Problems of the Statistician in Collaborative Research

HYMAN GOLDSTEIN, Ph.D.

COLLABORATIVE research in its finest sense can be an adventure in thinking, working, and giving together in science. The statistician has the rare opportunity of trying to make it work.

But the collaborative study of cerebral palsy and other neurological and sensory disorders of infancy and childhood, supported and coordinated by the National Institute of Neurological Diseases and Blindness of the Public Health Service, supplies an example of the difficulties of collaborative research and the resulting problems for the statistician.

This is a prospective study to determine the relationship of certain biological, genetic, and environmental factors in parents to the occurrence of abnormalities in the products of conception.

An adequate study of reproductive failure must fulfill four basic requirements if it is to succeed:

1. The study must be capable of evaluating simultaneously the many etiological factors which may be responsible for causing fetal wastage.

2. The varied abnormalities or forms of fetal wastage must be observed and differentiated.

3. The data obtained must not be subject to bias which would set up spurious correlations between prenatal events and postnatal defects.

4. An adequate number of cases must be studied to make possible the establishment of statistically significant correlations.

Since a number of separate institutions across the country are collaborating, it is essential that the study be conducted according to a single design, the data being collected and recorded in a uniform fashion and reported to a central office for continuing analysis.

In this investigation, the gravida (pregnant woman) is the starting point. Since it is not possible to determine in advance which outcomes will be normal and which defective, the study has built-in controls. There is interest in all products of conception and, in effect, disorders of all body systems, not merely the neurological, whether such abnormalities appear at time of delivery, during infancy, or during early childhood.

The study is still in the phase of pretesting study forms and data collection procedures. Of the 15 institutions in the study, 13 serve, and hence contribute to the study, obstetrical clinic populations heavily weighted with Negro gravidas of low socioeconomic status. The other two institutions collaborate by followup of babies born in project hospitals in their communities.

It is necessary to obtain from prospective parents detailed information on genetic, biological, and environmental factors which might be germane, both before and during pregnancy. The pregnant woman is subjected to detailed and meticulous examinations throughout pregnancy and during labor and delivery. An equally detailed evaluation of the products of conception, including both those resulting from

Dr. Goldstein is chief of the Biometrics Branch, National Institute of Neurological Diseases and Blindness, Public Health Service. *This paper was presented before the biostatistics division of the New York area chapter of the American Statistical Association, in New York City, December 10, 1959.*

uncompleted pregnancy and those at term, must follow. Periodic evaluation of the live offspring is to be continued throughout infancy and early childhood, including general and special examinations at regular intervals and pathological examinations whenever possible.

The factors and conditions to be investigated in the parents are:

1. Conditions of pregnancy itself, such as infection, trauma, bleeding, drugs, and progress of labor. This includes the normal and abnormal physiology of pregnancy, labor, and delivery.
2. Environmental factors influencing the mother, such as socioeconomic conditions, emotional stress, and medical care.
3. Biological factors, such as age, parity, medical and reproductive history, and immunological characteristics.
4. Genetic background.

It should be emphasized that the necessity for accurate detailed data is not confined to information regarding the causative factors during pregnancy, but applies also to the evaluation and categorization of presumed result or defect. The study will require accurate and detailed differentiation of the types of abnormalities observed.

What are some of the methodological issues that the statistician in collaborative research must face? Many are no different from those found in noncollaborative research, but the sources of variation in collaborative research are more numerous in all areas. Some of the issues may be described under size of sample, case selection, followup, the data and their reliability, and results.

Size of Sample

Using the expected incidence of 5 cases per 1,000 live births for cerebral palsy within the first year of life and assuming that approximately 40,000 live births will be studied, some 200 cerebral palsy cases are expected. This is considered the minimum number of cerebral palsy cases required for analysis, in view of the great number of variables to be cross-classified and the fact that cerebral palsy is a complex disorder manifesting itself in different ways. One thing is certain, because of the low incidence of many of the conditions and the multiplicity and interrelationship of the various factors of pregnancy under study, a large number of cases is essential. The goal of 40,000 live births is believed possible from a total of 50,000 pregnancies, making allowance for expected fetal loss. Even with a sample of this size, it is problematical whether the study will provide a sufficient number of damaged cases to permit the achievement of statistically significant conclusions.

If it is assumed that 40,000 live births will be studied, one may well ask: What sort of differences in incidence rates might be observed with this number? Lilienfeld and Parkhurst (*1*) observed in an extensive series of some 95,000 live birth certificates an incidence rate for cerebral palsy of 5.3 per 1,000 among children whose mothers had no complications associated with pregnancy and parturition. Among children whose mothers had such complications (slightly more than 1 percent), the incidence rate for cerebral palsy was 18.2 per 1,000 live births. If such a difference should exist in the population to be studied by the collaborative project, would it be possible to detect it by the usual statistical means? Or, thinking of it slightly differently, what would be our probability of detecting it with the study patients?

Assuming that the figures just given are true for the collaborative project, the difference would have about a 25 percent chance of being missed, at the 1 percent significance level. At the 5 percent level, the chance of being missed is 13 percent.

Tests of this sort were made for a number of other incidence rates in the literature. It should be pointed out that the incidence rates for conditions and defects are based on a spotty literature. Quite often the rates apply to a specific hospital or to a given locality, or the method of selection is biased or obscure. Furthermore, whatever is available represents only a few conditions or defects. There is no way of determining the chances of detecting significant differences for conditions or defects for which no incidence rates are reported.

It seemed fairly certain that for a number of particularly uncommon causes or conditions, the contemplated size of sample would be inadequate, especially with so many variables to

be included in this study. It seemed, therefore, that some method had to be found for permitting inclusion for analysis of a larger number of damaged or defective children. Such data could be obtained by using information already available from patients outside the sampling frame but in the collaborating institutions or from other institutions and agencies in communities where the collaborating institutions are located. Although of necessity such information might be less detailed and less definitive, it would be obtained from a larger number of pregnancies with a relatively small expenditure of effort.

Briefly, one such approach, retrospective in nature, consists of: (*a*) attempts to identify in the community, as nearly as possible, all damaged and defective children born in that community during the period of study; and (*b*) comparison of the events in the mothers' pregnancies obtained from the data available in the written records, mainly from hospitals, with data obtainable on pregnancies which resulted in presumably "normal" children.

The success of this type of undertaking depends largely on the ability to select the key items of information required, and to find ways of obtaining data on such cases which are identical to or at least comparable with the same items of information derived from the central core study. By this approach, it will be possible to study in detail certain specific suspected factors inadequately covered in the central core or "intensive phase," and to obtain valid data on incidence and prevalence of certain gross defects such as prematurity, cerebral palsy, blindness, deafness, and mental deficiency.

The various methods and studies by which the additional cases and types of data can be obtained are included under the term "extensive phase" of the collaborative project. Activities which are basic to fulfilling the objectives of the extensive phase are casefinding, record review, and the estimation of reliable and valid population parameters.

Case Selection

Because of the highly selected nature of the institutions in the intensive phase, their patients, as has been mentioned, are not a representative sample of the general population, and the experience reported, including incidence and prevalence figures for complications of pregnancy, outcome, and the like, cannot be considered representative. From this viewpoint the total sample in the project might be considered as a sample in search of a population. It was believed that the results obtained from a sample of a given hospital's obstetrical population should be generalized to the total obstetrical sampling frame of that hospital. Accordingly, in institutions not contributing 100 percent of their obstetrical population to the study, a form of systematic sampling was started, taking into account the size of the obstetrical population and the anticipated contribution. Although a basic scheme of unbiased sampling is used, it is not always possible to employ a single sampling design. There may also be slight variation from place to place regarding the basis for exclusion from the sampling frame.

For each obstetrical patient in the sampling frame, including those not registered in the study, a registration form, giving data on age, weeks of gestation, marital status, and race, is completed and sent to the central office by the institution. By periodic analysis of registration forms sent in for all obstetrical patients in the sampling frame, significant changes in number and type of obstetrical patients can be noted, and changes in the sampling ratio can be effected wherever necessary.

A number of modifications of the completely random or unbiased case selection method have been recommended by some of the collaborating institutions. The reasons for these recommendations have been several.

First, it has been suggested that the selection of patients should be based on the likelihood of a potentially higher yield of defective outcomes because of specific characteristics. Such a selection of high-risk cases might be based on parity, age, existence of previous complications of pregnancy, previous outcomes resulting in a number of abortions, and so forth. The evidence available shows that each of these factors might be associated with an increased incidence of defect in the offspring, and thus an increased likelihood of finding out more about the mechan-

ism through which these recognized factors produce defect.

However, there are a number of objections to a selection of this type. To the extent that cases are selected on the basis of known or suspected etiological factors, the likelihood of detecting presently unrecognized or unsuspected factors in perinatal morbidity is reduced when the number of patients to be studied is fixed. Furthermore, when a special selection basis of this type is set up, one can never be sure that some additional bias is not being introduced which is not evident superficially in the basis of selection used. Thus, one of the chief limitations of this technique lies in the possibility of overlooking important interrelated variables which may actually be determining factors.

A second recommended basis for case selection deals with improved ease and consistency of data collection. A number of collaborators have cautioned, for example, that if we take into the study those patients who do not report for obstetrical care until late in pregnancy, there will not be accurate data available regarding the greater part of their pregnancies. Other collaborators have used this type of reasoning to justify their desire to exclude patients from the study on the basis of unwillingness to participate or geographic factors which might make it difficult to maintain adequate followup.

The primary concern in selection procedures should be to make it possible to obtain desirable and feasible data. One must guard against the serious risk that inapparent factors may exert influence on such selection and that important influences relating to the cause of perinatal morbidity may be obscured or overlooked. Insofar as possible, a study of this type should attempt to sample as complete a population of pregnant women as possible in order to insure the broadest possible basis of experience.

As a result of analysis of data on age, marital status, and race, obtained on registration forms from all obstetrical patients coming into the hospital, it will be possible to recommend special sampling procedures, such as for patients reporting for care early in pregnancy, or for patients under 20 years of age or over 40 years of age. Furthermore, it will be possible to alter the kind of information requested on the registration forms to obtain population characteristics pertaining to new or special variables. Information of this type obtained from sampling the entire frame of patients will permit reaching decisions as quickly as possible regarding special sampling ratios.

Patients coming back to the study in subsequent pregnancies provide data of unusual interest. They give some clue as to the importance of genetic or constitutional factors in pregnancy outcome. They supply unusually valuable data in respect to exposure to virus disease, since serologic data are available over a period of time. In addition, they provide information on the reliability and consistency of some of the history items.

On the other hand, repeat pregnancies reduce the number of different pregnant women included in the study and thus reduce the potential detection of significant differences. There are several methods of dealing statistically with repeat pregnancies. Any woman previously registered may be excluded from the study. Or the study may include only those repeaters who, by chance, fall again into the sample. The number of repeat pregnancies brought into the study in this way would depend on the total number of pregnancies that come into a particular hospital and the sampling ratio.

Another method is to include in the study any woman who has previously been registered in the study; in other words, deliberately induct into the study all gravidas because of their prior inclusion in the study. The number of repeat pregnancies thus brought into the study would depend primarily on the reproductive patterns encountered in that particular obstetrical population.

The decision as to an optimum percentage of repeat pregnancies for inclusion in the study must involve a balancing of the merits of including these repeat pregnancies against the loss of independent observations. Except for certain genetic considerations and other questions where the repeat pregnancies alone are of interest, one must consider that the entire sample is reduced by the number of repeaters allowed. The fewer repeaters allowed into the study, the more factors it will be possible to detect as contributing to abnormal conditions in their children.

The effect of reducing the 50,000 independent cases by allowing repeaters among them can be measured only by noting the impact of such inclusions on the chances of missing a true difference in incidence rates for defects occurring in a population of gravidas with and without specified pregnancy conditions. The chances of missing true differences in incidence rates for specific defects associated with some eight pregnancy conditions, cited in the literature, by allowing a repeat rate not to exceed 25 percent, were determined. It was found, for instance, that, for nonpuerperal complications where no repeats were allowed, using a 5 percent level of significance, the chance of missing a significant difference was 13 percent; with 25 percent repeats, it was 17 percent. For toxemias of pregnancy, the chance was 40 percent with no repeats and 50 percent with 25 percent repeats. Even with no repeaters, there is a good chance of missing some differences for certain conditions and defects.

Unfortunately, there is very little information on which to base estimates of the number of repeat pregnancies likely to be found in the study over a 5-year period of enlisting gravidas. Analysis of previous experience of three hospitals in the study would indicate that approximately 20 percent of women delivering in these hospitals return for one or more deliveries within 5 years.

Followup

Some of the collaborators have pointed out that the contemplated long followup of numerous infants in the study is the most irksome feature of the whole undertaking. They believe that cases with a greater likelihood of some defect or abnormality being present or developing should be selected for followup. It is assumed that a carefully selected control case would be included in the study for comparison with the case selected on the basis of some presumed abnormality.

The criticisms leveled against selection of high-risk cases could also be leveled against selection of the presumed defective child for followup. Although this type of case selection would be useful in detecting the mechanism of factors or influences already suspected, it would definitely reduce the likelihood of detecting presently unrecognized causes of perinatal morbidity. Moreover, bias may easily be introduced in such a technique. The danger always exists that if a subsequent examiner learns that a case has been introduced into the study because of some presumed defect, this may have an effect on his objectivity.

A more serious disadvantage is the limited number of factors which can be used as a basis for selection of cases and controls. When different categories of cases plus controls are drawn into the study at various stages, the net result might very well be a great variety of procedures and relatively small groups of cases studied in different ways, with doubtful suitability for comparison with the main group of cases. Moreover, when all the special interest cases plus an equal number of controls are added, there is a good possibility of ending up with practically all the cases in the study. It would appear that the difficulties introduced into the study by a selection of infants based on their presumed defect outweigh any possible advantages.

Followup, one of the most critical problems of the collaborative project, was recently the subject of a careful review by one of the collaborating institutions. In a series of eighth-month examinations scheduled for a 2-week period, 86 percent were completed, 8 percent were temporarily delayed, and 6 percent were permanent losses, because patients dropped out of the study, moved out of town, or were lost for some other reason. From the foregoing it is obvious that, despite diligent effort, dropouts and delinquents are likely to pose a serious problem. Building rapport with the family, inculcating a feeling of contribution to a humanitarian effort, publicizing the project, stimulating identification with the project through various means, paying carfare for those unable to come for followup examinations otherwise, providing baby sitters where necessary, and the like are all mechanisms which can be used to try to keep the sample intact. In fact all mechanisms that have proved to be effective in tracing families should be employed, such as use of the social service exchange and skip-tracing services. It is essential not to lose the time, effort, and money invested in each neonate.

When a family is lost to followup, death certificates should be checked. This might supply endpoints for a fraction of such lost cases.

There might be differences of opinion with respect to the frequency of followup examinations, ranging from examinations to be made as frequently as possible to no followup examination until such time as the child is of sufficient age to permit a definite decision on the status of his nervous system through examination. The first approach is impractical because of the huge drain on the resources and personnel of the hospital and on the time and energy of both mother and child. The defects in the last approach are dual. The first possibility is that the nature and character of certain minor defects, evident at an early age, might be obscured by variations in training "compensation." For instance, the true nature of a speech defect can be confused when the child is not seen until after reactions to training and school experience are set. The other weakness of this approach lies in the fact that it makes no provision for evaluation of the importance of injuries and illnesses and other events which may lead to neurological damage subsequent to the birth of the child and prior to the "definitive" examination at the end of the study. The earlier a given defect can be demonstrated, the less likely its erroneous correlation with some subsequent event.

The Data and Their Reliability

In a prospective study the data recorded at any given time should not be biased by reference to preceding events or modified in the light of subsequent developments. To this end, data should be recorded at the earliest possible time after the event reported.

For the purposes of this collaborative study, it was stressed that copies of the record of an event must be forwarded to the central office for coding, processing, and analysis without delay. Examinations should be conducted, and the results recorded, without reference to previous events or examinations. Perhaps, however, the best that may be achieved is a series of mutually exclusive "bias-free" blocks, one for the prenatal data, another for the events of labor and delivery, and still another for the postnatal data, with different examiners for each of the three blocks. It is especially important that examinations of the infant are carried out without knowledge of possible favorable or unfavorable circumstances in the parents or in the environment. Sometimes the delivery room examiner of the neonate may be aware of events during delivery. If so, such awareness should be recorded so that it may be taken into account in analysis of the data.

It is recognized that the ideal is unobtainable. In many instances, the examination itself will elicit information relative to previous events. In other instances, the examiner himself may recall pertinent details from a previous contact with the patient. In addition, the necessities of time and personnel may require screening or selective procedures which of themselves are an indication to subsequent examiners of the possible existence of an abnormality. Finally, medical care and ethical considerations often require that every possible means of evaluation be utilized.

The achievement of uniformity of data collection and reporting represents the greatest difficulty in a study of this magnitude, where the collaborative efforts of a number of institutions and the coordinated activities of individuals of many disciplines are involved. In order to insure this essential uniformity, several measures are being used.

1. The study is being conducted according to a single design, and the data collected are assembled, coded, and analyzed within a single central office.

2. Training sessions for participants have been developed, and it is hoped that uniformity may be maintained by frequent exchange of personnel among the collaborators.

3. There is a continuing review of data and procedures by personnel of the central office. However, each institution must be on the lookout for biases, errors, and other inadequacies. These deficiencies, unfortunately, cannot be discovered as efficiently or as rapidly through central-office editing alone. Local editing permits checking for completeness of study forms and accuracy of the interviewer or recorder and affords early detection of consistent errors.

Periodically, the central office compares institution with institution for the percentage of

times a given item on a form is left blank and other aspects dealing with reliability and consistency of data. These comparisons are sent to the institutions so that each can compare itself with the others. Variations may be explained by differences in clinic populations (race, parity, geographic locations, and so forth) or by differences in standards, definitions, and clinical interests. Agreement must be reached among members of a given discipline on standards, definitions, and even abbreviations to be used.

The project director is considered responsible for the accuracy and quality of the data submitted. It is his responsibility to insure that all records for submission to the central office are checked promptly for completeness, legibility, and reliability.

Recently a review of Apgar Scores (a composite score based on an evaluation at a specific time of the neonate's heart rate, respiratory effort, muscle tone, reflex irritability, and color, compared with given standards, some of these components being more objectively rated than others) for 408 neonates from six collaborating institutions revealed a great variation in distribution of scores among the institutions. The least distressed baby could score 10, the most distressed, 0. The great variation in distribution of scores is indicated by the fact that the percentage of neonates scoring 8 or higher ranged from 12 percent in one institution to 84 percent in another. These great differences in score distribution were also evident in each of the five component scores of the Apgar test. Such great variation might conceivably be due to actual differences in neonate populations, but it is unlikely. It is probable that greater adherence to procedure in making the test, such as time after birth at which test was given, inaccessibility to prior knowledge by the rater of the gravida's pregnancy risk, or reduction in inter- and intra-rater variability, can reduce the variation. These and similar analyses will continue to be made.

Concerning validation of data, one of the project institutions has studied the degree to which birth weight information given by the gravida concerning her prior pregnancies compares with birth record data. Additional studies by other project institutions attempt to validate information given by the gravida about her sisters and their offspring, by comparing it with similar information obtained from the sisters themselves. Underway is a pilot study of the actual examination of the relatives reported by the gravida as having neurological disorders to determine over-reporting on the part of the gravida and the extent of inaccuracies in this type of reported data.

It is expected that a number of studies to determine the validity of the gravida's response to questions concerning her medical history, by comparison with records of medical care, will be undertaken with several prepaid health plans. Efforts to validate data will continue to receive urgent attention.

The Results

All records will be machine processed and tabulated centrally. It is planned to prepare a feedback summary statistical report periodically (probably annually) during the course of the study. These reports will include summary tabulations for each institution and for all institutions combined. Because of the complexity of this study and the multitude of variables to be analyzed, it would not be feasible or practicable for these reports to be prepared in great detail. The summary tabulation will be in the form of relatively simple distributions, such as two-way classifications for the items of major importance and interest. No attempt will be made to prepare detailed, periodic analytical reports describing associations or correlations of perinatal events with the development of neurological and other sensory disorders until data on a sufficient number of pregnancy outcomes have been collected to allow meaningful interpretation.

Reports and tabulations prepared periodically will also be used to determine the general distribution of the populations sampled in each institution with respect to the variables under consideration. They will be used, in addition, on a continuing review basis, to point up possible major differences between the institutions, due either to random variation, differences in respect to populations sampled, or differences in procedures, techniques, and methods of data collection. In interpretation of these data, con-

sideration must be given to these possible sources of differences.

This review of the preliminary distribution of data, obtained from the tabulations mentioned above, will, it is hoped, help to determine the most expeditious way of treating and analyzing the data. For example, the observed distribution of the various items can be used as a guide for setting up feasible cross-classifications of items, such as prenatal factors, for analysis and will indicate, to some extent, types of analytical methods and tests to be employed. One possible first approach would be a series of 2 x 2 chi-square tables. This type of massive analysis is now feasible through the use of high-speed computers which permit the otherwise laborious computations of large volumes of data in a relatively short time and at reasonable cost.

In this study it is necessary to determine the incidence of certain types of stress among the newborn through the first several years of life from gravidas with given prenatal conditions or events as compared with those without such conditions or events, controlling as nearly as possible for other relevant variables such as age of mother, race, and previous pregnancies. For example, the mothers will be subdivided into groups according to certain characteristics, environmental, biological, or genetic, which they or their husbands possess, and according to the course of pregnancy, labor, and delivery. The incidence of neurological and other conditions, of pregnancy wastage, and of childhood mortality will be studied in each of these groups. Associations or significant relationships which may be revealed from the analyses may require additional well-controlled studies of a specific nature.

In analyzing the data, consideration must be given to the fact that a large number of variables are being studied. Many of the variables to be studied are not independent of one another, which is a further complication. For these and other reasons, in the process of examining cross-classified tabulations, perplexing questions may arise. To arrive at answers, it may be necessary to employ complex analytical techniques, such as multivariate analysis.

These are some of the methodological issues that face a statistician in collaborative research. While statistical decisions must be uniform for all of the institutions in a study of this type, the application of these decisions and their feasibility may vary from place to place. Practicality may occasionally compromise the difference between what should be done, or the ideal, and what can be done, or the situation as it is.

Finally, a project of this magnitude represents a considerable burden superimposed on routine hospital administration. Under no circumstance can routine administration be disrupted by the study. Consequently, the study must be fitted into the hospital situation as it exists. Since no two hospitals function in the same way, this may, in itself, account for some of the interhospital variation.

One of the most important objectives of this study has been relatively unemphasized, that is, to obtain a greater knowledge of the methods used in conducting such long-term, interdisciplinary, collaborative undertakings. If successful methods can be devised, a completely new and almost limitless area of research will have been opened up, research where the united efforts of many individuals and institutions can be brought together in a common undertaking.

REFERENCE

(1) Lilienfeld, A. M., and Parkhurst, E.: A study of the association of factors of pregnancy and parturition with the development of cerebral palsy. Am. J. Hyg. 53 : 262–282, May 1951.