truly said: *"Wisdom hath builded her house, she hath hewn out her seven pillars."*

The physician, too, "is worthy of his hire". Beyond that, the responsibility is ours to see that all men, when they are patients, learn the meaning of compassion as well as the science of medicine. Whatever happens and however society may be reorganized, we will continue to serve the sick according to the high tradition of our art, putting his good before all else, guarding his secrets, understanding him, and guiding him as best we can.

REFERENCES

1. STALKER, M. R.: *Canad. M. A. J.*, **84**: 155, 1961.
2. MACDERMOT, J. H.: *Ibid.*, **83**: 331, 1960.
3. SCARLETT, E. P.: *Ibid.*, **69**: 324, 1953.
4. JONES, W. H. S., editor: Hippocrates, The Loeb Classical Library, Harvard University Press, Cambridge, Mass., 1952-58.
5. TERENCE: Phormio, 2nd ed., edited by A. Sloman, The Clarendon Press, Oxford, 1890.

SPECIAL ARTICLE

AN ASSESSMENT OF RESEARCH METHODS REPORTED IN 103 SCIENTIFIC ARTICLES FROM TWO CANADIAN MEDICAL JOURNALS

ROBIN F. BADGLEY, Ph.D.,* *Saskatoon, Sask.*

RECENTLY, four articles[1-4] dealing with techniques of statistical analysis and the design of experiments have been published in this journal. The purpose of this paper is to apply some of the methodological implications of these articles to an assessment of the research techniques used in 103 studies reported in the *Canadian Medical Association Journal* and the *Canadian Journal of Public Health*. This paper is not an analysis of the content of medical experiments but it is rather an evaluation of their methodological design.

This distinction between contents and design was made in 1937 by A. Bradford Hill in the Preface to the first edition of his text, "Principles of Medical Statistics".

"Statistics are curious things. They afford one of the few examples in which the use, or abuse, of mathematical methods tends to induce a strong emotional reaction in non-mathematical minds. This is because statisticians apply, to problems in which we are interested, a technique which we do not understand. It is exasperating, when we have studied a problem by methods that we have spent laborious years in mastering, to find our conclusions questioned, and perhaps refuted, by someone who could not have made the observations himself. It requires more equanimity than most of us possess to acknowledge that the fault is in ourselves."[5]

The studies which are analyzed in this paper were originally selected to provide illustrative data for seminars to be given to third-year medical students. During the academic year 1961-1962 these seminars on methodology will precede the Third-Year Project which is one phase of the third-year course presented by this department.[6, 7] This project involves students in the preparation of comprehensive reports focusing on specific diseases. Among other questions dealing with their topics, students are asked to consider the following points:

"1. In the course of your investigation you will have read many studies relevant to this disease. Criticize the methodology used in those studies.

"2. It is unlikely that you will have been able to find completely satisfactory answers to all the questions dealt with in this Project. In one area which you consider to be inadequately covered by present knowledge you are asked to make specific suggestions for the design of research to fill the gaps."

To date, students have been guided informally by tutors in answering these questions. During the next academic year the tutorial sessions will be complemented by seminars focusing on five aspects of the design of studies using group data. These five points provide the basis for the evaluation of the 103 studies which are analyzed in this paper.

Specifically, the questions posed concern (1) how terms are defined, (2) the selection of the population or the sample of cases described, (3) the use or non-use of control groups, (4) the statistical techniques used in the analysis of data, and (5) the derivation of conclusions.

MATERIALS AND METHOD

All articles published in the *Canadian Medical Association Journal* and the *Canadian Journal of Public Health* from January 2, 1960 to July 2, 1960 inclusive were surveyed. The criterion for the selection of the articles analyzed herein was whether authors used or did not use group data in reporting original research. Consequently, case reports, reviews, descriptive papers and articles providing a survey of the literature on a given topic were omitted from the study. The remaining 103 articles, all of which used group data, were either epidemiological surveys or clinical trials of drugs and descriptions of therapeutic procedures.

The following questions were applied to each article included in the study.

1. Are the terms defined in such a way that it is possible to replicate the study?

*Assistant Professor, Department of Social and Preventive Medicine, University of Saskatchewan, Saskatoon, Sask.

2. Who (or what) does the population (or sample) being studied represent? Are the criteria given by which cases were selected or rejected?

3. What type of control group was used in the study?

4. If the results were not analyzed statistically, could statistical analysis have provided additional descriptive and analytical measures?

5. Are the generalizations induced in the conclusions of the study limited to the findings of the study?

FINDINGS

1. *Definition of Terms*

Systems of classification and the terms used in a study should be defined precisely and should be appropriate to the problem under analysis.[8] In addition to these criteria Fletcher and Oldham[9] note that "Having ensured that the terms in which the diagnosis is to be defined are appropriate and clear and that any system of classification to be used is consistent and comprehensive, it is still necessary to ensure that the individual diagnostic criteria or tests which are to be employed satisfy certain other requirements. They must be repeatable, valid, discriminating and as simple as possible."

In this study the terms or systems of classification used were not explicit in 18 (17.5%) of the articles. Listed below are a few examples of terms or systems of classification for which no other criteria were provided.

1. Normal or acceptable patients.

2. The progress of patients being classified as: excellent, good, fair, poor, and no progress; or, as slight, moderate and complete improvement.

3. Characteristics of patients as being moody, apathetic, content, antisocial, co-operative, alert, adjusted, confused and stimulated.

On the basis of the terms or classificatory systems used in these 18 articles, it is neither possible to repeat these studies in other settings nor for a reader to determine what criteria were subsumed in these categories.

2. *Selection of Samples*

Fisher[10] contends that the process of inductive inference is the only one by which new information is brought into the world. The process of induction (deriving conclusions from the particular to the general) is the basic assumption underlying all sampling techniques. Given certain conditions, it is legitimate in parametric research to draw inferences from a sample to a population. When units of analysis are being selected or a sample is being drawn from a population, the following conditions should be met.[11]

1. The characteristics of the population under analysis are explicitly delineated. It is assumed that these characteristics follow a normal distribution curve in the population.

2. The unit of analysis is specified.

3. The procedure followed in the selection of the sample is described. It is assumed that the attributes of the sample which is selected are representative of the attributes found in the population.

If the above assumptions and procedures have been followed in the selection of a sample, then two additional steps are feasible.

4. From quantitative analysis of the sample, estimates are made of the distribution of specified attributes in the population. The reliability of these estimates is calculated and presented.

5. Hypotheses about the population are tested from estimates of attributes in the population. On the basis of the accuracy or reliability of these estimates, the hypotheses are accepted or rejected (at a prespecified level of significance).

Articles in this study were grouped into three categories. Articles which met the first three criteria were designated as employing "good sampling" techniques. "Inadequate sampling" covered those studies in which samples of cases were used but which did not meet the first three criteria. The category "inapplicable" included (1) descriptive studies, (2) studies in which a total population was used, and (3) reports of small numbers of unusual cases where sampling would not have been feasible.

Using the above criteria, the following results were obtained in this study:

Good sampling	10.8%
Inadequate sampling	41.9%
Inapplicable	47.3%

Several articles classified here as having inadequate samples omitted a description of one or more of the following points: (1) the population from which the sample was drawn or what the sample represented; (2) the time span involved, or (3) the techniques used in the process of sample selection. In two articles the size of the samples being studied was omitted. Several studies, which otherwise met the criteria for good sampling, omitted the reasons why cases were subsequently dropped from the sample. In four articles samples were composed of "unselected cases" or patients were chosen "on the grounds of common sense", without additional criteria being specified.

On the basis of the information provided, it would be impossible to replicate most of these studies with inadequate samples. Also, it would be difficult in these studies to draw reliable conclusions from the groups studied to the specific populations concerned, i.e. from patients with a certain disease to all patients with that disease.

3. *Control Groups*

Several authors in this study equated the inclusion of detailed clinical observations in their reports to the use of control groups. These two facets of research are not identical. The former should be a prerequisite in all research activity. Control groups provide the constants or known variables with which unknown variables may be compared. When a procedure or a therapy is being used for the first time or is being given a clinical trial, it is crucial to know not only the effects on the cases or patients who have been treated but also how these patients differ from those who have not received comparable therapy. Essentially, the use of a control group in an experiment provides the researcher with a basis of comparison of the unknown as gauged by the known.

O. B. Ross,[12] a physician, concluded from his analysis in 1951 of 100 randomly selected experiments that in only 27% of the articles were adequate controls used. Ross's criteria for the use or non-use of control groups were followed in this study.

"The articles were classified as: those which used adequate controls; those which used inadequate controls only; those which used no controls, and those which by nature precluded controls. A satisfactory control was defined as a number of untreated patients, or procedures, approximately equal to the number treated, with the specific form of therapy being tested as the only variable factor. Controls that were held to be inadequate were those in which the number of untreated patients was too small, or a different time or place or other variables were utilized in comparing the treated and the controls, or controls were not subjected to the same physical or emotional conditions as those treated (injections and roentgen radiation). Use of controls was held to be impossible when there were reported small numbers of unusual cases in which the use of controls obviously could not have been expected or when the severity of the disease was such that none should be left untreated."[13]

In this study the category "control impossible" was extended to include descriptive and epidemiological surveys where the use of control groups was inapplicable.

The results of this study of 103 articles were:

No control	35.5%
Inadequate control	12.5%
Well controlled	25.1%
Control impossible or inapplicable	26.9%

These results, as do those of Ross,[12] highlight the need for careful planning in the development of research design. The absence or inadequate use of controls does not mean that the studies concerned may not have elucidated specific clinical problems. Rather, it is suggested that the findings of 48% (no control and inadequate control) of the articles in this study would have been more reliable, or the conclusions possibly altered, had adequate control measures been used.

4. *Statistical Techniques*

Biostatistics and methodology are the warp and woof of quantitative medical research. These two techniques complement one another in formulating research problems, in postulating and testing hypotheses and in deriving conclusions. While the content of an experiment or of a study using group data may vary, the form, regardless of content, may be assessed by a few generally accepted logical principles. These principles pertain to the organization or logical structure of a study and the statistical techniques used in the analysis of its data. Since the area of statistical theory and techniques is extensive, no attempt is made here to offer a comprehensive statistical assessment of the articles under review.

The articles in this study were assigned to one of three categories. Studies which adhered to accepted principles of statistical argument and design were classified as using "appropriate statistical analysis". Studies which deviated from the above principles were designated as employing "inappropriate statistical analysis". The distinction between "appropriate" and "inappropriate" statistical analysis is described in the discussion of the results obtained in the latter category. Finally, the third category, "additional analysis required", encompassed those studies which reported on group data but in which no descriptive statistical measures were used.

The following results were obtained:

Appropriate statistical analysis	42.7%
Inappropriate statistical analysis	24.3%
Additional analysis required	33.0%

Under the category of additional analysis required, there were 11 articles which presented raw, ungrouped data and 23 articles which used grouped data. None of these studies used descriptive statistical indices. The use of such indices can assist the researcher in deriving his conclusions and can provide the reader with a brief statement of the findings. For example, in one study which described the use of a new drug on over 100 patients, the reader was required to derive his own calculations to assess the average effect of this therapy on different types of patients. The use of measures of central tendency (mean, mode, median) and of deviation would have provided useful summary indices of all the cases in this study.

It is suggested that if descriptive statistical techniques had been used in these 34 articles where no summary indices were employed, then the data might have been presented more efficiently, and

possibly more effectively, than in the absence of such summary measures.

Twenty-five articles (24.3%) were classified as using statistical techniques inappropriately in the analysis of group data. These were studies where (1) the data were internally inconsistent, (2) the definitions of incidence and prevalence were confused, (3) several levels of significance were followed, and (4) the techniques of association were only partially used.

The data in four articles were internally inconsistent, i.e. the findings in the tables or in the descriptive reports did not tally with the totals presented. For example, in one study seven cases were dropped from the analysis. The remaining 96 cases were described, but the total number of cases initially under analysis was listed as 112 individuals. Thus no information was presented on nine cases.

In five articles the concepts of incidence and prevalence were either confused for one another or misused. The concept of incidence refers to the number of cases *arising de novo* in a specified population within a given time interval. In contrast, prevalence refers to the number of cases *existing* in a specified population at or over a given period of time. Two authors reported the incidence of cases in given diseases while actually describing the concept of prevalence in their studies. No time interval was specified in two studies and in one instance the total population under consideration was not given.

When tests of statistical probability (e.g. t-test, chi-square) are used in a study, the results are usually expressed in terms of a prespecified level of significance (e.g. .05, .01). Although the selection of a given level of significance is arbitrary, it is usual to specify it before the results are analyzed and to follow this level of significance consistently in the interpretation of the results. If the level of significance is changed to suit the data throughout the course of a study, then no consistent pattern is established for the acceptance or rejection of the hypotheses. Three articles in this study used various levels of significance in reporting their findings. An example illustrates the confusion which this situation may engender. In one study 13 different levels of significance were cited (e.g. 0.1, 0.2, .05, .01, .001) in using the t-test (testing the significance of the difference between means). When the authors of this article subsequently concluded that their findings were statistically significant, it was impossible to ascertain to which of the 13 levels of significance reference was being made.

Thirteen articles evinced confusion in the use and interpretation of the concepts of relationship (e.g. contingency, analysis of variance and correlation). In four articles authors reported that there was a significant correlation between two variables and in each instance one of the variables had not been described. In nine articles authors established the relationship between variables by using bargrams, scattergrams and graphs. Although pictorial evidence is useful, it is not a substitute for the more precise coefficients of association which could complement these findings. Statistical techniques have been devised to measure the existence, direction and degree of association between two or more characteristics. The use of such techniques (e.g. Q, r) would yield more concise statements of association than the conclusions based solely on pictorial data in which "close" or "significant" correlations were reported.

5. *Derivation of Conclusions*

Although conceptually there are different approaches to the nature of statistical inference, there is some agreement in practice about several of the common difficulties which should be avoided in research studies. These difficulties and the rules of statistical inference have been thoroughly outlined in several reports.[14-17] Only three types of conclusions which may be questioned on the basis of logical inference are examined here.

1. Conclusions drawn to units of analysis not specified in the original terms of reference of a study (e.g. from animals to humans, from a tested drug to an untested drug, etc.).

2. Conclusions derived from a single trial (a) when the population from which the sample was drawn has not been fully described, (b) which are generalized to all possible units, and (c) in which an insufficient number of cases has been used (e.g. reporting that a drug or therapy is effective for all patients with a given disease where the population has been unspecified and the results have been obtained from only 10 to 20 cases).

3. Conclusions pertaining to cause and effect (a) where only variation in the relationship of the variables has been shown and (b) where no control group has been used (e.g. concluding definitively about the effect of a particular event on the structure of human personality from a study of 12 families. Families not undergoing the event were not considered.).

The difficulties listed above are additive, not mutually exclusive categories. In 41.5% of the articles analyzed, *one or more* of the difficulties was noted. Since conclusions in research may be stepping-stones for action, this finding suggests for the studies considered that their results and conclusions should be carefully re-examined for possible alternative interpretations.

SUMMARY

To provide medical students with a basis for assessing the methodology of research studies, 103 articles using group data were reviewed. The analysis focused on (1) the definition of terms, (2) the selection of a population or sample, (3) the use of controls, (4) types of statistical analysis, and (5) the derivation of conclusions. For the articles under review, this assess-

Canad. M. A. J.
July 29, 1961, vol. 85

ment revealed the need for greater precision in the design of many studies using group data and for caution in the interpretation of results.

REFERENCES

1. PHILLIPS, A. J.: *Canad. M. A. J.*, **84**: 376, 1961.
2. MORRISON, R. T.: *Ibid.*, **84**: 487, 1961.
3. *Idem*: *Ibid.*, **84**: 545, 1961.
4. *Idem*: *Ibid.*, **84**: 591, 1961.
5. HILL, A. B.: Principles of medical statistics, 6th ed., Oxford University Press, New York, 1955, p. vii.
6. BADGLEY, R. F.: *Canad. M. A. J.*, **84**: 705, 1961.
7. ROBERTSON, A.: *Ibid.*, **83**: 1100, 1960.
8. FLETCHER, C. M. AND OLDHAM, P. D.: Diagnosis in group research. *In*: Medical surveys and clinical trials, edited by L. J. Witts, Oxford University Press, London, 1959, p. 24.
9. *Idem*: *Ibid.*, p. 27.
10. FISHER, R. A.: The design of experiments, 6th ed., Hafner Publishing Co., New York, 1951, p. 7.
11. HAGOOD, M. J. AND PRICE, D. O.: Statistics for sociologists, revised ed., Henry Holt & Co., New York, 1952, p. 188.
12. ROSS, O. B., JR.: *J. A. M. A.*, **145**: 72, 1951.
13. *Idem*: *Ibid.*, **145**: 73, 1951.
14. HILL, A. B.: Principles of medical statistics, 6th ed., Oxford University Press, New York, 1955, p. 177.
15. COHEN, M. R. AND NAGEL, E.: An introduction to logic and scientific method, Harcourt, Brace and Company, New York, 1934, pp. 316, 376.
16. HOGBEN, L.: Statistical theory: the relationship of probability, credibility and error, George Allen & Unwin Ltd., London, 1957.
17. FLETCHER, C. M. AND OLDHAM, P. D.: Diagnosis in group research. *In*: Medical surveys and clinical trials, edited by L. J. Witts, Oxford University Press, London, 1959, p. 23.

REVIEW ARTICLE

MYOTONIA DYSTROPHIA: A REVIEW OF 17 CASES

BERNARD SLATT, M.D., *Toronto*

MYOTONIA dystrophia is an heredo-familial degeneration characterized by myotonia, selective atrophy of the muscles, and dystrophic signs in other tissues including baldness, cataracts, testicular atrophy and dysfunction of the endocrine glands. This report, listed in Tables I and II, details observations on 17 patients with this condition, encountered at Sunnybrook Hospital, Toronto, between 1948 and 1960.

HISTORY

Since Erb's monograph in 1886, describing atypical forms of Thomsen's disease in which the association of myotonia with atrophy of the muscles was recognized, our understanding of this condition has burgeoned from its original narrow concept to that of a multisystem process involving various tissues and organs. Deleage, in 1890, was the first to describe this curious combination of symptoms as a distinct entity, while Rossolimo, a few years later, offered the nomenclature of myotonia atrophia for those cases in which atrophy supervened. In 1909, Batten and Gibb,[5] and Steinert,[41] simultaneously but independently, stated that the myotonia was limited in its distribution and that the atrophy showed a characteristic pattern with involvement of the facies, the sternocleidomastoid muscles, the muscles of the forearm, the extensors of the legs, and the dorsiflexors of the feet. Batton gave priority to the atrophy, stating that the myotonia was a secondary symptom. Steinert stressed the myotonia but also drew attention to the importance of the widespread dystrophic process which included testicular atrophy, baldness and acrocyanosis. Adie and Greenfield,[2] in 1911, showed that cataracts were an integral part of the syndrome when he described a family of 13 brothers and sisters of whom five suffered from myotonia dystrophia; two of the affected five had cataracts, while two others, otherwise unaffected, also suffered from cataracts. Curschmann[13] emphasized the extramuscular manifestations and coined the name myotonia dystrophia. However, the etiology remained obscure until Fleischer,[17] in 1918, showed that myotonia dystrophia was an heredo-familial degenerative disease exhibiting anticipatory signs through several generations before it developed entirely in one generation. He traced the disease through six generations and demonstrated how it burst forth in a number of families at the same distance from the common ancestor.

ONSET

The onset of dystrophia myotonia occurs commonly in the second or third decade, with weakness as the prevailing complaint. However, the age of onset varies markedly with the generation affected, since in antecedent generations the disease begins at a later age and rarely progresses to a point where the diagnosis is made before death. When the diagnosis is made in an affected child, it is often in retrospect that the features of baldness and senile cataracts are recognized as strongly suggestive evidence that the parent also suffered from the disease. Conversely, the offspring of those affected in the third and fourth generations may be expected to show the disease in the second or even first decade of life. The insidious nature of the disease and the slowness with which it progresses allows a considerable lapse of time between the commencement of symptoms and the request for relief because of disablement. Patients often compensate and accept early symptoms, so that initial hospital admission may be for treatment and in-