

AMERICAN Journal of Epidemiology

Formerly AMERICAN JOURNAL OF HYGIENE

© 1977 by The Johns Hopkins University School of Hygiene and Public Health

VOL. 105

JANUARY, 1977

NO. 1

Reviews and Commentary

JUDGMENT AND CAUSAL INFERENCE: CRITERIA IN EPIDEMIOLOGIC STUDIES

MERVYN SUSSER

The processes of statistical inference, those of causal inference, and those of reaching decisions often overlap, but the principles that govern them are not the same (1). Clinicians, epidemiologists and applied statisticians often experience a tension between the formal requirements of a statistical test of a result on the one hand, and the practical requirements of a judgment about the application of the result on the other. Formal statistical tests are framed to give mathematical answers to structured questions leading to judgments, whereas in any field practitioners must give answers to unstructured questions leading from judgment to decision and implementation. These questions of decision generally hinge around judgments about causality and prediction (2). Once-famous historical controversies will serve to illuminate them. The approach is thus that of the case-history and not of the epidemiologic survey, with all the selectiv-

ity the method implies. Nonetheless, from our position of after-knowledge, we have the advantage of observing major figures exercising their judgments as part of the historical evolution of decisions of which we know the consequences.

An early illustration of a statistician's mathematical answer to a structured question is William Farr's formula, the first for an epidemic curve. Farr, a physician appointed as the Compiler of Statistical Abstracts of the newly-founded General Register Office in London in 1839, analyzed the data for the smallpox epidemic of 1837-1839 to show how the disease regularly rose to a peak and declined (3, 4). He then calculated the equivalent of a normal frequency curve and showed that it fitted the observed frequencies of smallpox deaths by quarter-year (5, 6). Exhibit 1 is taken from Farr's report.

An illustration, some 25 years later, of an epidemiologist making a judgment about a practical question is again provided by William Farr, in a letter to the *London Daily News* of February 17, 1866 about the raging and unfamiliar rinderpest epidemic among cattle (6). As mortality among animals continued on a sharply rising curve, the Right Honourable Robert Lowe, a major figure in the parliamentary

From the Division of Epidemiology, Columbia University School of Public Health, 600 W. 168th St., New York, NY 10032.

Read at the meeting of the Biometrics Society, St. Paul, MN, March 25, 1975.

The author acknowledges the contributions to the development of this paper made by Drs. Agnes Berger, Joseph Fleiss and Zena Stein, as well as by several members of his seminar classes.

EXHIBIT 1

Extracts from William Farr's letter to the Registrar General, Second Annual Report of the Registrar General, 1840 (from Humphreys NA (ed.): Vital Statistics: A Memorial Volume of the Reports and Writings of William Farr. London, Sanitary Institute of Great Britain, 1885, pp. 318-319 (reference 4))

Small-Pox										
	1837			1838				1839		
Periods	1	2	3	4	5	6	7	8	9	10
Seasons	Summer	Autumn	Winter	Spring	Summer	Autumn	Winter	Spring	Summer	Autumn
Deaths	2513	3289	4242	4489	3685	3851	2982	2505	1533	1730

If the latent cause of epidemics cannot be discovered, the mode in which it operates may be investigated. The laws of its action may be determined by observation, as well as the circumstances in which epidemics arise, or by which they may be controlled.

Amidst the apparent irregularities of the epidemic of small-pox, and its eruptions all over the kingdom, it was governed in its progress by certain general laws. The deaths in the early stage of the epidemic were not registered. To avoid circumlocution, it will be convenient to call the ten quarters in which the deaths were registered the ten periods, the first quarter the first period, the second the second period, etc., etc. The mortality increased up to the fourth registered period; the deaths in the first were 2513, in the second 3289, in the third 4242; and it will be perceived at a glance that these numbers increased very nearly at the rate of 30 per cent. For multiply 2513 by 1.30 and it will become 3267; multiply 3267 by 1.30 and it will become 4248. The rate of increase is retarded at the end of the third period, and only rises 6 per cent in the next, where it remains stationary, like a projectile at the summit of the curve which it is destined to describe.

The decline of the epidemic was less rapid than its rise, and the mortality was somewhat greater in the autumns of 1838 and 1839 than in the summers. But by taking the mean of the deaths in the third and fourth period, the mean of the deaths in the fourth and fifth period, etc., etc., a regular series of numbers is produced.

Deaths observed in the decline of the Epidemic						
1	2	3	4	5	6	7
4365	4087	3767	3416	2743	2019	1631
Deaths in regular series						
1	2	3	4	5	6	7
4364	4147	3767	3272	2716	2156	1635

The 4365 may be considered to represent the deaths that happened between the middle of February and the middle of May. The regular series of numbers has been calculated upon the hypothesis that the fall of the mortality took place at a uniformly accelerated rate.

The calculated numbers are sometimes a little too high, and sometimes too low; but, on the whole, the agreement is remarkable. The second number (4147) is nearly 5 per cent lower than the first; and the decrease is successively 5, 10, 15, 20, 26, and 32 per cent. The rates of decrease are 1.052, 1.101, 1.152, 1.205, 1.260, 1.318. The division of 4364 by 1.052 reduces it to 4147; the division of 4147 by 1.101 produces 3767, etc. The mortality decreased at accelerated rates; and the rate of acceleration was 1.046, which, by successive multiplication, will reproduce all the rates, 1.052, 1.101, etc., etc. The rate 1.046 may be called the constant.

opposition, announced impending national doom. Mr. Lowe's argument was that "... there is no reason why the same terrible law of increase which has prevailed hitherto should not prevail henceforth." Farr, even though he was a longtime civil servant, did not hesitate to confute Mr. Lowe in the press. "No one," he wrote, "can express a proposition more clearly than Mr. Lowe but the clearness of a proposition is no evidence of its truth." Previous experience of epidemics, Farr argued, showed that the alarm was false, and that a decline from the peak would soon follow. He made his prediction of a downturn merely on the basis of a series of four 4-week averages of mortality. Nothing but judgment founded on his 30-year experience could have allowed Farr to venture so much on so little, and his prediction, although a little optimistic, was soon confirmed (see Exhibit 2).

In another such controversy that I will use as an example, the protagonists, Karl Pearson and Almroth Wright, were more evenly matched. In 1896 Wright had developed a vaccine against typhoid (more or less at the same time as Pfeiffer) (7). Wright's vaccine had been tried out among volunteers in the British Army first in India and later in the South African war (8). In 1904 a specially appointed Committee of the Medical Advisory Board to the War Office, including Wright, recommended on the basis of the data collected by these means that the Army adopt routine inoculation against typhoid.

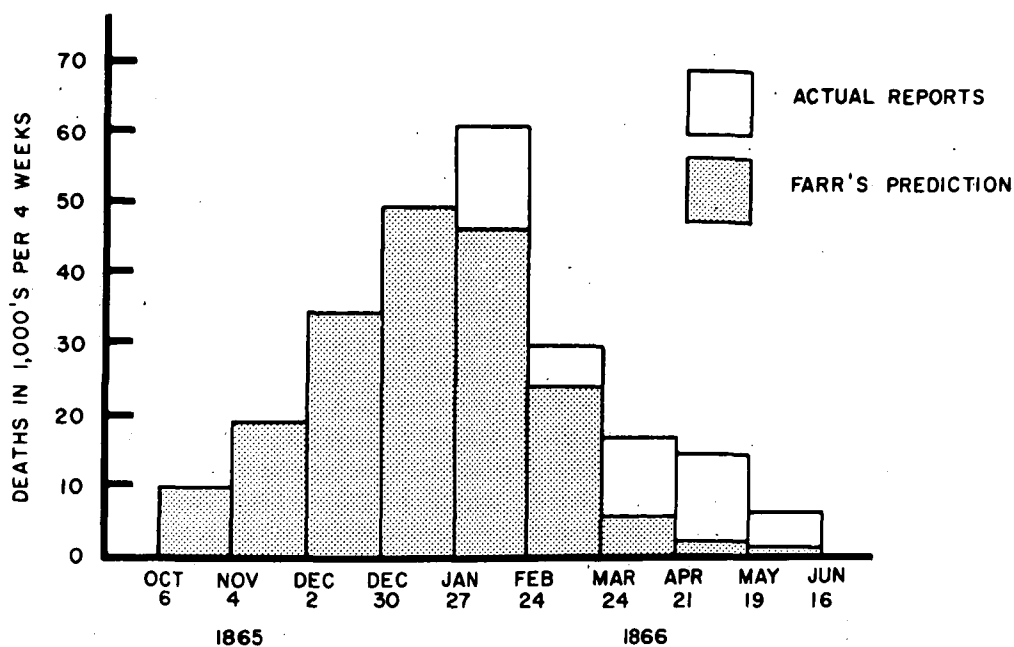
The Medical Advisory Board referred the Committee's report to Professor Pearson (who, Wright said, was a Rhadamanthys to whom all inoculators must come to have judgment passed). Pearson's review of the data was published in the *British Medical Journal* of November 5, 1904 (9). He separately analyzed the data for incidence (these came from five sources) and for case-fatality (these came from six). In each case, he calculated a (tetrachoric) coefficient of correlation.

Pearson appreciated the insufficiency of statistical significance alone for making decisions. In advance of his analysis, he ingeniously derived a judgmental criterion from the available experience with smallpox and diphtheria inoculation, both of which were by then in use. Pearson decided that with a correlation coefficient of 0.6 or more, as with smallpox, the new procedure would be clearly justified; he noted that even with a coefficient between 0.24 and 0.47, as with diphtheria antitoxin treatment, the procedure had been "universally" adopted by the medical profession, and the typhoid vaccine could be recommended above this level.

The results were as follows (Exhibit 3): In all but two of the eleven instances (IV and VII on the table), "the correlations were at least twice, and generally four, five, or more times their probable errors", and hence statistically significant by Pearson's own criterion. But the mean value of the correlations fell below the acceptable preset level ("0.2 or even 0.3"). Unlike smallpox vaccination, wrote Pearson, the effect of typhoid inoculation was inconsistent: in his words, "largely influenced by differences of environment or of treatment," and he did not think the results yielded a pattern sufficiently coherent to be trustworthy.* He recommended further investigation and prophylactic trials, and the suspension of inoculation as a routine measure. He wrote further that "if it were not presumption . . . I should say that the data indicate a more effective serum or effective method of administration must be found before inoculation ought to become a routine practice . . ."

Wright responded to Pearson's report, and an accompanying editorial in the *British Medical Journal*, with a tirade (10,

* Both Pearson and Wright were aware of the problems of using volunteers, in terms of self-selection and comparability with controls. Pearson thought that the caution of volunteers for vaccination may have accounted for the variation in the South African results between immobile and mobile troops.



<u>Periods of Four Weeks ending</u>	<u>Reported Attacks</u>	<u>Calculated Series by "Law"</u>	<u>Actual Figures</u>
-------------------------------------	-------------------------	-----------------------------------	-----------------------

1865

November 4 9,597

December 2 18,817

December 30 33,835

1866

January 27 47,191

February 24 43,182 57,004

March 24 21,927 27,958

April 21 5,228 15,856

May 19 494 14,734

June 16 16 5,000 (about)

EXHIBIT 2. Epizootic cattle plague (rinderpest) epidemic, London, October 1865 to June 1866. (Data taken from Langmuir A: The role of William Farr in the development of the concept of surveillance (unpublished paper).)

got Wright a knighthood to shore his reputation as a scientist† and the Army proceeded with use of the vaccine (7).

Pearson's doubts were not ignored entirely. In a program directed by Wright's former co-worker, Leishman, vaccine use was monitored over a three-year period by trained medical officers assigned to every unit going abroad during this time (20, 21). Exhibit 4 shows the results of this experience published by Leishman. They were eloquent, as Leishman wrote, but less conclusive than he thought (22). The inoculated were volunteers; the diagnostic criteria were not definitive and the medical officers who made the diagnoses in affected cases were not necessarily blind to inoculation status; and the composition of the exposed population was not stable because men were posted to and from units. Further, any attempt to measure duration of exposure was confounded because some men volunteered for vaccination after posting abroad, sometimes in the face of a threatened epidemic or after it, and these records were not collected (22). The Army Report (21) assigned men to inoculated and uninoculated groups according to their status at the end of the observation period, thus inflating the duration of exposure for the inoculated (some of whom were uninoculated for part of the time) and artefactually lessening their apparent risk of contracting the disease over a given period.

Greenwood and Yule (23) corrected for this error by reversing this procedure, a stringent test; they assigned men to inoculated and uninoculated groups according to their status at the beginning of the ob-

servation period, thus deflating duration of exposure for the inoculated and artefactually raising their apparent risk. This calculation halved the relative risk of the uninoculated compared with the inoculated from six- to three-fold, still a significant effect. By the time of World War I, these further data had much strengthened the case for the vaccine, and inoculation was in regular use in the British and French armies. The vaccine probably saved many lives.

As Cockburn pointed out on the countervailing side, however, there is no way of quantitatively assigning the reason for the reduced incidence among vaccination and the concurrent changes in clean water supplies and personal and food hygiene (22). One should note how large the possible costs of wrongly introducing such a vaccine could have been. A half century passed before anyone had the courage again to conduct a controlled prophylactic trial of typhoid vaccine. In the 1950's such a trial became possible because of the discovery of an antibiotic (chloromycetin) effective against typhoid fever; happily the vaccine proved efficacious (24).

The angels were on Wright's side. We might with profit ask ourselves why so great a statistician as Pearson did not perceive them hovering about the head of the man come to judgment. He rendered a sensible initial opinion. The pugnacity of the polemics and a self-serving element in some of Wright's arguments forced Pearson into an increasingly negative stance as the debate proceeded. The question of judgment must be examined in terms of reason rather than affect, however, and it seems to me that on a number of points Wright sustained his argument better than did Pearson.

Pearson used two quantifiable criteria. The results satisfied his first strictly statistical criterion; they were well beyond the usual *bounds for chance events*. They did not satisfy his second criterion: the *strength of the association* did not meet the

† Colebrook (7, p.40) states: "John Freeman has told how it came about — on going home from work in the early hours of the morning, they found a letter from Lord Haldane. . . . He said there was no agreement among the Army Medical authorities about the inoculation policy, but it had got to be adopted. To achieve that he must build up Wright as a great man, and the first step in that process was to make him a knight." In this role Wright can be recognized as Sir Colenso Ridgeon, newly knighted in the opening scenes of George Bernard Shaw's play, *The Doctor's Dilemma*.

EXHIBIT 4*
Recent Results of Antityphoid Inoculation: Statistical table showing the results of antityphoid inoculation in sixteen units of the British Army, up to June 1, 1908

Unit	Medical Officer	Station	Date of arrival	Total strength (actual)	Inoculated		Non-inoculated		
					No.	Deaths	No.	Deaths	
2nd Roy. Fus	Capt. A. B. Smallman	Trimulgherry	Jan., 1905	1,013	10	1	815	59	
17th Lancers	Capt. E. J. Luxmore	Merrut	Oct., 1905	616	3	0	294	71	
Brigade, R.A.	Capt. E. G. Lithgow	Pindi (from Transvaal)	Nov., 1905	370	0	0	310	7	
14th Hussars	Lieut. C. E. Fawcett	Bangalore	Oct., 1906	647	2	0	261	4	
2nd Dorsets	Lieut. E. G. Anthonisz	Wellington	Nov., 1906	1,107	1	0	908	6	
3rd Coldstream Guards	Lieut. J. H. Graham	Cairo	Oct., 1906	705	1	0	136	13	
2nd Leicesters	Lieut. H. S. Sherren	Belgaum	Oct., 1906	963	3	1	617	17	
1st Connaught Rangers	Lieut. A. D. O'Carroll	Dagshai (from Malta)	Mar., 1907	483	0	0	183	2	
3rd Worcesters	Lieut. W. H. Forsyth	Wynberg	Dec., 1907	900	0	0	680	3	
1st Dragoon Guards	Lieut. G. H. Stevenson	Umballa	Dec., 1907	592	0	0	142	0	
1st Yorks	Lieut. S. deC. O'Grady	Cairo	Jan., 1908	893	0	0	423	0	
1st Suffolks	Lieut. J. B. G. Mulligan	Malta	Dec., 1907	900	0	0	500	0	
3rd Roy. Rifles	Lieut. R. W. D. Leslie	Crete	Feb., 1908	879	0	0	689	0	
2nd Bedfords	Lieut. C. M. Drew	Gibraltar	Sept., 1907	700	0	0	380	3	
Brigade, R.A.	Lieut. A. S. Littlejohns	Pretoria	Nov., 1907	375	1	0	128	2	
1st Lan. Fus.	Lieut. F. D. G. Howell	Chakrata	Dec., 1907	940	0	0	144	0	
Totals				12,083	21	2	6,610	187	
				Case-incidence per 1,000.					
				Inoculated		Non-inoculated			
				3.8		28.3			
				6.6		39.5			
				3.7		32.8			

1) Among the whole of the above sixteen units
2) Among the "exposed" units, i.e., in which cases of enteric fever had occurred
3) "Exposed" units, less Royal Fusiliers (the unit inoculated with the "old vaccine")

* Taken from Leishman, W. B: Statistical table of the recent results of antityphoid inoculation. J. R. Army Med Corps 12:163-167, 1909 (reference 20).

preset standard he had estimated from the only two pre-existing vaccines. While strength of association is a quantitative criterion, its application requires judgment.

Much of the passion Wright and his medical supporters felt was indirectly inspired by this criterion of strength of association. The criterion, in the form of a correlation coefficient, ignores attributable risk; it does not bring into the judgment, as we have since learned to do, the scale of benefits and costs for the population. In consequence, Pearson could be, and was, accused of callousness to the life-saving and preventive potential possessed even by this relatively inefficient vaccine.

Wright showed superior intuition with regard to a third quantifiable criterion, the ability of the trial of the vaccine to detect an effect (which I shall loosely describe as *power*), and the effect of random measurement error upon it. Wright recognized, as Pearson seemed not to, that measurement error must have suppressed the strength of the positive association that was found. As a practical consequence, had Pearson recognized the effect of the greater chance of misclassifying typhoid than of misclassifying smallpox or diphtheria, he might have been prompted to relax the level of his criterion of strength of association.

Readers will not need reminding of the importance of the power of a test (25, 26). They may need reminding, however, of the bias toward skepticism in conventional procedures of inference. Analytic strategies for avoiding false negatives are few, aside from the statistical criterion of power, and aside from the analytical elaboration of data that tests internal validity and reveals conditional or suppressed or distorted associations (2). Much statistical strategy aims to avoid false positives, inferences that give credence to causality where none exists.

Human minds seem to be more credulous than skeptical, and most people need

protection against being gulled. Yet undue skepticism can be as dangerous to scientific progress as credulity. Statisticians and epidemiologists are properly professional skeptics. But we must be aware of the trained incapacity to believe in positive results. As the White Queen implied to Alice, one may have to practice believing. Why, one might ask, is the conventional level above which we reject the null hypothesis and accept a positive result set at one chance in 20, and the conventional level above which we accept the null hypothesis and reject a positive result set at one chance in five? Any disparity should rest on a weighing of the gravity of the error of accepting or rejecting an hypothesis, but the costs of rejecting a positive result seem rarely to be considered (26, 27).

Wright's intuition was superior also with regard to an unquantified criterion of judgment, namely external validation of the data by *consistency on replication*. The epidemiologist does not have the opportunities for exact replication which enable the physical scientist to demonstrate consistency. His most powerful alternative to exact replication is consistency of a finding on repeated tests (28, 29). The strength of the argument rests on the fact that diverse approaches produce similar results. There were, by Pearson's grouping, 11 independent replications of the test of the vaccine, and nine of these 11 yielded statistically significant associations. Pearson combined all the tests, and based his judgment only on the single quantitative measure of the unweighted mean value of the 11 correlation coefficients.

The force of external validation by consistency on replication was emphasized by Selvin (30). He explored the principles involved in Durkheim's studies of suicide and religious affiliation, and applied a crude probability test to show the much higher degree of significance obtained than could be attributed to any one study. Long ago, J. S. Mill stated this criterion as

the "method of agreement; all the situations are different but they have one circumstance in common." Here lies the particular cogency of the 36 studies reviewed by the Surgeon General's Committee on *Smoking and Health* (31). This is not "the mere repetition of evidence of the same kind," as R. A. Fisher described these studies (32).

Another non-quantitative criterion of judgment necessarily applied by scientists to the interpretation of data is their *coherence*, in the sense of the reasonableness of the association in biologic terms (28). Here Wright was fortunate in his intimate knowledge of his experimental results on the bactericidal properties of the vaccine *in vitro*, as well as of the field conditions of the trial (for instance, the variability of batches of vaccine, the weaknesses of data collection and so forth). On the other side, retrospectively, Pearson seems rash in his readiness to judge the data for their coherence when he had not made a close examination of their collection and processing. He seems to have exercised what amounted to a self-denying ordinance. When Wright charged that Pearson had refused an invitation to discuss the results with Wright and his fellow committee members some time before he undertook the inspection and analysis of the results, Pearson defended his position with the claim that a statistician needed time for reflection and calculations. Noting the variability in the efficacy of the vaccine, he wrote: "It is difficult to explain this on the basis of any real theory of inoculation" (9, p. 1244). In other words, incoherence detracted from Wright's results. This was perhaps a point he thought better of, because he qualified it in a footnote and suggested some alternative explanations.

Coherence is an ultimate and yet not a necessary criterion for causality. It is, of course, essential to the overall constructs within which scientific investigation proceeds. Without it, the Baconian injunction to come to a conclusion "by proper rejec-

tions and exclusions" must remain meaningless. Only then can one conclude that an explication is coherent with other known facts. But coherence supports pre-existing inference and theory. Incoherence may have a parochial or incidental explanation, as, for example, variability in the manufacture and strength of batches of vaccine. Incoherence may also have a more general explanation, in which instance it will generate new theory.* As Lilienfeld has said: "the finding of a biologically implausible association may be the first lead to this extension of knowledge" (29). In the case of the typhoid vaccine, the existing theory proved sound, and the incoherence incidental and attributable, as Wright thought, to incidental conditions.

The next example reinforces the usefulness of some of the criteria applied to the first. In April 1955, amid great fanfare, the Poliomyelitis Vaccine Evaluation Center sponsored by the National Foundation for Infantile Paralysis published its "Summary Report on the Evaluation of the 1954 Poliomyelitis Vaccine Trials" (33). In December 1955, K. A. Brownlee at the University of Chicago published an invited critique of the trial (34).

The number of schoolchildren involved in the trial was 1,829,916, and the trial cost in the region of five million dollars. Two study plans were put into effect (Exhibit 5a).

An initial plan was described as the "Observed Control" trial. Second grade children whose parents agreed they could participate were to be vaccinated. The controls were to comprise the "corresponding" first and third graders. In the event, the

* Sartwell (28) illuminates the point with a quotation on "nonsense correlations" from D. W. Cheever's book, *The Value and Fallacy of Statistics in the Observation of Disease* (1861): "It could be no more ridiculous for the stranger who passed the night in the steerage of an emigrant ship to ascribe the typhus which he there contracted, to the vermin with which bodies of the sick might be infested. An adequate cause, one reasonable in itself, must correct the coincidence of simple experience."

"corresponding" first and third grade volunteers needed for controls could not be adequately identified from the registers, and all first and third graders were used as controls. This study covered 127 areas of 33 states with a total population of 1,080,680 children in the first, second and third grades (or 59 per cent of the total in both studies).

A second plan was described as the "Placebo Control" trial. R. Korns and M. Levin in New York State, and L. Schuman in Illinois had refused to participate unless there was a placebo control. Thomas Francis then instituted the randomized controlled trial (35).† Children of the first, second and third grades in the remaining study areas were invited to participate; one half of volunteers received the vaccine, the other half a placebo injection. This study covered 84 areas of 11 states with a population of 748,236 children in the first, second and third grades (41 per cent of the total of both studies).

To take first the "Placebo Control" trial (see Exhibit 5b), these results showed the vaccine to be 72 per cent effective ($100 \times (1 - R_1/R_2)$) against paralytic poliomyelitis (row *d*, third column). This result, Brownlee thought, left little doubt as to the effectiveness of the vaccine, given that there was independent random sampling and that the vaccinated and controls differed only in vaccination status. He was not satisfied, however, that this assumption held. On closer examination of the data, Brownlee noted a bias (albeit a slight one) towards greater susceptibility to poliomyelitis among the controls, as judged by pre-inoculation antibody levels.

Turning to the "Observed Control" trial, Brownlee dismissed the data out of hand

† Morton Levin and Abraham Lilienfeld provided this version of events. Paul (35) incorrectly attributes the controlled trial to the advice Austin Bradford Hill gave Francis who, he says, was in London when invited to direct the trial. In fact, Lilienfeld says he himself was with Francis at a conference in Atlanta when Francis received a telephone call asking him to direct the trial.

with such descriptions as "worthless" and "total folly"; the 59 per cent of the total effort devoted to this part of the trial was "stupid" and "futile". In the light of the results, and in the face of the panic and publicity about the poliomyelitis scourge at the time, it surely took courage for Brownlee to write that he felt "... the need for an independent confirmation."

Nonetheless, the vaccine was put into use without further trial. It proved effective (although not as effective as the oral vaccine which later replaced it), and it undoubtedly saved thousands from crippling and death. In retrospect, then, events proved Brownlee's judgment to be too skeptical, even intemperate. No doubt, he was provoked by the heat of debate and publicity, for the National Foundation for Infantile Paralysis had brought large-scale publicity techniques to bear on a scientific issue for the first time, and the pressure to produce had led to risky decisions, even wrong decisions, with some unhappy consequences for the vaccinated (35-37). It was a time, too, when for many the randomized controlled trial had still to establish itself as the preferred means of testing a mode of intervention.

Brownlee, doing battle on behalf of the randomized controlled trial, seems to have blinkered himself from all other evidence. He pointed to two results that demonstrated the "futility of the work in the observed areas." First, non-participants (mainly refusals) had different rates of school absenteeism and were of lower social class than volunteers, and second, they had lower rates of poliomyelitis (as determined from the rate among those given placebo). Thus, it was obviously improper in the "Observed Control" trial, all will agree, to compare the vaccinated second grade *volunteers* with the *total population* of the first and third grades, as was done in the summary report (Row *i* in Exhibit 5b). This comparison runs the risk of confounding by social class and other factors related to poliomyelitis incidence,

EXHIBIT 5a

Study plans for poliomyelitis vaccine field trial, 1954*

"Placebo Control" Trial		"Observed Control" Trial
a. Half 1st, 2nd, 3rd grade volunteers; randomly assigned. <i>N</i> = 200,745.	<i>Vaccinated</i>	e. 2nd grade volunteers. <i>N</i> = 221,998.
b. Half 1st, 2nd, 3rd grade volunteers, randomly assigned. <i>N</i> = 201,229.	<i>Controls</i>	f. All 1st and 3rd grade children. <i>N</i> = 725,173.
c. 1st, 2nd, 3rd grade refusals. <i>N</i> = 338,778.	<i>Non-participants</i> (Refusals, absentees, etc.)	g. 2nd grade. <i>N</i> = 123,605.

* Adapted from Vaccine Evaluation Center: Evaluation of the Field Trial of Poliomyelitis Vaccine. Summary Report. Ann Arbor, MI, University of Michigan, 1955 (reference 33).

EXHIBIT 5b

Comparison of poliomyelitis among vaccinated and controls under two study plans: Rates per 100,000*

	All cases	All polio	Paralytic	Non-para-lytic	Non-po-lio
<i>"Placebo Control" Trial</i>					
a. Vaccinated: 1st, 2nd, 3rd grade	41	28	16	12	12
b. Placebo: 1st, 2nd, 3rd grades	81	71	57	13	10
c. Non-participants†: 1st, 2nd, 3rd grades	54	46	36	11	7
d. Ratio of Vaccinated to Placebo	.506	.394	.280	.923	.120
<i>"Observed Control" Trial</i>					
e. Vaccinated: 2nd grade	34	25	17	8	9
f. Unvaccinated: 1st and 3rd grades	61	54	46	8	6
g. Non-participants†: 2nd grade	53	44	35	9	10
h. Total: 2nd grade	41	32	23	8	9
i. Ratio of vaccinated 2nd grade to all 1st and 3rd grades	.557	.462	.369	1.0	1
j. Ratio of all 2nd grade to all 1st and 3rd grades	.672	.592	.50	1.0	1.5

* Adapted from Vaccine Evaluation Center: Evaluation of the Field Trial of Poliomyelitis Vaccine. Summary Report. Ann Arbor, MI, University of Michigan, 1955 (reference 33).

† Mainly refusals.

in addition to the risk created by the imperfect age match across school grade.

Yet, by the criterion of consistency on replication, and by the criterion of coherence, the "Observed Control" trial provided valuable support, unrecognized by Brownlee, for the "Placebo Control" trial. Thus, we are back with the question of judgment. First, with regard to external validation by consistency on replication, note that the incidence of poliomyelitis in several categories of increasing diagnostic refinement is virtually identical in the

"Placebo Control" and "Observed Control" areas both among the vaccinated and among the non-participants.

Further, in the "Observed Control" trial a legitimate comparison can be made between the *total* second grade population, vaccinated and unvaccinated, and the *total* first and third grade populations (Exhibit 5b, row *j*). The comparison can be taken as a stringent test of the total program in the field, since the closer observation of vaccinated children, solely in the second grade, was likely to bring more poliomyelitis

cases to light in the second grade than in the first and third grade controls.* The vaccine remains effective, although less so (50 per cent), by this test ($\chi^2 = 29.14$; d.f. = 1; $p < .001$).

This comparison, in the double-barrelled design of "placebo" and "observed" controls, would also have proved a safeguard against a contingency that did not, but might have occurred. It is not inconceivable that in the "Placebo Control" trial a fully effective vaccine could have produced a level of herd immunity in the population that would have protected the placebo controls to a degree that significant differences between them and the vaccinated did not clearly emerge. Contamination of the control group was a lesser likelihood in the "Observed Control" design, since first and third grade controls were less likely to commingle with the vaccinated second graders, whereas in the "Placebo Control" design the vaccinated and the controls were in the same classroom.

With regard to the criterion of coherence, the most refined diagnosis, and hence the most precise results, were to be expected among paralytic cases, and in both studies it is these cases which contribute the great part of the positive result. Further, as noted above, lower socioeconomic classes had a known lower incidence of poliomyelitis and a demonstrated lower level of participation in the study. For this reason, the dilution of the result seen in the "Observed Control" trial was coherent and to be anticipated, since the control incidence rates were lowered by the inclusion of non-participants among them (not by intention, it seems, but because of flawed execution of the registration of acceptance and refusal among controls in the first and third grades).

Nonetheless, the existence of the "Observed Control" provided a coherent rebut-

tal against one critical interpretation of the results of the "Placebo Control" design. The higher incidence of poliomyelitis among the placebo controls than among the vaccinated, it was argued by the critics, could be the result, not of the protection of the vaccinated, but of provocation by inoculation, just as previously poliomyelitis had been found to be provoked by other inoculations. Francis (38) was able to refute the suggestion from the data (unpublished) on volunteers who received no injections in the "Observed Control" areas. The estimated risk in the "Observed Control" areas for volunteers who received no injections (as compared with those who refused participation) was the same as the risk in the "Placebo Control" areas of those who received placebo injections (as compared with those who refused participation).

Brownlee took the position that the difference between the evidence from a blind randomized prophylactic trial and any other is absolute. He cited the Summary Report: "In observed areas where only those second grade children whose parents requested participation were vaccinated, the problem of establishing the control population was more complex", and he commented: "It is perfectly true to say that it is more complex, but to indulge in understatement of this order of magnitude is to be misleading. The plain fact is that it is impossible." Brownlee was misled by Fisher's central dogma of the randomized trial. Differences in strength of inference from experimental and observational studies are relative, not absolute, as anyone who has conducted large experimental field trials will know, and to test their validity requires that we bring to bear all possible criteria of judgment to all the data.

The next example bears on the controversy about the effects of smoking. Judgment failed two major statisticians, one of whom many class among the greatest of all. Joseph Berkson and R. A. Fisher both

* Although the Summary Report (33) makes little of this comparison, Alexander Langmuir tells me that he and William Cochran insisted that the essential data be included in the report.

disputed the causal associations of smoking with lung cancer. Berkson (39, 40) advanced a multitude of arguments, some good and most less good—like Pearson, at one point he proposed the mere unreliability of data (smoking histories and death certificates) as a reason for the finding of a spurious positive association and, at another point, the apparent incoherence of some of the results—but he devoted his major attack to the judgmental criterion of specificity. Here Berkson took his stand on ground prepared for him by Yerushalmy and Palmer (41) who argued that “specificity of effect” was one essential of Koch’s postulates. In chronic disease too, he held, specificity should be sought. A characteristic is specific to a disease “when . . . similar relationships do not exist with a variety of characteristics and with many disease entities when such relationships are not predictable on physiologic, pathologic, experimental or epidemiologic grounds. In general, the lower the frequency of these other associations, the higher is the specificity of the observed association and the higher the validity of the causal inference” (41, pp. 38–39).

By the term *specificity of association*, then, we describe the precision with which the occurrence of one variable will predict the occurrence of another. The ideal, a one-to-one relationship, encompasses the element of strength of association as well as of precision, and might be better reduced to the statement that one thing, and only one thing, causes one effect and only one effect. Departures from the ideal occur first where many things are posed as causes of a single effect, and second where a single thing is posed as the cause of many effects. In the smoking controversy, Berkson was perturbed by the second type of departure from the ideal in the smoking data, i.e., where one thing has many effects. He noted the congeries of manifestations found in association with smoking, and he argued that such non-specificity casts doubt on any causal relationship. He

carried the attack forward by analogy. He pointed to the extraordinarily regular association of marital status—in the order of married, single, divorced—with overall mortality from many specific diseases. He offered, as more likely than some of the causal associations considered by others, such explanations as systematic bias in registered data. The equally non-specific associations of smoking with disease, Berkson held, led to equally infirm inferences of causality.

Arguments that demand specificity are fallacious, if not absurd. There can be no logical reason why any identifiable factor, and especially an unrefined one, should not have multiple effects (2, 28, 29, 42). To take Berkson’s own analogy, marital status, in the form of transition into widowhood, has since been shown to be a highly probable cause of suicides, of entry into psychiatric care, and of cirrhosis of the liver (43–48). Other associations will not prove causal, but there is surely more to come. By now it is evident that the associations of health disorders with smoking depend on a variety of mechanisms, some causal and some not. Specificity enhances the plausibility of causal inference, but lack of specificity does not negate it.

Fisher’s major attack on the smoking and lung cancer hypothesis was around the judgmental criterion of the *time-order and logical structure* of causal and outcome variables (although he too attacked on grounds of incoherence) (32). Predisposition poses a major problem in inferring the time-order of variables. As a latent manifestation of an outcome variable, predisposition must cast doubt on the precedence in time of supposedly independent causal variables. Fisher suggested that smokers could be self-selected from among persons with a genetic predisposition both to smoking and to lung cancer, and pointed for supporting evidence to the association between genetic constitution and smoking habits.

These associations between personality

type and smoking habit pose the same difficulties for inferring time-order among variables as all others measured at the same point in time. While predisposition cannot be excluded, its presence has certainly not been demonstrated. Vital links in the logical structure of this counter-hypothesis are missing. The relevant personality traits have even now not been linked with lung cancer, let alone been shown to antedate the disease. If this link with the disease were to be supplied, the association of the rising trend in the frequency of lung cancer with smoking and its distribution among the sexes, and, most of all, the overwhelming strength of the association (49, 50), would require at the least that smoking be a crucial intervening variable between the predisposition to lung cancer and its manifestations. (There are several other criteria of judgment by which this argument can be pursued, as has been ably done by authors beginning in the early days of the controversy (31, 49)).

Fisher chose a criterion for attack on which many inferences of causality are vulnerable; in observational studies the criterion of time-order is perhaps the most difficult to defend. In order finally to determine time-order and logical structure among variables, our main reliance must be on design: we use incidence rather than prevalence measures, cohort rather than case-control studies. Here, cohort studies and the effects of stopping smoking on mortality have settled the question for most of us (51). Recent attacks on the smoking hypothesis add nothing essentially new to the arguments of Fisher and Berkson.

In conclusion, this paper should not be construed as an attack on statistical inference or on major statisticians, but as a demonstration of the fallibility of judgment even among those with superb technical equipment, and of the need for developing the means of decision-making. Some injunctions may perhaps be drawn from

the review of these few case histories. One should be aware of bias toward negative judgments, which require as much caution as positive judgments, recognize that the difference between types of evidence is relative and not absolute, and apply all available criteria of judgment to any particular instance. Perhaps we may advance beyond our present limitations by systematizing our criteria of judgment, and by expanding the number and type of available criteria.

REFERENCES

1. Raiffa H: *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. New York, Addison Wesley, 1970
2. Susser M: *Causal Thinking in the Health Sciences: Concepts and Strategies of Epidemiology*. New York, Oxford University Press, 1973
3. Susser M, Adelstein AM (eds.): *Introduction to Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr*. Edited by NA Humphreys. London, Sanitary Institute of Great Britain, 1885. Reprinted for the New York Academy of Medicine. Metuchen, NJ, Scarecrow Press, Inc., 1975
4. Humphreys NA (ed.): *Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr*. London, Sanitary Institute of Great Britain, 1885
5. Serfling RE: Historical review of epidemic theory. *Hum Biol* 24:145-166, 1952
6. Brownlee J: Historical note on Farr's theory of the epidemic. *Br Med J* 2:250-252, 1915
7. Colebrook L: *Almroth Wright: Provocative Doctor and Thinker*. London, Heinemann, 1954
8. Wright AE: *A Short Treatise on Anti-typhoid Inoculation*. Westminster, Archibald Constable & Co, London, 1904
9. Pearson K: Report on certain enteric fever inoculation statistics. *Br Med J* 2:1243-1246, 1904
10. Wright AE: Antityphoid inoculation. *Br Med J* 2:1233, 1904
11. Editorial. *Br Med J* 2:1259-1261, 1904
12. Wright AE: Antityphoid inoculation. *Br Med J* 2:1343-1345, 1904
13. Pearson K: Antityphoid inoculation. *Br Med J* 2:1432, 1904
14. Wright AE: Antityphoid inoculation. *Br Med J* 2:1489-1491, 1904
15. Pearson K: Antityphoid inoculation. *Br Med J* 2:1542, 1904
16. Wright AE: Antityphoid inoculation. *Br Med J* 2:1614, 1904
17. Pearson K: Antityphoid inoculation. *Br Med J* 2:1667-1668, 1904
18. Wright AE: Antityphoid inoculation. *Br Med J* 2:1727, 1904
19. Pearson K: Antityphoid inoculation. *Br Med J* 2:1775-1776, 1904
20. Leishman WB: Statistical table of the recent

- results of antityphoid inoculation. *J R Army Med Corps* 12:163-167, 1909
21. Report of Antityphoid Committee, G. B. (1912). London, Her Majesty's Stationery Office, 1913
 22. Cockburn WC: *The early history of typhoid vaccination*. *J R Army Med Corps* 101:171-185, 1955
 23. Greenwood M, Yule GU: Statistics of antityphoid and anticholera inoculation and interpretation of such statistics in general epidemiology. *Proc R Soc Med* 8:113-194, 1915
 24. Cvjetanovic BB: Field trials of typhoid vaccines. *Am J Public Health* 47:578-581, 1957
 25. Neyman J, Pearson ES: On the use and interpretation of certain test criteria of statistical inference. *Biometrika* 20:175-240, 264-299, 1928
 26. Neyman J, Pearson ES: On the problem of the most efficient of statistical hypotheses. *Philos Trans R Soc Lond* 231A:289-337, 1933
 27. Lehmann EL: Some principles of the theory of testing hypotheses. *Ann Math Stat* 21:1-26, 1950
 28. Sartwell PE: "On the methodology of investigations of etiologic factors in chronic diseases": Further comments. *J Chron Dis* 11:61-63, 1960
 29. Lilienfeld AM: "On the methodology of investigations of etiologic factors in chronic diseases": Some comments. *J Chronic Dis* 10:41-46, 1959
 30. Selvin HC: Durkheim's suicide and problems of empirical research. *Am J Sociol* 63:607-619, 1958
 31. United States Dept. of Health, Education and Welfare: *Smoking and Health: Report of the Advisory Committee to the Surgeon General*. Washington DC, USGPO, 1964
 32. Fisher RA: *Smoking and the Cancer Controversy*. Edinburgh, Oliver and Boyd, 1959, p 11
 33. Vaccine Evaluation Center: *Evaluation of the Field Trial of Poliomyelitis Vaccine*. Summary Report. Ann Arbor, MI, University of Michigan, 1955
 34. Brownlee KA: Statistics of the 1954 polio vaccine trials. *J Am Stat Ass* 50:1005-1013, 1955
 35. Paul JR: *A History of Poliomyelitis*. New Haven, CN, Yale University Press, 1971
 36. Meier P: Safety testing of poliomyelitis vaccine. *Science* 125:1067-1071, 1957
 37. Langmuir A: The surveillance of communicable diseases of national importance. *N Engl J Med* 268:182-192, 1963
 38. Francis T: Symposium on controlled vaccine field trials: Poliomyelitis. *Am J Public Health* 47:283-287, 1957
 39. Berkson J: Smoking and cancer of the lung. *Proc Staff Meeting of the Mayo Clinic* 35:367-385, 1960
 40. Berkson J: Mortality and marital status. *Am J Public Health* 52:1318-1329, 1962
 41. Yerushalmy J, Palmer CE: On the methodology of investigations of etiologic factors in chronic diseases. *J Chronic Dis* 10:27-40, 1959
 42. MacMahon B, Pugh TF, Ipsen J: *Epidemiologic Methods*. Boston, Little, Brown, 1960
 43. Ciocco A: On the mortality in husbands and wives. *Hum Biol* 12:508-531, 1940
 44. Ciocco A: On the mortality in brother-sister and husband-wife pairings. *Hum Biol* 13:189-202, 1941
 45. Kraus AS, Lilienfeld AM: Some epidemiologic aspects of the high mortality rate in the young widowed group. *J Chronic Dis* 10:207-217, 1959
 46. MacMahon B, Pugh TF: Suicide in the widowed. *Am J Epidemiol* 81:23-31, 1965
 47. Stein ZA, Susser MW: Widowhood and mental illness. *Br J Prev Soc Med* 23:106-110, 1969
 48. McNeil D, Kelsey JL: Cirrhosis mortality among widows. Paper presented at the Seventh International Meeting of the International Epidemiological Association, University of Sussex, England, August 17-21, 1974
 49. Cornfield J, Haenszel W, Hammond EC, et al: Smoking and lung cancer: Recent evidence and a discussion of some questions. *J Natl Cancer Inst* 22:173-203, 1959
 50. Bross IDJ: Pertinency of an extraneous variable. *J Chronic Dis* 20:487-495, 1967
 51. Doll R: The age distribution of cancer: Implications for models of carcinogenesis. *J R Stat Soc* 134:133-166, 1971