Statistics in Clinical Appraisals of Nutritional Status

EDWIN B. BRIDGFORTH, M.S.*

I SHALL consider briefly a variety of points in the area of nutritional appraisal. It is easy to obscure the most important considerations by concentrating attention on minor details, while neglecting consideration of the over-all rationale of approach. It is not true, in terms of current approaches to nutritional assessment, that our results are no stronger than some of our very weak links. Many of the measures commonly recorded are likely totally irrelevant to nutriture, and the weightings of the various bits of evidence on status are not equal. Some physical signs or biochemical levels yield high diagnostic probability, while others affect only slightly the chance of accuracy of a diagnosis.

When our objective is "group diagnosis," as in the epidemiologic approach of the nutrition survey, we can reach conclusions only by analogy from the evidence obtained in the examination of subjects which points toward an "individual diagnosis" for some. In a field survey we usually aim at broad coverage of many subgroups, obtaining limited individual information from a large number of subjects. We do not, in general, expect definitive diagnoses of specific deficiencies in individual subjects, as might be made for certain nutrients by further detailed physical examination or biochemical testing which we are not prepared to attempt, or by therapeutic trial.

There are two grave difficulties in proceeding from recorded nutrition survey data to its interpretation. First, nutritional deficiency-adequacy is not a simple dichotomy of present-absent. If deficiency is taken to include all inadequacies that reflect a "suboptimal" health status, then most persons called deficient will be only marginally so. Marginal states are strongly dependent on environmental and host factors, as well as on dietary intake of nutrients, and the practical definition of marginal states, in terms of tissue, blood, excretion products, clinical signs or symptoms, is still unknown. Therefore, we are left, at best, with hazy semi-quantitative conclusions of the magnitude of specific nutritional deficiencies that we think we see in the subjects examined. Interpretation of clinical findings is guided by the results of simultaneous biochemical and dietary measurements and by the evidence of previous investigations. This evidence which we take from the literature for use as a guide in interpretation is of varying quality as scientific truth in its own setting; we remove it further for uncertain generalizations as applicable to the time and locale of the survey we need to interpret. Second, although lacking definite diagnoses on a certain percentage of our survey sample, we still need to make "group diagnoses" of the status of the population. Some surveyors have by *tour de force* accomplished this feat by equating the percentage of their sample which was recorded as having some "key" lesion, or the percentage having biochemical levels below some chosen value, to be a reasonable approximation of the percentage of the population with a certain nutritional deficiency, due to an insufficient supply of that nutrient in the diet. A more conservative approach is to recognize the many limitations and uncertainties, and to tread cautiously in this area of *applied* research with an uncertain base of knowledge, admitting in interpretation the mere plausibility of posited deficiencies. An even more conservative approach is to confirm by controlled experimental trial the rela-

From the Division of Biostatistics, Department of Preventive Medicine and Public Health, Vanderbilt University School of Medicine, Nashville, Tennessee.

* Assistant Professor of Biostatistics.

Presented at the Symposium on Recent Advances in the Appraisal of the Nutrient Intake and the Nutritional Status of Man, on March 6–7, 1962, at the Massachusetts Institute of Technology, Cambridge, Massachusetts.

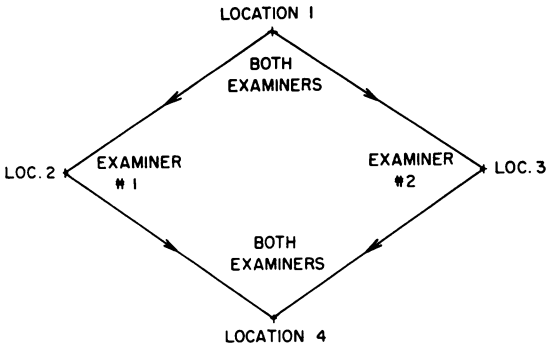

FIG. 1. A hypothetical survey schedule with two examiners and four locations.

tion of suspicious signs to the nutrients of the diet.

EXAMINER DIFFERENCES

I turn now to problems of recognition of clinical signs. How objectively can this be done and with what degree of accuracy and reproducibility? An examining schedule is diagrammed in Figure 1, for a survey with two examiners, initially working together, then departing to separate locations and finally returning again to work together. Assume that at each location 200 men are examined, with each examiner seeing 100 men of similar status at locations 1 and 4. Figure 2 shows what may happen. Examiner 1 finds 5 per cent of the subjects at each location with lesion X, and examiner 2 finds 35 per cent at each location with this lesion. Unless one or both examiners have oscillated in their criterion for recording lesion X, and assuming that the time interval between examinations at the different locations is not a material factor, there are no differences in the prevalence of this lesion at the four locations. But, if we examine these rates without regard to the examiner, we find locations 1 and 4 average at 20 per cent, with location 2 distinctly lower and location 3 distinctly higher—statistically significant, as they say. (Such erroneous differences may also occur with a single examiner recording a sign which is present only in a certain subgroup of the population, such as elderly men. If a high proportion of the sample is elderly men in some locations, and a low proportion in others, the total prevalence for each location may be quite different even though the age-sex specific rates are identical.)

How can we interpret such data? First, we should try to prevent such discrepancies. This could be attempted by setting up a training program for examiners, followed by trials in which the examiners make independent duplicate examinations of subjects in order to determine their areas of disagreement. Then, retrials should be carried out until a specified standard of agreement is attained on all signs being sought. There are some obvious limitations to such presurvey trials, such as nonavailability of subjects having certain of the signs. To compensate for this lack, color photographs or physical models might be used as a substitute. In addition, the duplicate examination of a subsample of the subjects could also be adopted as part of the routine survey procedure.

Granted that the events have occurred, such data can be separated by examiner, or we can attempt to weight the results in some way. With the present example, we might average the examiners' findings and equate 5 per cent findings by examiner 1 to 20 per cent to be reported, and 35 per cent findings by examiner 2 also to 20 per cent to be reported. This striking of an average and inflating the estimates of the examiner with below-average finding rates while discounting the estimates of the examiner with above-average rates has no sound rationale. The modification of the actual finding rates of the examiners could be accomplished by use of either additive constants or multipliers, both of which can produce anomalies of estimates sometimes falling below 0 or above 100 per cent. If, instead, we consider one of the examiners to be more competent and

EXAMINER 1	EXAMINER 2	TOTALS
LOCATION 1	LOCATION 1	
5% in 100 men	35% in 100 men	Loc. 1 20% in 200 men
LOCATION 2	LOCATION 3	Loc. 2 5% in 200 men
5% in 200 men	35% in 200 men	Loc. 3 35% in 200 men
LOCATION 4	LOCATION 4	
5% in 100 men	35% in 100 men	Loc. 4 20% in 200 men

FIG 2. A hypothetical set of survey findings, for the survey of Figure 1.

adjust the other's finding rates to his as the standard, we admit that the quality of the data of the less competent examiner is poor, and face the same problem of method of adjustment.

The difficulties of interpretation are much greater if, for example, a survey is made by ten clinicians, who examine unequal numbers of subjects in each of twenty locations. It would be necessary to tabulate the findings of each examiner at each location (possibly further by age and sex), then to inspect for examiner difference and its effect on the total finding rates, then to proceed cautiously with recombination of the data, possibly "disqualifying" the findings of some examiners on certain signs and excluding their data. It is neither easy nor scientifically sound, but the evidence of the findings may, with luck, be sufficiently clear cut to survive the process.

Figure 3 shows a hypothetical situation in which 1,000 subjects underwent independent duplicate examinations by two examiners, A and B. The results for a certain sign are given in the bottom half of the chart. Although both examiners record positive findings in seventy of the 1,000 subjects, a 7 per cent prevalence, they agree on the presence of the sign in only forty-five subjects, and each recorded positive findings in an additional twenty-five not recorded by the other. If we assume both examiners to be equally expert in identifying this sign, then the signs in the subjects not agreed upon must have been on the borderline. In the upper part of the figure is postulated a simplified explanation, with 900 subjects unequivocal in status (880 with no evidence and twenty with distinct evidence); the remaining 100 subjects show doubtful indistinct evidence. I have ignored sampling variation and assumed that the two examiners apportioned half the findings in these men as positive and half as negative. The over-all results show considerable correlation in the examiners' classification of the subjects. However, the group in whom classification was disagreed upon is larger than the group in whom there was agreement on presence of the sign. The forty-five subjects in whom the lesion was considered present by both examiners is a heterogeneous group: an indistinguishable mixture of the twenty with definite lesions and the twenty-five with "borderline" lesions.

		AGREEING ON 900 DEFINITE SIGNS			"COIN-TOSSING" ON 100 BORDERLINE SIGNS		
		EXAMINER A			EXAMINER A		
		−	+	TOTAL	−	+	TOTAL
EXAMINER B	−	880	0	880	25	25	50
	+	0	20	20	25	25	50
TOTAL		880	20	900	50	50	100

TOTAL GROUP:

		EXAMINER A		
		−	+	TOTAL
EXAMINER B	−	905	25	930
	+	25	45	70
TOTAL		930	70	1000

FIG. 3. Findings in 1,000 duplicate examinations by two examiners and assumed situation which might produce such results.

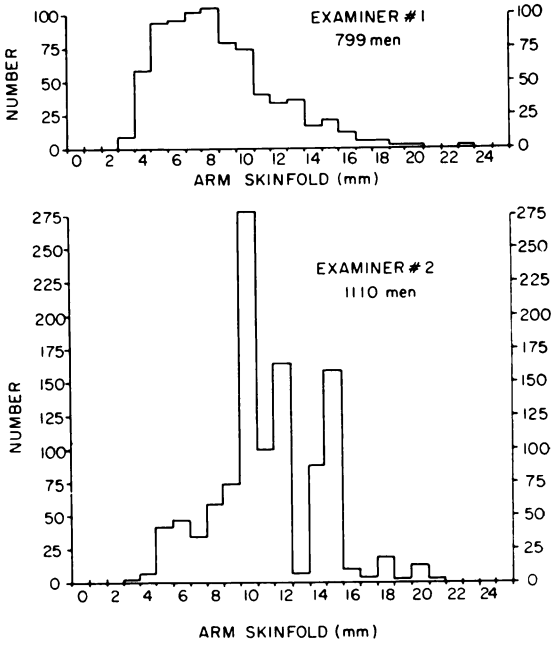

FIG. 4 Arm skinfold measurements in a survey by two physicians examining equivalent groups of subjects.

I have tried this general model on numerous sets of actual data and have found that it seldom fits. Usually the examiners have different "levels of suspicion" of evidence of signs present and so produce differing prevalence estimates. Effectively, they do not agree on the doubtful group for which they figuratively toss a coin. Some also have levels of suspicion which oscillate from day to day. Generally I

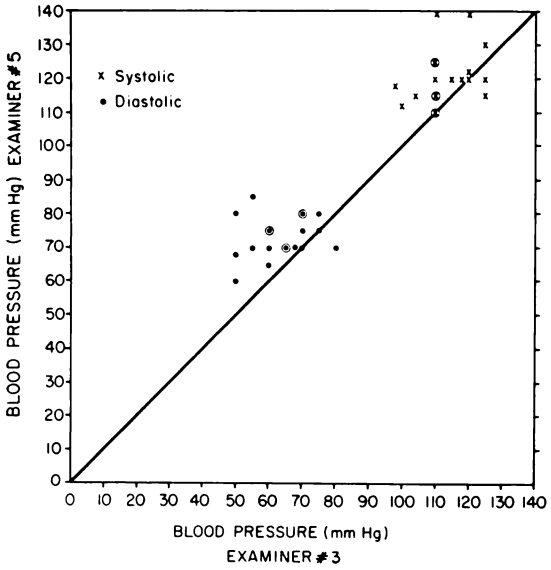

FIG. 5. Blood pressure standardization trial by two physicians preceding a survey. Duplicate measurements of nineteen men.

have found the groups in whom the classification was disagreed upon to be much larger than those in whom positive findings were agreed upon, presumably due to a large proportion of subjects examined having doubtful lesions. The other oversimplification of this model is the combining of those with doubtful status into a single group as if homogeneous, whereas in fact there are degrees of doubtfulness as to the presence of lesions.

If we are defeatists we can give up on the inherent subjective problems of definition and standardization of clinical signs and concentrate on getting some "really objective" physical measurements. Let us measure the skinfold thickness by pinching the arms of our subjects with a caliper. This measurement has been taken in many surveys. Figure 4 shows results obtained when two physicians made these measurements on equivalent groups of men. Examiner 1 has a histogram of results with respectable shape, having a broad peak of measurements in the 5 through 10 mm. range and moderate positive skewness, but examiner 2, for the most part, obtained readings in the 10 through 15 mm. range, except for a partiality against 13. Both these physicians are eminent experts on nutritional appraisal, but most certainly at least one of them is *not* an expert in measuring the skinfold thickness of the arms. Solely on the evidence of the data shown, I would wager that examiner 1 is the expert. Note, however, that the gross irregularity of the distribution of the measurements made by examiner 2, although it may cause esthetic dissatisfaction, may merely indicate that he has biases in his habit of rounding the readings of the caliper scale. A relative discrimination between subjects higher or lower in arm fatness may be achieved, separately for each examiner's subjects, but the absolute values are noncomparable. The findings cannot be pooled, nor safely compared with those of other surveys. The handling of such data in analysis and interpretation becomes a tedious project, easily avoided by some training and standardization at the onset.

Let us try another objective measurement, one of vague relevance to nutriture but frequently included in nutrition surveys—blood pressure. Figure 5 shows the results of a standardization trial, preliminary to a recent survey. Two examiners made independent measurements on nineteen men. No instructions were given since, as everyone knows, all physicians know how to measure blood pressure. However, on these nineteen men, examiner 5 consistently obtained readings higher than those of examiner 3. They did not learn from this, however, as the results of the survey (Fig. 6) show that the discrepancy in levels

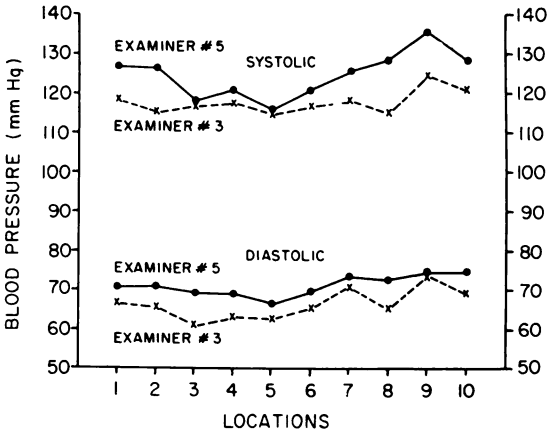

FIG. 6. Blood pressure averages at ten locations by two physicians during survey of Figure 5.

recorded by the two examiners continued throughout the whole survey.

Thus, to the need for standardization trials, we must add that these trials should not be considered an idle ritual but should be carried on until specified standards of consistency are met, with training sessions interspersed.

That nutritional signs are rarely objectively and reproducibly identifiable has been intermittently publicized since 1929.[1,2] The problems deserve more attention than they have received in most clinical nutrition studies, past and present.

SURVEY DESIGN AND SAMPLING PROBLEMS

A survey is carried out to enable inferences from the sample to the population. Generally stratifications are made of the whole population, based on prior information suggestive of subpopulations more nearly homogeneous in nutritional status than the total. These may be based on differences in amounts and types of food available, urban-rural divisions, and the like. Great difficulties may be, and commonly are, encountered in obtaining sample subjects as random representatives of identifiable population subgroups. In some areas, population rolls of sufficient accuracy may be available as the basis for a genuine probability sample, or a sampling plan already in use for other studies may be utilized. Samples of known adequacy have been obtained by these means in several recent nutrition surveys. In the absence of such opportunities, a wide variety of procedures may be applicable. A number of recent publications have included advice on sampling in medical surveys,[3-7] giving suggestions of possibilities and pitfalls and generally advising that expert help be sought or a sampling text consulted.

A sample total of *one* subject is sufficient if the whole population is homogeneous on the measures taken, but if one knew the population status in advance, there would be no need for a survey. Therefore, we have to explore by our sampling to find the actual degree of heterogeneity. We *should* accordingly set our sample sizes on a sequential basis,[8,9] but aside from problems of scheduling, there is some impracticality in a sequential approach when many different items are studied; some will require much larger samples than others to attain equivalent precisions of estimation, and we may also find large differences in heterogeneity as we encounter different population subgroups.

VERIFICATION OF SUSPECTED NUTRITIONAL INADEQUACIES

In the progress of a survey, the clinical, biochemical and dietary results, or one or two of these, may presage the final problems of interpretation by indicating possible or probable evidences of deficiencies. Clinical trials, of sound design and adequate supervision, should be set up.[10,11] Dietary intakes should be recorded. Preferably at least two levels of supplementation should be included, along with a placebo control. A double-blind procedure is essential. While it is true that a negative result is not definitive in such trials,[10,12] and great effort may be required if several trials are run for different nutrients or combinations of nutrients, the lack of such explorations may prevent any but fuzzy interpretation of the survey data.

HANDLING OF SURVEY DATA

In recent years methods of processing survey records have made use of coding of the information and machine tabulation. Data are reduced to a small number of items by coding combinations of findings together. Possibilities need to be explored for further increase in efficiency of data processing, such as use of high-speed computers for faster tabulations, and possibly more efficient coding procedures.[13] Methods are being explored for using computers for dietary calculations, and multiple regression calculations of size prohibitive for calculation by hand are now conveniently being done by computers.[14]

DIRECTIONS FOR IMPROVEMENT IN METHODOLOGY

Further work on standardization of clinical examinations to achieve better agreement in recognition of signs is needed. It has been suggested that medical students given a week of training will usually be better survey examiners than physicians. More

careful attention to the subjects for whom duplicate examinations result in disagreement may clarify the extent to which this is unavoidable due to truly borderline lesions or to poor training of the examiners. Careful descriptions and high-quality photographs showing the various stages of severity of signs may increase the likelihood that the same standards will be used in different surveys. Training for clinicians taking part in surveys should be better organized, should include trials on subjects and on slides[15] and should continue until differences in examiner criteria are reduced to a minimal level. Use of more elaborate mathematical models than the one I have mentioned should be explored to provide a basis for a more sophisticated handling of the examiner difference problems which occur. Ashford[16] has recently developed such a model for analysis of findings from x-ray films read by several observers. The pros and cons of grading lesions according to severity rather than the present-absent classification only should be re-explored. Consideration should be given to reducing clinical examination in surveys to a *very* brief screening for evidence of frank deficiency states.[17]

The use of mathematical models for diagnosis, which has received much attention lately,[18–20] but little practical application, should be explored for possible limited application in nutritional diagnosis. Such attempts will promote clearer thinking in the area of multiple deficiency states and better recognition of the extreme complexity of the problems of nutritional appraisal. The examination of models postulated to explain the time sequence of the interrelation of clinical, biochemical and dietary measures of nutriture may give a better basis for reconciling the small or absent associations between these measures, as seen on comparison of levels found in different groups[21] and also frequently in the measurements of individual subjects. The need for experimental work to clarify some of the confusion in this area would then be more evident. It may become clear that cross sectional nutrition surveys cannot provide enough evidence for a sufficiently high probability of accurate and useful conclusions, and that longitudinal follow up is required—that is, that optimal allocation of resources (money and people) can be more nearly attained by a change in the basic methods of approach.

REFERENCES

1. Bean, W. B. An analysis of subjectivity in the clinical examination in nutrition. *J. Appl. Physiol.*, 1: 458, 1948.
2. Bransby, E. R. and Hammond, W. H. Reliability of the clinical assessment of "nutritional state." *Brit. M. J.*, 2: 330, 1951.
3. Interdepartmental Committee on Nutrition for National Defense. Manual for Nutrition Surveys. Washington, 1957. U. S. Government Printing Office.
4. Doll, R. (Ed.) Methods of Geographical Pathology. Oxford, 1959. Blackwell Scientific Publications, Ltd.
5. Immunological and haematological surveys. Report of a study group. In: World Health Organization Technical Report Series No. 181. Geneva, 1959.
6. Perez, C. Scrimshaw, N. S. and Munoz, J. A. Classification of goitre and technique of endemic goitre surveys. *Bull. World Health Organ.*, 18: 217, 1958.
7. Muench, H. Catalytic Models in Epidemiology. Cambridge, 1959. Harvard University Press.
8. Hegsted, D. M. and Drolette, M. E. Sequential analysis in nutrition studies. *Am. J. Clin. Nutrition*, 8: 112, 1960.
9. Armitage, P. Sequential Medical Trials. Oxford, 1960. Blackwell Scientific Publications, Ltd.
10. McGanity, W. J. and Darby, W. J. Some considerations in making therapeutic trials. In: Symposium on Methods for Evaluation of Nutritional Adequacy and Status. U. S. Quartermaster Food & Container Institute for the Armed Forces, Chicago, Surveys of Progress on Military Subsistence Programs, S. II, No. 2, p. 82. Washington, 1954. National Research Council.
11. Council for International Organization of Medical Sciences: Controlled Clinical Trials. Oxford, 1960. Blackwell Scientific Publications Ltd.
12. Yudkin, J. The nutritional status of Cambridge school children. *Brit. M. J.*, 2: 201, 1944.
13. Davis, J. F. Notes on punched-card systems for medical research data. I. The binary-ternary-decimal code. *J. Canad. M. A.*, 82: 24, 1960.
14. Young, C. M., Martin, M. E. K., Tensuan, R. and Blondin, J. Predicting specific gravity and body fatness in young women. *J. Am. Dietet. A.*, 40: 102, 1962.

15. MERROW, S. B., CLAYTON, M. M., NEWHALL, C. A. and FOSTER, W. D. Examiners' ratings of color transparencies of clinical signs associated with vitamin deficiencies. *Am. J. Clin. Nutrition*, 5: 56, 1957.
16. ASHFORD, J. R. A problem of subjective classification in industrial medicine. *Appl. Statis.*, 8: 168, 1959.
17. Interdepartmental Committee on Nutrition for National Defense. Lebanon, Nutrition Survey, May 1962.
18. LEDLEY, R. S. and LUSTED, L. B. Computers in medical data processing. *Operations Res.*, 8: 299, 1960.
19. LUSTED, L. B. and LEDLEY, R. S. Mathematical models in medical diagnosis. *J. M. Educ.* 35: 214, 1960.
20. WARNER, H. R., TORONTO, A. F., WEASEY, L. G. and STEPHENSON, R. A mathematical approach to medical diagnosis. Application to congenital heart disease. *J.A.M.A.*, 177: 177, 1961.
21. PLOUGH, I. C. and BRIDGFORTH, E. B. Relation of clinical and dietary findings in nutrition surveys. *Pub. Health Rep.*, 75: 699, 1960.