Editorial

The Love-at-First-Sight Effect in Research

PROMISING results and the favorable outcome of an investigation lead naturally to a euphoric response, particularly when they exceed expectations. Frequently, however, other investigators (or sometimes the same investigator) will attempt to reproduce these results, and they see none of the nice things that were first seen. Often, the second time around, slightly different materials or methods are used, or we find the first report was a misreport. But more often, the early enthusiasm is the result of an artifact which reasonable repetition cancels out, and reasonable statistical precaution might have hinted at. There are even some situations in which a repeat *must* give poorer results than the first try. The reason is obvious once it is seen—and yet people have devoted much experimental work to trying to find out why a material "deteriorated"—when the putative deterioration was a pure artifact.

Three examples, derived from widely different research areas, illustrate the problem of the sometimes misleading nature of early work. In the first two, a relatively homogeneous material, based on electrocardiograms of "normal" subjects was chosen.[1] The researcher (call him A) attempted a correlation between the subjects' ages and the various ECG measurements. The spatial magnitude of the maximal QRS vector was selected as the first candidate measurement. There was little reason to believe that the two variables were related, other than indirectly, and a low correlation was expected.

Using 25 records from subjects ranging in age from 20 to 85 years, A found a large negative correlation of −0.63. He was careful, and he looked up the 95% confidence limits on $r = -0.63$, $N = 25$ (the Geigy tables[2] are an excellent source), and found them to be −0.31 to −0.82. This outcome was statistically significant, new, surprising and, as might have been suspected, he loved the result. Publication followed shortly thereafter.

A colleague, B, read about the promising lead in correlating ECG data and age. He collected 50 "normal" ECG records for correlation with age. Using the same ECG measurement, his results turned out to be most disturbing. The correlation coefficient was not more than −0.03. This result did not differ significantly from zero. The 95% confidence limits on r were −0.33 to +0.25. These limits just overlap the 95% limits on r found in A's smaller experiment, and thus these two results might have been considered consistent with each other even though, as a simple single comparison the two sample correlations are statistically significantly different from each other. Many serious discussions and arguments ensued between the two friends. Many colleagues raised not-too-helpful voices. The skeptics rallied to B, the believers to A.

Following up his initial result, investigator A went beyond the QRS complex and attempted a correlation between an S-T vector and age. To guard against criticism, this time he collected 75 "normal" records which yielded a correlation coefficient of −0.53. The 95% confidence limits on this correlation are −0.66 to −0.34. Our friend B, however, had not yet given up completely, but now he limited his efforts to 25 cases. The result was again devastating, yielding only a nonstatistically significant coefficient of −0.15. This time the confidence limits were from −0.51 to +0.26, clearly overlapping A's results. But B didn't look at his material this way. His disappointment was great, and he decided to abandon further ECG studies. The believers apparently had won.

A, however, was a little uneasy. He pursued his correlations with larger samples. His love-at-first-sight reaction alternately cooled

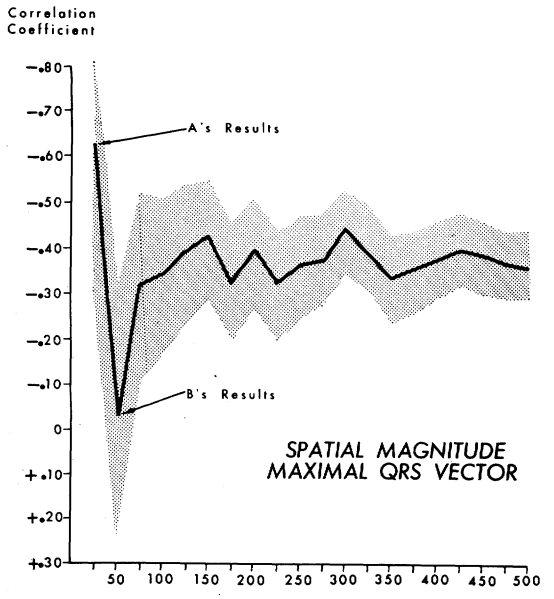

Figure 1

The heavy line indicates correlation coefficients (r) for various sample sizes, shown on the abscissa. The shaded area represents the extent of the confidence limits of r.

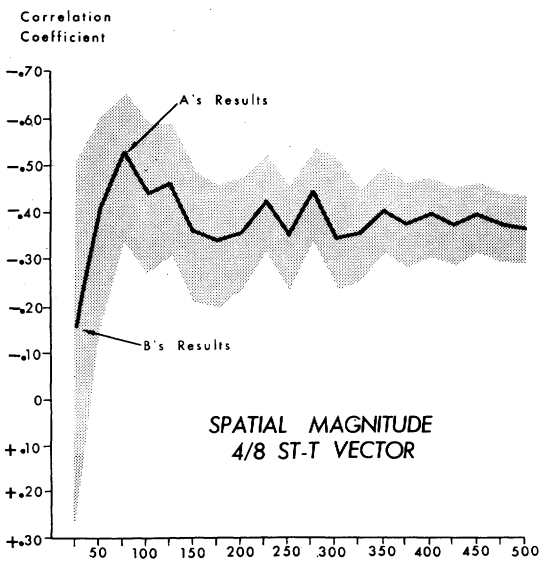

Figure 2

Correlation coefficients and confidence limits for S-T vector.

and reheated. Adding cases led to fluctuations of results with parallel fluctuations in his enthusiasm (figs. 1 and 2). After years of data collection, he was able to obtain 500 "normal" records, resulting in correlation coefficients of −0.36 for QRS (95% confidence limits, −0.43 to −0.28) and −0.37 for ST (95% confidence limits, −0.44 to −0.29). B's early results were as much to the "left" of true as A's were to the "right." Both A's and B's early results had been consistent with the "true" result. What A learned was how very large a sample size he needed to iron out the wide fluctuations. Figures 1 and 2 show the data with 95% confidence bands. B may have learned the power of positive thinking.

The first two examples were ones in which some closer attention to statistical concepts (confidence limits on a correlation coefficient, in particular) would have reduced misunderstanding. They also showed how long (in terms of the numbers of observations) it takes for data really to "settle down."

The third example is from the field of cancer chemotherapy. In this field, the first results which experimenters try to reproduce are almost always "too good." The reasons become obvious once the results are looked at carefully. The results in this kind of experiment are expressed as the ratio of tumor weights of treated animals to the tumor weight of control animals (T/C). The closer this ratio is to zero, the more "effective" the material. The criteria for selection for follow-up are largely biological. The experiments are designed so that a material truly capable of giving a low T/C has a high probability of "passing" the screen. The nature of the screen and its operating characteristics have been described elsewhere.[3] Experimenter B, unhappy in electrocardiography now entered this field. He screened many compounds on two different tumor systems. After much work he found 117 materials that gave low T/C ratios on one system, and 115 others that had low T/C's on the other screen. As shown in columns 2 and 4 of table 1, the results appear most encouraging.* Could he have fallen in love with them at first sight?

*Data supplied by the Cancer Chemotherapy National Service Center. The data are real. Only the names of the investigators have been changed.

Table 1

	Animal Tumor System			
	Lewis Lung		Walker IM	
100 T/C	First trial	Repeat	First trial	Repeat
0-5	2		8	4
-15	1	1	18	14
-25	3	3	22	15
-35	24	2	26	17
-45	37	9	26	8
-55	50	22	15	9
-65		13		2
-75		19		4
-85		13		9
-95		12		8
-105		6		9
> 105		17		16
Total	117	117	115	115
2nd trial as good as or better than 1st:		$\frac{12}{117} = 10\%$		$\frac{27}{115} = 23\%$

As Dr. B had tried to verify Dr. A's work, so attempts were made to reproduce B's results on the 232 new anti-cancer drugs. The retest results, shown in columns 3 and 5 of table 1, look as if B, having once erred by finding too little, had now erred by reporting too much.

What did happen here? Materials were retested because the initial results were good. If a material that really was ineffective perhaps looked "good" the first time because of experimental variations, upon retesting it should do worse. Thus, in the Lewis lung system, 105 of 117 materials did worse on the second trial than on the first (if our sample were really a random sample, we would have expected about a 50-50 split, half better, half worse). In the Walker IM system, 88 of 115 did worse the second time. This would seem to say that most of the "leads" from this screening program were materials that only accidently had low T/C's. Many of the 232 were probably artifacts.

Had materials which seemed poor on the first trial been retested, we would have found a parallel *upward* movement. The materials would seem to get better when retested. But poor-appearing materials are rarely retested. After all, no one wants to waste time and effort in retesting unpromising drugs. Thus, the sample for a second stage experiment in a screening situation is always biased, and the results in repeat studies are almost always poorer. This disheartening result comes about with no gross errors, no incompetence, no dishonest reporting. If the retested materials are biologicals (that is, not well-defined chemicals, but natural products formed through fermentation, and so forth), the producer of the material will often explain the difference between the first "good" test and the second "poor" test by arguing that the material had deteriorated while waiting to be retested. This argument may lead the investigator to attempt to remake the material (go through the fermentation process again, and so on)—usually at much cost, and usually to no avail. If too much hope is hung on the first result, it has the result of being swept away by love-at-first-sight. Sir Francis Galton talked about this many years ago, noting the tendency of extreme results to regress toward the mean—hence "regression analysis."

These examples raise some serious questions not only about the sample size but also the selection of the sample in medical research and reporting. As shown in figures 1 and 2, results of correlation investigations can vary tremendously when small samples are used. Application of statistics with "significant" *t*-tests, correlations, and other embellishments done without real thought and care may appear to strengthen results from inadequate samples. Sometimes, they only obscure the meaning of results. A thoughtful application of proper statistical techniques, tempered by good temper and subject-matter sense, can help clear up apparently conflicting reports.

There are no fixed rules for determining when a sample is sufficiently large to be representative for the population under study. Size of itself is no assurance of representativeness. In the example with correlations from "normal" subjects, fluctuations in the computed correlation decreased as more records were used. After about 100 records the results seemed to become stable. For electrocardiograms with abnormalities, such as left ven-

tricular hypertrophy, as many as 300 records may be needed to obtain relatively stable results.[4] Each pathological entity appears to behave differently in a statistical sense. A usable rule of thumb is that sick individuals are more variable, from one to the next, than well individuals. A sample size adequate to characterize "normals" rather nicely, will often be too small to characterize ill persons equally well.

This leaves us with the question of what inferences you draw when adequate samples cannot be obtained. The abnormality under study may be too infrequent or the experimental procedures too complex to be repeated on large numbers of subjects. Usually, the validity of results can be tested by an independent second sample, even of small but roughly equal size. This obvious procedure is unfortunately not used very often in clinical investigations. If duplicate experiments cannot be mustered, it would be preferable to report the available observations without claims for generalization. If statistical support is needed, it is possible to combine the probabilities[5] from several experiments, to see if they add to meaningful results. Valid observations on small samples from various sources may still add up in the literature and can result in a fairly reliable picture. They become misleading only when overstated. On the other hand, one must not over-rely on statements of statistical significance. In a screening situation large numbers of compounds are screened. Many of these looked promising, only by accident. Those selected for retesting may only represent the 5% (or 1%) of statistically significant results when no real difference exists. To stake too much on this first look is to succumb to love-at-first-sight. Sometimes the first result holds up. Far too often it leads to disillusionment. Being prepared in one's mind for a "regression" back toward no effect can reduce the possible disillusionment, and free an investigator for more fruitful work.

HUBERT V. PIPBERGER
MARVIN A. SCHNEIDERMAN
JACK D. KLINGEMAN
Washington, D. C.

References

1. PIPBERGER, H. V., GOLDMAN, M. J., LITTMANN, D., MURPHY, G. P., COSMA, J., AND SNYDER, J. R.: Correlations of the orthogonal electrocardiogram and vectorcardiogram with constitutional variables in 518 normal men. Circulation **35**: 536, 1967.
2. DIEM, KONRAD (Ed.): Documenta Geigy, Scientific Tables, ed. 6. Ardsley, New York, Geigy Pharmaceutics, Division of Geigy Chemical Corp., 1962.
3. ARMITAGE, P., AND SCHNEIDERMAN, M. A.: Statistical problems in a mass screening program. Ann NY Acad Sci **73**: 896, 1958.
4. KLINGEMAN, J., AND PIPBERGER, H. V.: Computer classifications of electrocardiograms. Comput Biomed Res **1**: 1, 1967.
5. FISHER, R. A.: Statistical Methods for Research Workers, ed. 13, revised. New York, Hafner Publishing Co., Inc., 1967.