose in the switch group that one patient was seen 16 times and was out of range 16 times. Thirteen of the remaining 14 patients were each within range once in a total of 58 visits (45 in range of 58 visits = 78%). This is a better proportion than that observed in the coumadin group. One or more extreme patients can bias the proportion. Each patient is an independent unit, and patients are different with respect to their responses to the drug. By combining all of the data from many patients, we lose the distinction between patients; and the dramatically increased apparent sample size (e.g., 15 patients now become 77 observations) results in dramatically increased and exaggerated significance. If, in fact, all patients were identical, then the pooling of all data may be justified. This concept can be illustrated by another fictitious example. Suppose there was only one patient in each group, but each patient was seen many times, 74 and 177, respectively, with the results seen in the published study. Would you say that switched patients showed more out-of-range results than coumadin patients? Of course not! One might say that one patient had a more erratic response (but the reason would not be obvious based only on these data). Therefore, one could not make any judgments on these data without seeing data for individual patients (assuming all patients were selected for inclusion in an unbiased way).

The other significant observation ($p < 0.025$) compared the proportion of months that a patient in the switch group was out of range for prothrombin time prior to switch and during the switch. Apparently, the data were analyzed by computing the proportion of months in which an out-of-range prothrombin time was observed prior to and during the study for each patient, although the exact analysis is not apparent based on the data given (this should be some kind of paired analysis). An obvious severe bias that would be introduced without a comparative group (the coumadin group) is a result of the less frequent visits prior to and after the study. The fact that there were more visits during the 6 months of the study (19% for both groups, as stated in the paper, and perhaps more than that for the switch group) would result in a greater chance of an out-of-range result even if the same product were given in both periods. As noted, the exact statistical analysis is not specified so that one cannot comment on the correctness of the analysis. However, the analysis could be valid if a patient was seen for an equal number of visits each month both before and during the study (which clearly did not happen). Then, one might take the difference of the proportions of months in which "at least one" out-of-range value was observed between the pre-study and study periods. The analysis might test the average difference versus 0, although without a comparative group, any differences could be attributed to the time of year or circumstances that would differ during these two study periods. Since the number of visits per month was not the same before and during the study, some obvious bias could occur (see below). It is curious why the author chose this different method to analyze the data rather than use the "proportion of visit" method in the previous analysis discussed above. One may wonder what the results of such an analysis would reveal. Nevertheless, some other problems that would be expected in the analysis include: i) the elimination of two patients who were included in other analyses for apparently arbitrary reasons; ii) the decision of the duration of time measured prior to the switch (was there a cut-off point?); (iii) there is no comparison to the coumadin group (which also may have had more out-of-range values per months in the study period); and iv) there is no comparison for the switch group following the switch back to coumadin. In the spirit of the problem of independent observations illustrated in the previous example,

we have a potential similar misinterpretation of the data in this analysis. Patients are seen for different periods of time in the pre-study and study periods, with different numbers of visits per month. This immensely complicates any statistical evaluation. Some examples again reveal the flaws in this analysis. If a patient is seen 10 times, once each month for 10 months, the proportion of months in which an out-of-range result will be observed will almost certainly be smaller than if a patient is seen 10 times in 1 month. Since the pre-study data had fewer visits and more time on the average, this extreme example illustrates the possible bias. Suppose that a patient missed some months in the pre-study period. Were those months not counted? Suppose that the patient had a visit on the 1st and 31st of the month. Could this 1-month data be compared with results of visits on the 1st day of the month and the 1st of the following month? If there was a failure only on the 1st day of the 1st month, the former patient would have a proportion of 100% and the latter 50%. Clearly, there is a bias in this evaluation. Although it is not clear how the data were handled, the fact that there were more patient visits in the study period as noted above would bias the results, unless this were somehow accounted for in the analysis.

Although these comments are not complete, one can conclude from the commentary that articles that appear in print do not necessarily imply fact. One could conclude from this study that there is a suspicion of lack of equivalence (without "statistical significance") between two sodium warfarin products (assuming that there was no severe bias in patient selection, as well as other biases to which retrospective studies are prone). However, as is often true with retrospective studies, a carefully controlled, prospective study would be needed to "prove" the hypothesis suggested by the data. The fact that no such inequivalence had been documented elsewhere during the many years of panwarfarin use suggests the importance of a follow-up study before conclusions are made. This is merely another example of the importance of careful examination with a skeptical eye and an open mind when evaluating the literature.

REFERENCES

1. S. Richton-Hewett, S. Foster, and C. A. Apstein, Medical and economic consequences of a blinded oral anticoagulant brand change at a municipal hospital, Archives of Internal Medicine, 148, 806 (1988).
2. S. Bolton, Pharmaceutical Statistics, 3rd ed., Marcel Dekker, New York (1997).