SOURCES AND VALIDITY OF MEDICAL STATISTICS WITH SPECIAL EMPHASIS ON DIAGNOSES

Barkev S. Sanders, Public Health Service

While we cannot map out the exact boundaries of medical statistics, all of us will agree that incidence and prevalence of disease represent a sizeable segment of such statistics, however defined.

The following are methods that have been used to ascertain prevalence of disease in a population:

1) A census or survey of the population to be studied;
2) Analysis and interpretation of causes of death related to fatality rates;
3) Hospital statistics by diagnosis;
4) Canvassing of physicians, or record-keeping by them, of the kinds of illnesses they treat;
5) Diseases notifiable by law;
6) Registers of persons with certain diseases or defects; and
7) Examination of representative samples of the population by a trained team of physicians with agreed-upon standards for differential diagnoses.

Each of these approaches has advantages and disadvantages. Since time does not permit a systematic consideration of them all, I shall limit myself to the first and last: population survey and clinical examination.

In the United States prior to 1950, the household survey was the preferred method of obtaining morbidity information, and to most people it still is. In this method an adult member of the household generally is asked to report for a specified time period the illnesses and conditions suffered by herself and other family members. The diagnosis, duration, and severity of illness are also reported, the latter chiefly by length of incapacitation and receipt of medical care, including hospitalization.

Even though morbidity surveys, historically speaking, are comparatively recent, the method of counting certain sick or handicapped persons in the population can be traced at least to Biblical times when periodic census became an instrument of statecraft.

In the United States in the 19th Century, some censuses included a count of sick and incapacitated persons with a few broad categories. The deaf, the mute, and the blind were enumerated by many Federal decennial and some State censuses even as late as 1930.

With the introduction of the categorical assistance program for the blind (Aid to the Blind) in the '30s, statisticians were startled to find many more blind persons eligible for aid than expected from the Census count. (1) While it would seem the Census took this disparity to heart and abandoned the counting of handicapped persons, many statisticians have been unwilling to abandon this method for obtaining prevalence of diseases. Therefore, much of our information on prevalence rates comes from specially designed morbidity surveys.

I. Evidence of Limitation of Surveys and Census

Despite the widespread use of this method, numerous examples indicate it is deficient, at least so far as prevalence 1/ of disease is concerned. To illustrate these deficiencies I will cite briefly instances from different types of studies.

Lack of Validation

Table 1, taken from the Commission on Chronic Illness Study in Baltimore, (3) illustrates the extent of discrepancy in morbidity findings derived from clinical examinations and from a household survey. It shows that, even in terms of broad diagnoses (2 place code), on the average, only 17 per cent of diseases diagnosed by a single clinical examination were reported specifically enough in the household survey to be given the same general code number. The percentage of matching varies widely for different diseases, ranging from 99 per cent for asthma, to 0 for syphilis, rheumatic fever, cervicitis, and arthritis.

[Insert Table 1]

It is probable that this matching would be even lower if it were possible to re-examine the sample, or at least, doubtful cases in the sample.

Another shortcoming of the survey method, as I see it, is the significant proportion of false positives, i.e., reporting of diseases which cannot be confirmed by clinical examination. Some statisticians, interested solely in a prevalence rate, have treated this deficiency as an asset, in that it offsets in part the large proportion of false negatives. But health workers, concerned not only with prevalence rates but with other characteristics of individuals with various diseases, find this a further deficiency.

Table 2, also based on the Baltimore Survey, shows that nearly one-half of all conditions reported in the household survey could not be validated by a single clinical examination. The extent of false positives in the Baltimore Study ranges from zero per cent for tuberculosis, diabetes, psychophysiologic disorders, rheumatic fever, angina pectoris, arteriosclerosis, and rheumatoid-and osteo-arthritis to 91.9 per cent

1/ The author has ready for publication a monograph entitled Evaluation of Morbidity Surveys as a Measure of Disease Prevalence, which presents these deficiencies in much detail based on recent surveys, including the survey of Kit Carson County, Colorado, with which the author was associated. See A Health Study in Kit Carson County, Colorado, Public Health Service Publication No. 844; especially Chapter II.

Table 1 - Percentage Match of Evaluation Diagnoses 1/2/3/ with Diagnostic Information Reported in the Household Survey, Baltimore, 1953-1955

Evaluation Diagnoses	Percent of Evaluation Diagnoses Matched by That Reported by Survey
All diagnoses	17.2
Asthma	99.2
Rheumatoid arthritis	97.8
Hay fever	73.3
Symptoms referable to limbs and back	58.3
Other allergies	57.5
Chronic sinusitis	44.7
Blindness and impaired vision	44.2
Other diseases of central nervous system	42.5
Deafness and impaired hearing	42.3
Osteoarthritis	42.0
Diabetes mellitus	37.2
Hemorrhoids	34.6
Coronary artery disease and angina pectoris	31.0
Diseases of gallbladder	30.9
Varicose veins of lower extremities	27.0
Other forms of arthritis	22.9
Orthopedic impairments (n.e.c.) except cerebral paralysis	21.9
Hernia of abdominal cavity	21.2
Vascular lesions of central nervous system	21.0
Low back strain	19.7
Cerebral paralysis (n.e.c.)	17.1
Malignant neoplasms	15.9
Other symptoms, senility, and ill-defined causes	14.5
Diseases of female genital organs (except cervicitis)	13.8
Diseases of kidney	15.3
Heart Disease	13.5
Rheumatic heart disease and rheumatic fever with heart involvement	13.3
Hypertension without heart involvement	13.2
Anemia	13.0
All other diagnoses	12.7
Benign neoplasms of uterus	12.6
Other heart disease	12.0
Other mental, psychoneurotic and personality disorders	10.9
Psychoneuroses	10.2
Psychoses	9.5
Cataract (not causing blindness)	9.0
Tuberculosis	7.8
Neoplasms 4/	7.8
Diseases of thyroid	7.7
Other diseases of circulatory system	6.5
Hypertensive heart disease	6.1
Obesity 5/	3.7
Benign neoplasms of other sites	3.4
Arteriosclerosis	1.2
Psychophysiologic autonomic and visceral disorders	0.7
Syphilis	--
Migraine	--
Rheumatic fever without heart involvement	--
Cervicitis	--
Arthritis	--

1/ Based on weighted number of evaluation diagnoses.
2/ Excludes conditions which began during the interval between the household survey and the clinical evaluation. Includes could-report and could-not-report conditions.
3/ Reported in any terms which came in the definition of "high degree agreement". For definitions of "high" and "low degrees of agreement", see Appendix C of the source indicated below.
4/ Includes neoplasms of unspecified nature.
5/ For the special definition of obesity used in the evaluation clinic, see Chapter 12 and Appendix B of the source indicated below.

Source: Derived from the work of the Commission on Chronic Illness. Chronic Illness in the United States, Vol. IV -- Chronic Illness in a Large City - The Baltimore Study, The Harvard University Press, Cambridge, Mass., 1957. Table 113, pp. 304-305.

for diseases of the kidney. While repeated examinations would probably somewhat reduce the proportion of false positives, their net effect in increasing false negatives would be greater. Other validation studies like that of Baltimore, citations 2 to 6 inclusive, confirm in general this low level of diagnostic specificity from surveys.

[Insert Table 2]

Table 2 - Percentage of Survey Diagnoses 1/2/3/ which could not be matched with that from Clinical Evaluation. Baltimore 1953-1955.

Survey Diagnoses	Percentage False Positives in the Survey Using Evaluation Diagnoses as Criterion	Survey Diagnoses	Percentage False Positives in the Survey Using Evaluation Diagnoses as Criterion
All diagnoses	45.6	Hypertensive heart disease	38.4
Rheumatoid arthritis	0.0	Cerebral paralysis (n.e.c.)	43.3
Osteoarthritis	0.0	Heart disease	44.8
Tuberculosis	0.0	Diseases of gall-bladder	45.7
Obesity 4/	0.0	Neoplasms	46.2
Psychoses	0.0	Benign neoplasms of other sites	48.6
Rheumatic fever without heart involvement	0.0	Chronic sinusitis	50.4
Coronary artery disease and angina pectoris	0.0	Diseases of female genital organs (except cervicitis)	50.6
Psychophysiologic autonomic and visceral disorders	0.0	All other diagnoses	52.3
Cataract (not causing blindness)	0.0	Other heart disease	59.6
Arteriosclerosis	0.0	Rheumatic heart disease and rheumatic fever with heart involvement	59.8
Low back strain	1.6	Malignant neoplasms	60.5
Hernia of abdominal cavity	3.7	Vascular lesions of central nervous system	61.1
Diseases of thyroid	5.6	Anemia	64.2
Varicose veins of lower extremities	6.0	Symptoms referable to limbs and back	65.6
Benign neoplasms of uterus	6.2	Hypertension without heart involvement	73.4
Diabetes mellitus	6.6	Other symptoms, senility, and ill-defined causes	80.6
Hemorrhoids	16.7	Migraine	81.8
Deafness and impaired hearing	21.6	Other mental, psychoneurotic and personality disorders	82.8
Orthopedic impairments (n.e.c.) except cerebral paralysis	22.3	Other diseases of circulatory system	91.2
Other allergies	26.2	Diseases of kidney	91.6
Arthritis	27.6	Cervicitis	--
Other forms of arthritis	28.5	Syphilis	--
Psychoneuroses	29.5		
Blindness and impaired vision	34.0		
Hay fever	34.3		
Asthma	36.3		
Other diseases of central nervous system	37.3		

1/ Based on weighted number of evaluation diagnoses.

2/ Excludes conditions which began during the interval between the household survey and the clinical evaluation. Includes could-report and could-not-report conditions.

3/ Reported in any terms which came in the definition of "high degree agreement". For definitions of "high" and "low degrees of agreement", see Appendix C of the source indicated below.

4/ For the special definition of obesity used in the evaluation clinic, see Chapter 12 and Appendix B of the source indicated below.

Source: Derived from the work of the Commission on Chronic Illness. Chronic Illness in the United States, Vol. IV -- Chronic Illness in a Large City - The Baltimore Study, The Harvard University Press, Cambridge, Mass., 1957. Table 121, pp. 324-325.

Incompleteness of Survey Reported Conditions

Now let us consider how complete our diagnostic information is when drawn from surveys. During the 40's and 50's a number of studies were made to determine the effectiveness of multiphasic screening and of periodic health examinations (19) in finding undiagnosed conditions in various population groups.

I shall use findings from Elsom, et al (13) as illustrative of this deficiency.

In their study these authors analyze findings from periodic health examinations of 1513 executives of various firms in and around Philadelphia and one in the South. These examinees represented 96 per cent of all executives eligible for the periodic examinations. In the first examination, 906 conditions, previously unknown, were diagnosed, and in 822 executives who returned for a second examination 16 to 28 months later, an additional 389 new conditions were diagnosed. These 906 and 389 conditions were classified by three of the authors according to (A) potential seriousness of the disease or condition, (B) optimal effect of treatment, and (C) urgency of medical attention required. 2/ Final scoring represented the agreed-upon rating by the three authors who first classified each disease independently in terms of A, B, and C. Ratings for the most frequent diagnoses not known previously are shown in Table 3.

[Insert Table 3]

Of the 1513 persons receiving the initial examination, 612 were found to have previously unrecognized diseases. Of the remaining 901 persons, 428 (28 per cent) were diagnosed as healthy while 473 (31 per cent) did not have undiagnosed diseases discovered.

Of the 612 group, 57 per cent of the diagnosed diseases were regarded as serious, i.e., would result in death or major disability if unchecked; 34 per cent, minor; and 9 per cent insignificant. Effective therapeutic measures were deemed available for 93 per cent of the 906 conditions. Immediate treatment was considered urgent for only a small proportion. Of interest, too, is the fact that over half of the newly-diagnosed diseases were found in 13 per cent of those examined.

These findings are typical of other studies, some listed in references 7-19, showing that any population subjected to screening or physical examination yields many individuals with potentially serious diseases unknown to them and to their physicians. In the Baltimore Study, (3) for instance, the over-all morbidity rate as revealed by physical examination was 2.3 times that reported by the survey--counting false positives.

Elsom's study also shows that in this group of executives more suffer from undiagnosed conditions than from known diseases. The prevalence rate following the examination was more than twice what it would have been in terms of previously diagnosed conditions. If this is true for men in executive positions, plainly it would hold true, to even greater degree, for comparable age groups lower on the economic ladder, and in the sparsely populated areas with fewer health facilities and personnel and higher relative costs for health care.

Non-Replicability of Response

Aside from specific studies in the health field, a much larger volume of general information indicates that, by and large, unsolicited questioning of sample populations shows a low level of replicability in general, though this level varies for different types of information, for different groups, etc. Examples of these are found in the references 19 to 30.

A study by the Public Health Service in Nashville associating air pollution with prevalence of certain diseases is an example. In the household interview, in addition to other questions, a list of 32 diseases, mostly chronic, was used. A sub-group of those interviewed were later invited to the University clinic for medical examination.3/ The interval between the household survey and clinical examination was, on the average, one week.4/ Before examining the patient or taking his history, the physician used the same list of 32 diseases to ask the individual if he had ever had any of them. (For children under 15, questions were put to the mother, in both instances).

Table 4 shows that of 714 replies given to lay interviewer and the doctor in which there was an affirmative answer, only in 33 per cent the answer was consistent, i.e. "Yes" "Yes" to both. In 57 per cent the answer to the physician was Yes while the prior answer to the interviewer had been No. In 10 per cent the answer to the doctor's questioning was No whereas it had been Yes to the lay interviewer.5/ Thus in two-thirds of the cases the reply was reversed within one week, despite the fact that we may assume some correlation if only for the sake of self-consistency. And if there were some way of getting at the truth, it is highly probable that discrepancies would be even greater.

[Insert Table 4]

We can scarcely write off such discrepancies as due to "memory failure", "misunderstanding", or "changes in the disease picture". Moreover, for children under 15 for whom the mother was the respondent, in both instances, consistency of response was lowest of all, suggesting a purposeful distortion of information on the part of many. Only 26 per cent gave the same reply to both interviewer and physician.

Another example of non-replicability is from a study of respiratory symptoms among 144 mail carriers in Great Britain, (21) with interviews conducted by three physicians and three trained sick visitors. Two separate interviews were held approximately six weeks apart, based on the assumption that in four weeks consistency attributable to memory would be negligible. To pin down sources of inconsistencies, interviews

2/ For further elucidation of A, B, and C, see footnote to Table 3.

3/ Results of the clinical findings were not analyzed as far as is known.

4/ The maximum interval was 23 days.

5/ In this analysis all the "No" "No" replies and replies where the response to the doctor or to lay interviewer was different from "Yes" or "No" are excluded.

Table 3 - Ten Most Frequent Diagnostic Categories Among 1,513 Executives, with Grading and Frequency of Occurrence

Class and Grade* A B C	Diagnosis	No. of Times Diagnosed† 1st Examination	2nd Examination	Total
3 1 2	Obesity	137	45	182
2 3 2	Hypertension	124	33	157
	Anorectal lesions	94	37	131
4 3 4	Cryptitis	1	0	1
4 3 4	Fissure	4	0	4
4 3 4	Granuloma	1	0	1
3 1 4	Hemorrhoids	76	33	109
4 3 4	Papillitis	3	1	4
4 3 4	Papillae, hypertrophied	8	1	9
4 3 4	Proctitis	1	2	3
	Colonic polyps	87	25	112
1 1 2	Benign	84	25	109
1 1 1	Malignant	3	0	3
	Prostatic lesions	47	20	67
2 1 3	Benign hypertrophy	37	14	51
1 1 1	Carcinoma	1	2	3
1 1 2	Nodule	1	0	1
3 2 3	Prostatitis	8	4	12
2 1 2	Inguinal hernia	35	5	40
1 3 2	Diabetes	28	7	35
3 3 3	Anxiety state, mild	16	9	25
	Arteriosclerotic heart disease	14	4	18
1 3 2	Angina pectoris	4	1	5
1 3 2	Arteriosclerotic heart disease (Unspecified)	6	1	7
1 3 2	Coronary heart disease	2	1	3
1 2 1	Myocardial infarction	2	1	3
2 2 2	Peptic ulcer	11	10	21
	Subtotal	593	195	788
	Total, all Diagnoses	906	389	1,295

* Class A, potential seriousness of disease. Results if not treated: grade 1, deaths; grade 2, major disability; grade 3, minor disability; and grade 4, insignificant disability. Class B, optimal effect of known treatment, likely to result in improvement of: grade 1, eradicated; grade 2, arrested; grade 3, ameliorated; and grade 4, outcome not affected. Class C, urgency of treatment required; grade 1, urgent within days; grade 2, not urgent but promptness still indicated; grade 3, early therapy helpful though not presently required; and grade 4, time not important in terms of inauguaration of therapy or progress of disease on basis of knowledge of disease.

† Figures to the right in columns "1st Examination", "2nd Examination", and "Total" indicate number of times a diagnosis listed under a general heading occurred.

Source: Elsom,K.A. et al: Periodic Health Examination, JAMA, Vol. 172, No. 1, Jan. 1960. p. 56/6

Table 4 - Responses to Doctor and Nonmedical Interviewer on Having Had Specific Diseases, by Response Groups, Nashville, Tenn. 1959*

Response Groups	"Yes" to Both	"Yes" to M.D. "No" to Interviewer	"No" to M.D. "Yes" to Interviewer	Total "Yes-"Yes" "Yes-"No" "No"-"Yes"
(1)	(2)	(3)	(4)	(5)
Number				
Responses of all individuals	239	404	71	714
Women responding for self	126	181	21	328
Men (wife responding)	53	95	20	168
Children under 15 (mother responding)†	35	79	20	134
All others, age 15 and over, not included elsewhere	25	49	10	84
Per cent				
Responses of all individuals	33.5	56.6	9.9	100.0
Women responding for self	38.4	55.2	6.4	100.0
Men (wife responding)	31.5	56.6	11.9	100.0
Children under 15 (mother responding)†	26.1	59.0	14.9	100.0
All others, age 15 and over, not included elsewhere	29.8	58.3	11.9	100.0

* Excludes "No-No" answers, doubtful answers, and failures to answer either the interviewer or the doctor.

† The mother was also the respondent for children under age 15 in the doctor's office; otherwise all persons in the doctor's office were their own respondents.

Source: Derived from Keilin, J.E.: The Use of the Information Statistic as a Measure of Conformity in Comparing Two Sets of Responses. (Processed.) Air Pollution Medical Program.

were taped and comparison of answers to identical questions showed that:

> "In one-third of the disagreements, the subject gave definite yet different answers to the correctly-asked question at the two interviews." (p. 184)

The authors further observe:

> "Much of the inaccuracy with which answers are reported is clearly irreducible by efforts on the part of the observer.

Consistent false positive and false negative answers, and some of the random variation in the two answers to the same question cannot be avoided." (p. 187) 6/

I have been fortunate to obtain from the authors of this British study the tabulations of the first and second responses. In Table 5 I show the replies on the first and second interview to one of the questions, i.e., "Does your breathing ever sound wheezy or whistly?"

[Insert Table 5]

Table 5 - "Yes" or "No" Response to the Question, "Does your breathing ever sound wheezy or whistly?" on two separate occasions.

		Response to 2nd Questioning 1/		
		Yes	No	Total
Response to 1st Questioning	Yes	65	12	77
	No	27	40	67
	Total	92	52	144

1/ Second questioning was 6 weeks after the first.

Source: Personal communication; for description of the study see Fairbairn, A.B.; Wood, C.H.; Fletcher, C.M.: Variability in Answers to a Questionnaire on Respiratory Symptoms. Brit. J. of Prev. & Social Medicine, Vol. 13, No. 4, Oct. 1959, pp. 175-189. (The figures for the four-fold tables are not given in the published article).

In the first interview, 77 of the 144 mail carriers answered Yes and 67 No. In the second questioning 92 answered Yes and 52 No. The full measure of discrepancy is not obtained by comparing 77 and 92. In the two questionings only 65 men said Yes both times. Twelve of the postmen who said Yes first, reversed themselves in the second questioning. On the other hand, 27 who said No first replied Yes the second time.

In studies where the primary purpose is to identify certain individuals in a population with specific attributes or characteristics, I have proposed the following replicability index which I have developed.

This index consists of $\frac{a-b}{a+c} = I$, where a and b are the first row of frequencies in a four-fold table and a and c are the frequencies in the first column. For the question of wheezing, this index would be $\frac{65-12}{92} = .58$. I shall not, however, attempt to discuss the index at this time, except to indicate that its limiting values are 1 where there is complete agreement between the replicates, and its minimum value approaches minus infinity. A pragmatic rule in using the index would be never to use a test which does not yield a positive index when related to some appropiate criterion which is being tested.

6/ See Ref. 21 (Fairbairn et al), Emphasis added.

For a third example I shall use information obtained from identifiable individuals in the Current Population Survey (CPS) sample in April, 1950, and April, 1960, compared with the information on unemployment with respect to these same individuals obtained in the decennial census of 1950 and 1960, respectively.(24) 7/

In Table 6 we can see that both in the 1950 and 1960 comparisons, unemployment was reported consistently for only about one-half of the males. For the other half, a different status, i.e., employed or not in the labor force was reported either in the CPS or in the Census. For women, unemployment was reported consistently for about one-third. For the other two-thirds the status was given either as employed or not in the labor force.

[Insert Table 6]

Psychological Studies of Interviewer Bias

Another indication of bias in surveys is demonstrated by various psychological experiments. Some of these findings are touched on in two recent papers by Rosenthal, and Rosenthal and Fode. (32)(33)

In one of the studies by these authors, conducted at the University of North Dakota, subjects rated numerically a number of photographs. Ten experimenters were used and 206 subjects. The experimenters were matched and assignment of subjects was random. Half of the experimenters, however, were led to expect a mean numerical score of plus 5 and the other half, minus 5 for the photographs that were to be rated. While experiments were conducted in identical manner, Table 7 shows that the means of experimenters with a minus bias for all 5 experimental pairs were appreciably below the means obtained by experimenters given a bias of plus 5. Obviously, the experimenters' belief influenced his subjects to respond differently. This influence was effected by experimenters' voice and gestures, with or without realizing it. Still other experimental studies found that experimenters in their turn are influenced by the replies. In other words, the interviewer does not remain a constant factor during the course of the survey. As a rule, the effect of such biases are overlooked in the interpretation and use of interview results.

[Insert Table 7]

Differentials in Diagnostic Skills of Doctors

Another consideration unheeded in the mass morbidity surveys is the marked differences among physicians in their diagnostic skills, practices,

7/ It should be observed that the figures given in Table 6 are the blown up estimates, apparently the comparisons could not be made in terms of individuals sampled--individuals interviewed.

Table 6 - Cross Classification of Estimated Number of Unemployed Based on Employment Status as Reported to CPS [1] Interviewers, for April, 1950 and 1960, Compared with that Reported to Census Enumerators for 1950 and 1960, Respectively, by Individuals Who Could Be Matched, by Sex of These Individuals.

Estimated unemployment based on replies to CPS of individuals identified in the Census		1950 Census [2]				1960 Census [3]			
		Male		Female		Male		Female	
		Number	Percent	Number	Percent	Number	Percent	Number	Percent
		2,551,000	100.0	874,000	100.0	1,985,000	100.0	1,146,000	100.0
Estimates of employment status of these same individuals as reported to the Census	Unemployed	1,271,000	49.8	262,000	30.0	1,027,000	51.8	360,000	31.4
	Employed	668,000	26.2	179,000	20.5	477,000	24.0	176,000	15.4
	Not in the labor force	612,000	24.0	433,000	49.5	481,000	24.2	610,000	53.2
Estimated unemployment based on replies to Census of identified individuals in CPS		2,057,000	100.0	699,000	100.0	1,897,000	100.0	1,028,000	100.0
Estimates of employment status of these same individuals as reported to CPS	Unemployed	1,271,000	61.8	262,000	37.5	1,027,000	54.1	360,000	35.0
	Employed	515,000	25.0	111,000	15.9	554,000	29.2	268,000	26.1
	Not in labor force	271,000	13.2	326,000	46.6	316,000	16.7	400,000	38.9

1/ CPS - Current Population Survey in April 1950 and April 1960, respectively.
2/ Employment and Unemployment Hearings before the Subcommittee on Economic Statistics of the Joint Committee, Congress of the United States, 87th Congress, 1st Session, Pursuant to Sec. 5(a) of Public Law 304, 79th Congress, Washington, 1962, p. 242.
3/ Measuring Employment and Unemployment, President's Committee to Appraise Employment and Unemployment Statistics, Table K. 4, p. 392.

Table 7 - Mean score ratings given to photographs by subjects in which 5 of the experimenters were told to expect a mean score of plus 5 and 5 other experimenters were told to expect a mean score of minus 5.

Results of Exp. 1: Mean Obtained Ratings for Each E

E Pairs	Sex of Es	High (+5) Bias N	High (+5) Bias M	Low (-5) Bias N	Low (-5) Bias M	Difference of Means
A	M	19	+ 3.47	21	+ 1.81	+ 1.66
B	M	20	+ 6.60	24	- 3.71	+10.31
C	F	21	+ 4.48	22	- 4.23	+ 8.71
D	M	18	+ 2.50	20	+ 1.70	+ 0.80
E	M	21	+ 3.05	20	+ 0.40	+ 2.65
Sums		99	+20.10	107	- 4.03	+24.13
Weighted Ms			+ 4.05		- 0.95	+ 5.00*

* Significant at .007 level.

Source: Rosenthal, R.; and Fode, K.L.: Psychology of the Scientist: V. Three Experiments in Experimenter Bias. Psychological Reports Monograph Supplement 3-V12 1963.

and in their inclination to examine patients for conditions not directly associated with their immediate complaint. Physicians also vary in how they advise their patients about examination findings. This is the information on which the survey interviewer must rely. The extent or importance of these differences is not yet known. I am familiar with only two studies where a systematic attempt has been made to study the qualifications of general practitioners and their diagnostic interest and skills; (34)(35) both show wide variability. 8/ 9/

Were such studies to be extended to all practitioners (including specialists), the range of variability of skills and insight would be materially greater. Without gauging the extent of these differences, comparisons which have been made between urban 10/ and rural groups have dubious significance for health workers. (36)(37)

8/ Thus, the North Carolina study found that 45 per cent of the physicians gave physical examinations to patients fully clothed. Findings in the Hunterdon County Survey (2) showed that of the doctors (85 per cent) who responded to questionnaires, 40 per cent made various diagnoses without the use of laboratory tests; 25 per cent said that they kept no written records.

9/ These studies were made not by students interested in morbidity surveys and their usefulness but by physicians interested in the quality of general practice and practitioners.

10/ These are NHS studies comparing prevalence of disease in different geographic regions; urban, rural-farm, rural-non-farm; and metropolitan areas. While these studies purport to show the effect of geographic differences in the prevalence of specific diagnoses, actually the differences could be attributed to numerous other variables, such as threshold of complaint; variation in physician skills; cooperation level, etc.

II. Superiority of Clinical Examination

In the health field where comparable diagnostic information is desired for different population groups, I believe that evidence shows the clinical examination of an appropriate sample (7) is decisively preferable to the household survey. To reduce clinical variability, the examining team should be adequately trained and every effort made to develop objective measures and agreed-upon standards for different diagnoses. Re-examination, where essential for accurate diagnosis, would be included.

For historical and other reasons, the clinical examination method has not been widely adopted. One of the objections is the higher unit cost. This should not be considered a barrier today, however, when large-scale surveys with costs running into millions are being conducted, considering the fact that far more reliable diagnostic information would be obtained perhaps at no greater aggregate cost.

A second limitation pointed to is the low rate of participation. In studies of the Commission on Chronic Illness (2)(3) only about two-thirds of the sample selected consented to the examination. The NHS has shown, however, that with the proper approach and preparation this percentage can be increased to 85-95. The participation rate is particularly likely to suffer if the desired procedure is followed and participants are scheduled for more than one examination. While this limitation is serious, it is not decisive, and again, participation can be improved with proper preparation. Moreover, the alternate survey method with its lack of precision has shown no significant difference in morbidity rates of examination participants and non-participants in any study where this has been tested. (2)(3)(4)

The third objection, which one can scarcely take seriously, is that the physical examination is not infallible. We grant that in science, as in life, nothing is perfect, but this does not mean that alternate procedures cannot be ranked according to their comparative precision. In fact, the main end of science, I suppose, is to reduce the margin of error in measurement. If it is true that certain diseases cannot be diagnosed with any degree of accuracy with our present clinical skills, let us ascertain which these are and attempt to develop differential diagnostic criteria. Let us not, however, use this as a justification for accepting the household survey as an effective method for determining prevalence of disease. Is it reasonable to expect accurate results from a method which we can see is subject to lack of quality control; to wide differences among physicians, both in their skill in diagnosis, and their practice in informing the patient about his condition; to distortions intentionally or otherwise introduced by the patient or the respondent; and to inherent variance among interviewers themselves? 11/

11/ See footnote 1.

References

1. Sanders, B.S.: The Blind - Their Number and Characteristics. Social Security Bulletin, Vol. 6, No. 10, Oct. 1943, pp. 17-26.

2. Trussell, R.E., and Elinson, J.: Chronic Illness in the United States, Chronic Illness in a Rural Area. The Hunterdon Study, Harvard University Press, Cambridge, 1959, Vol. III. 440 pp.

3. Commission on Chronic Illness in the United States. Chronic Illness in a Large City. The Baltimore Study, Harvard University Press, Cambridge, 1957, 620 pp.

4. Cobb, S.; Thompson, D.J.; Rosenbaum, J; Warren, J.E., and Merchant, W.R.: On the Measurement of Prevalence of Arthritis and Rheumatism from Interview Data. J. of Chronic Diseases, Vol. 3, No. 20, Feb. 1956. pp. 134-139.

5. Thompson, D.J., and Tauber, J.: Household Survey, Individual Interview, and Clinical Examination to Determine Prevalence of Heart Disease. Am. J. of Public Health, Vol. 47, Sept. 1957, pp. 1131-1140.

6. Eichhorn, R.L., and Morris, W.H.M.: Respondent Errors in Reporting Cardiac Conditions on Questionnaires. Proceedings of the Purdue Farm Cardiac Seminar. Sept. 10-11, 1958, pp. 46-50.

7. A Study of Multiple Screening. Council on Medical Service, American Medical Association, Revised 1955, 91 pp.

8. Journal of Chronic Diseases, Oct. 1955, pp. 363-490.

9. McDonald, G.W.; Remine, Q.R., and Durdick, E.J.: Results of Diabetics Screening Activities, Fiscal Year 1959, Public Health Reports, Vol. 76, pp. 825-831, Sept. 1961.

10. Remine, Q.R.: A Current Estimate of the Prevalence of Diabetes Mellitus in the United States; Ann. New York Academy of Science, Vol. 82, Sept. 1959, pp. 229-235.

11. Wilkerson, H.L.C.; Krall, L.P., and Butler, F.K.: Diabetes in a New England Town. Journal of American Medical Association, Vol. 169, Feb. 28, 1959, pp. 910-914.

12. McDonald, Glen W.; Hozier, J.B.; Fisher, G.F., and Ederma, A.B.: Large-Scale Diabetes Screening Program for Federal Employees. Public Health Reports, Vol. 78, No. 7, July 1963, pp. 553-560.

13. Elsom, K.A.; Schor, S.; Clark, T.W.; Elsom, K.O., and Hubbard, J.P.: Periodic Health Examination. Journal of American Medical Association, Vol. 172, No. 1, Jan. 2, 1960, pp. 55/5-60/10.

14. Lipkind, J.B.: Evaluation of Continuous Diabetes Screening in a Hospital Outpatient Department. Public Health Reports, Vol. 78, No. 6, June 1963, pp. 471-476.

15. Schenthal, J.E.: Multiphasic Screening of the Well Patient. The Journal of American Medical Association, Vol. 172, No. 1, Jan. 2 1960, pp. 51/1-54/4.

16. Roberts, N.J. (1957). Periodic Health-Maintenance Examinations in the Early Detection and Prevention of Disease, Edited by Hubbard, J.P. The Blackiston Division. McGraw-Hill Book Company, Inc., New York, pp. 27-57.

17. Mooney, H.W.: Methodology in Two California Health Surveys. Public Health Monograph No. 70, Public Health Service Publication No. 942, U. S. Department of Health, Education, and Welfare, Public Health Service, 1962, 143 pp.

18. Cochrane, A.L.; Chapman, P.J., and Oldham, P.D.: Observers Errors in Taking Medical Histories. The Lancet, Vol. 260, No. 6662, May 5, 1951, pp. 1007-1009.

19. Siegel, G.S.: Periodic Health Examinations, Abstracts from Literature, Public Health Service Publication 1010.

20. Keilin, E.J.: A Use of the Information Statistics as a Measure of Conformity in Comparing Two Sets of Responses; Division of Air Pollution, Public Health Service (processed).

21. Fairbairn, A.S.; Wood, C.H.; Fletcher, C.M.: Variability in Answers to a Questionnaire on Respiratory Symptoms. Brit. J. of Preventive & Social Medicine, Vol. 13, No. 4, Oct. 1959, pp. 175-189.

22. Douglas, J.W.B., and Blomfield, J.M.: The Reliability of Longitudinal Surveys. The Milbank Memorial Fund Quarterly, Vol. XXXIV, July 1956, No. 3, pp. 227-252.

23. Eckler, A.R.: Extent and Character of Errors in the 1950 Census, The American Statistician, Vol. 7, No. 5, Dec. 1953, pp. 15-21.

24. President's Committee to Appraise Employment and Unemployment Statistics; Measuring Employment and Unemployment; Government Printing Office, Washington, D.C. 1962, 412 pp.

25. Sanders, B.S.: Have Morbidity Surveys Been Oversold? American Journal of Public Health, Vol. 52, No. 10, 1962, pp. 1648-1659.

26. Health Interview Responses Compared with Medical Records. National Health Survey, Public Health Service Publication No. 584, Series D, No. 5.

27. Lilienfeld, A.M., and Graham, S.: Validity of Determining Circumsion Status by Questionnaire as Related to Epidemiological Studies of Cancer of the Cervix. J. of the National Cancer Institute, Vol. 21, No. 4, Oct. 1958, pp. 713-720.

28. Feldman, J.J.: The Household Interview Survey as a Technique for the Collection of Morbidity Data. Journal of Chronic Diseases, Vol. II, No. 5, May 1960, pp. 535-557.

29. Hyman, H.: Do They Tell the Truth? Public Opinion Quarterly, Vol. 8, 1944, pp. 557-559.

30. Parry, Hugh J., and Crossley, H.M.: Validity of Responses to Survey Questions. Public Opinion Quarterly, Vol. 14, 1950, pp. 61-80.

31. Scott, Christopher: Research on Mail Surveys. J. of the Royal Statistical Society, Series A, Vol. 124, Part 2, 1961, pp. 143-205.

32. Rosenthal, R.: On the Social Psychology of the Psychological Experiment: The Experimenter's Hypothesis as Unintended Determinant of Experimental Results. American Scientist, Vol. 51, No. 2, June 1963, pp. 268-283.

33. Rosenthal, R., and Fode, K.L.: Psychology of the Scientist: V. Three Experiments in Experimenter Bias. Psychological Reports, Monograph Supplement 3-V12 1963, pp. 491-511.

34. Peterson, O.L.; Andrews, L.P.; Spain, R.S.; and Greenberg, B.G. (1956): An Analytical Study of North Carolina General Practice. J. Med. Education, Vol. 31, Part 2, 12 pp.

35. Clute, K.F.: The General Practitioner. University of Toronto Press, 1963.

36. NHS Health Statistics, Series C, No. 5. Geographic Regions and Urban-Rural Residence, United States, July 1957-June 1959.

37. NHS Health Statistics, Series C, No. 6, Geographic Divisions and Large Metropolitan Areas, United States July 1957-June 1959.