USE OF CASE RECORDS IN THE STUDY OF THERAPY AND OTHER FEATURES IN CHRONIC DISEASE

I. PLANNING THE SURVEY

BY

DONALD MAINLAND

From the Department of Medical Statistics and Study Group on Rheumatic Diseases,
New York University College of Medicine

(RECEIVED FOR PUBLICATION OCTOBER 6, 1955)

Experiments and Surveys

Probably few physicians now fail to appreciate the need for a properly conducted experiment if there is doubt whether a treatment is beneficial, or if the relative merits of two treatments are not known. The essential features of such a controlled therapeutic or clinical trial are easily stated:

1. Allocation of treatments (e.g. two drugs or drug and placebo) to patients by an automatic random technique, such as random numbers.
2. "Double blindfold" method, whereby patients are kept unaware of the differences in treatments, and those who assess the results are unable to identify patients with treatments.

The conduct of the experiment is not easy (Daniels, 1951; Green, 1954; Herdan, 1955; Hill, 1951b, 1952; Mainland, 1954b), and it is natural to ask what information can be gained by a non-experimental method, such as a survey of case records, supplemented, where possible, by follow-up studies (interviews or mail questionnaires). In the comparison of treatments, Densen and others (1952) have shown that even under exceptionally favourable conditions a case record survey can produce very misleading conclusions. Although the records often contain the results of "trial and error" experiments, these cannot be utilized as experiments because they do not insulate the control of hidden bias.

Possible Uses of Case Record Surveys

There are, however, two other common problems:

1. Estimation of the amount of benefit obtainable from a treatment already known to be beneficial;
2. Estimation of the deleterious effects of such a treatment.

The exploration of the possible value of case records in such problems was one purpose of the American Rheumatism Association's Co-operative Study of Cortisone Therapy in Rheumatoid Arthritis, described elsewhere in this issue (p. 325) and hereafter referred to as the "Study". This Study should by no means be taken as a model. It was made extensive in order to ascertain the quality and the amount of response possible in a forward-going survey or in a therapeutic trial, but it retains the features of a preliminary exploration, including many defects that would have to be removed before even a pilot study (a small-scale dress rehearsal) could be started. Nevertheless, its scope and faults render it useful in an attempt to answer three questions:

1. What can be safely inferred, from a case record survey, regarding the benefits and drawbacks of a particular treatment, especially in a complex chronic disease?
2. What reliable information can such a survey give regarding aetiology or clues thereto, and inter-relationships of phenomena, between diseases and within one disease?
3. How far can a clinician conduct such a survey, even on his own records, without the help of an epidemiological statistician?

The third question is vital, because such statisticians, even medical statisticians of any kind, or biological statisticians with time for medical projects, are far too few to help in more than a very small fraction of current researches. This problem was examined recently (Mainland, 1954a) from the point of view that every investigator is already to some extent his own statistician, but that many investigators need to improve their skill by grasping certain general principles and by acquiring a few simple techniques of design and analysis appropriate to their particular projects. This would limit the scope and complexity of many inquiries and eliminate some entirely, but the general quality of such work would improve.

A clinician who desired such skill in case record surveys could obtain some general help from text-
books (where principles are often enmeshed in arithmetic), from the literature of public health and industrial medicine (Cochran, 1951, 1953; Cornfield, 1951; Doll, 1951; Dorn, 1953; Greenberg, 1953; Hansen and Hurwitz, 1951; Hill, 1951a, 1953; Sartwell and Merrell, 1952), and from a report by the American Public Health Association’s Committee on Sampling Techniques (1954). In the clinical field are introductory discussions by Asher (1954), Mainland and Herrera (1954a) and White (1953), whose article on sampling in medical research is very illuminating.

The present article, and one now in preparation on analysis of data, are intended to meet the need for more detailed guidance. They do not attempt to enumerate all the errors in the Cortisone Study (or rather, the features that would have been errors in a final survey); nor is it implied that all the errors mentioned were committed in that Study.

Nine Questions for Use in Planning

If treated with latitude, nine questions (Mainland, 1952, p. 12), found useful in criticizing reports, are helpful also in planning investigations. The questions, and their implications in case record surveys, are as follows:

1. Who?—Persons responsible for the original observations and for the survey.
2. Why?—The purposes of the survey.
3. What?—The population sampled.
4. Where?—Location and environment (e.g., of a clinic and its patients).
5. When?—Time factors.
7. How much?—Measurement. (In the broad sense of assessing results, this is covered by Question 6.)
8. How many?—Enumeration (numbers of patients and observations).
9. Why?—Why did this happen?—Causal relationships.

Looking at this list, a clinician might well say: “All I wish to do is take stock of my experience with a certain treatment up to date, and find the percentages of patients improved, unchanged and worse. Must I consider in detail nine questions before doing this simple arithmetic?” The best answer would be: “What, precisely, would the percentages tell you?” Before replying he could profitably read (in full or in extensive quotations—Mainland, 1952, p. 13) an article by a perceptive surgeon on “The Use of Experience” (Ogilvie, 1949), and he might then seriously doubt the value of any such percentages, laden with multifarious bias. The nine questions are designed to show:

1. How far such pessimism is justified,
2. How forethought can reduce biases, which can be defined here for reference as “things that make a sample different from what it purports to be”.

The notion that a survey, giving only “approximate” results, needs less care than a first-class experiment is erroneous. Particularly, it needs clear thinking throughout, and muddled thinking is common (Asher, 1954).

Persons Responsible for Observations and Survey

The effects of personalities on data and inferences are best discussed under Methods, but anyone who contemplates sharing in a survey should ask himself if he has enough time to do his part properly, and what his attitude will be.

Time Required.—There must be ample time for thought, for frequent personal consultation (not merely by letter or telephone) of all participants, and for detailed supervision of all assistants’ work.

Time is hard to estimate, but it may be mentioned that in this Department, when a clinical investigator colleague estimates the time required by a statistician to study a problem (without any computation) the estimate must often be multiplied by 10.

Experience in the Cortisone Study and other surveys suggests that, because of clinicians’ numerous other duties, at least 2 years should be allowed for a definitive survey designed to answer no more than five questions from a hundred pre-existing case records in one clinic. Certainly no less time would be needed for a preliminary exploration in a proposed co-operative study involving half a dozen clinics situated within a few travelling hours of each other.

Attitude.—The investigator should ask such questions as the following:

(i) Shall I be impeded in research by my need, in clinical practice, to be definite, helpful, optimistic, anxious for positive results?
(ii) Do I wish to hunt for bias, though it may mean discarding the whole survey?
(iii) When there is equal doubt regarding the validity of various results, shall I be more likely to accept, and propose publishing, a result that agrees with my previous impressions, rather than one which disagrees with them?
(iv) Am I inclined to brush aside, as too complex, subtle, sophisticated, or academic, a bias-causing mechanism that I do not understand?
(v) Am I inclined to react likewise to methods of analysis that I do not understand?
(vi) Shall I, a laboratory worker or statistician, fail to appreciate the difficulties of clinical research?*

* The New York University Department of Medical Statistics is peculiarly fortunate in not having to depend solely on recollected clinical experiences for its awareness of difficulties in research on patients. It meets the difficulties directly in its own x-ray investigations.
USE OF CASE RECORDS IN STUDY OF CHRONIC DISEASE

Purposes of the Survey

Two general purposes of case record surveys should be kept distinct:

(1) To increase experience regarding individual patients. Although called "anecdotal", this method of learning is not negligible. The reports of such surveys should, however, omit percentage frequencies, averages, and other estimates.

(2) To evaluate experience—the so-called "statistical" method—the subject of this article.

Before beginning an evaluational survey the investigator should express its purpose quite specifically. For every question that he intends to put to the data he should prepare a skeleton Table or statement, for example:

<table>
<thead>
<tr>
<th>Result</th>
<th>Number of Cortisone-Treated Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved</td>
<td>?</td>
</tr>
<tr>
<td>Unchanged</td>
<td>?</td>
</tr>
<tr>
<td>Worse</td>
<td>?</td>
</tr>
<tr>
<td>Total</td>
<td>?</td>
</tr>
</tbody>
</table>

As planning proceeds, each term in the Table will have to be defined precisely; the Table itself may be modified, and it will probably be multiplied to show results by different criteria, such as Grade and Class.

It will be noted that the Table avoids the post hoc propter hoc fallacy committed in the Cortisone Study data sheet, which asked for "Results of Treatment", and in the official titles of the A.R.A. Grades ("Grades of Response to Therapy"). Even members of the Cortisone Study Committee caught themselves forgetting to allow for concomitant therapy, undetectable factors, and the natural fluctuations in the disease.

The investigator is not an archivist, but intends to use the information obtained—to be able to say, for example: "If I have a patient like those in the Table, and treat him in the same way, his chances of being in a better condition at the end of 12 months are about 30 per cent."; or again, if a co-operative survey revealed that only about 10 per cent. of moderately severe arthritics improved under cortisone, this would strengthen the rheumatologists' efforts to find something better. Therefore, since his purpose is practical, the investigator should at the outset, and throughout the planning, ask: "Shall I have enough patients to make my estimate of practical value?" This question is discussed later under Confidence Limits.

Definition of the Population

Any therapeutic study, whether by experiment or survey, is based on an assumption which has made medical practice possible, and has thereby been justified—the assumption that future patients will be sufficiently like the observed patients to make the results of the study applicable. That is, the sample is assumed to be representative of a fairly constant population. The investigator cannot define the population fully, but he should define it and its segments (subpopulations) as fully as he can. Otherwise, his estimates will mislead him and may perplex him by differing from other observers' estimates. He should note carefully what the definitions imply regarding the risk of bias in generalization.

Definition by Selection of Patients.—The Cortisone Study was conducted on the patients of rheumatic disease specialists and, with few exceptions, on patients affected only in the peripheral joints.

The samples would rarely include patients in whom cortisone, prescribed previously by other physicians, had been manifestly useless or very toxic, because after such a history a physician at one of the clinics in the Study would rarely try cortisone again. Therefore, as compared with the general population of rheumatoid arthritics, the patients would presumably possess three characteristics:

(1) A higher ratio of potential successes for cortisone treatment.

(2) A higher proportion of cases treated soon after onset, because cases of longer duration would have had a longer time during which they could have demonstrated the failure of cortisone, and thus have excluded themselves from the Study (6 months' minimum period, discussed below, would counteract this tendency to an undeterminable extent).

(3) A higher proportion of cases treated early in the period covered by the Study (1950-1952), just after cortisone had become available.

Very important was the selection introduced by two stipulations:

(1) Presence of arthritis for at least 6 months before cortisone administration—an effort to reduce the number of mistaken diagnoses, but tending to exclude successful cases.

Many of these would be mild cases, and even if they had been included their records should have been examined separately, because many of them go into remission, even without treatment. The automatic exclusion of cases of very recent onset, however, made it impossible to answer the question: "What happens to a severe case treated by cortisone within 6 months of onset?" Moreover, 6 months is an arbitrary dividing line.

(2) Observation for at least one year after the start of cortisone therapy—desirable to show how many such cases were available, but excluding an unknown number who had stopped attending a
ANNALS OF THE RHEUMATIC DISEASES

clinic because they felt much better or because they were dissatisfied.

Any restriction can be placed on the class of patients surveyed, provided that:

1. it is not ambiguous,
2. it will give useful results,
3. the results are not extrapolated beyond the subpopulation so defined.

In a final survey the restrictions imposed in the Study would impair the usefulness of the results by risk of bias—if, for instance, at a different time or place the factors responsible for loss of patients were different, or if a different proportion of severe cases were treated in the first 6 months. To avoid this, all cortisone-treated cases should be included, but different categories (possible subpopulations) could be studied separately (e.g. patients treated within a few weeks of first symptoms; diagnosis doubtful; diagnosis reversed during treatment).*

Definition by Patients' Characteristics.—Severity of disease at start of treatment is the most obvious basis for subdivision of the original population, and it is here that rheumatoid arthritis presents its major difficulties: difficulty of accurate diagnosis; lack of uniform diagnostic standards; and lack of reliable criteria of status and progress. Until better methods have been discovered, the patient's history, inaccurate though it may be, must play an important part in the physician's decision regarding the episodic or sustained character of the disease (Ragan, 1955).

Other subdivisions of the population, as by sex and age, will be discussed in the article on analysis. It should be noted here, however, that such subdivision is justified simply by the knowledge, or likelihood, that separate estimates are more applicable to their respective subpopulations than are pooled estimates. It does not imply profound causal relationship. Thus, females in a certain age group may, at a certain clinic, do better under treatment than females of other ages or males of the same age, simply because of a local socio-economic or occupational difference.

Definition by Therapy.—In the Cortisone Study the population comprised “cortisone-treated patients” with certain specifications (e.g. that cortisone must have been the main therapeutic agent, that a minimum daily dose of 15 mg. should have been continuous, with certain exceptions, including reduction or discontinuance of dosage for observation). These stipulations would be too arbitrary and complex for a final survey, in which records of all cortisone-treated patients (or all those in a well-defined class, such as adults) should be collected. Then, if division according to minimal dosage or continuity of treatment were desired, the rules should be made suitable to cover data from all the clinics.

Therapy of a chronic disease is, quite rightly, multiple, and assignment of rank may be very difficult. In rheumatoid arthritis, for example, when does aspirin (taken with or without the physician's knowledge) cease to be an occasional pain-killer and become an anti-rheumatic treatment concomitant with cortisone?

When the administration of a drug is followed by improvement and no serious side-effects it will usually be continued, and the drug will be called the main (or sole) agent. Again, if there is no doubt about the physician's intention to continue the drug as the sole agent unless he is induced or compelled to stop it for such reasons as toxicity, inadequate benefit, or loss of the patient, such patients would clearly qualify for the survey.

Many cases, however, are not clear cut. A patient is started on Drug A, but receives from time to time another drug, B, either as part of the routine or because he seems to need additional help. At a certain time he appears to benefit greatly from Drug B; therefore its dose may be increased, while Drug A is cut in dosage or even stopped, temporarily or permanently. The process may be reversed, according to the patient's response; then a third drug, C, perhaps newly on the market, may be tried.

To eliminate from the survey such ambiguous cases, or all cases that received more than one drug, would give obviously biased estimates of that drug's efficacy. Dosage level gives little help in selecting the main therapy, because x mg. of Drug A daily for 2 months, without Drug B, is not equivalent to the same amount along with Drug B.

In the Cortisone Study no questions were asked regarding the grounds for stating that cortisone was the main therapeutic agent, and this may have been, in general, a safe procedure. However, one physician, who administered multiple therapy, stated that he had submitted records only of patients in whom he was convinced that the therapeutic response was due to cortisone, and this raised two questions that would have to be answered before his data could be used in a final survey:

1. Would the record of a patient, receiving the same therapy, have been submitted if he or she had not shown a response?
2. Would the lack of response have prompted the use of another drug, and thus have automatically excluded the patient's record from the survey?

Many patients in the Study had received concomitantly drugs other than cortisone, and, not infrequently, benefit was recorded. Therefore some of the responses, recorded at 3-month intervals, must have been due partly to those drugs. Even full information on times and doses would have permitted only partial, and somewhat speculative, separation of these effects from the cortisone effects.

* If a diagnosis is reversed at autopsy, the clinical diagnosis should be retained, for treatment is not based on a post-mortem diagnosis. (Densen and others, 1952).
USE OF CASE RECORDS IN STUDY OF CHRONIC DISEASE

The Study disregarded physiotherapy because most patients receive it, much or little. A final survey could not neglect it, because of its effects on functional capacity; and no therapy of a psychological kind can be overlooked, even in a survey of a predominantly physical disorder.

To cope with the complexity of therapy, perhaps the best skeleton Table would have heading and side-heads such as the following:

<table>
<thead>
<tr>
<th>PATIENTS STARTED ON CORTISONE—CONDITION AT 12 MONTHS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Treated by cortisone alone, or with concomitant treatment which had been administered, not because of inadequacy of cortisone, but as part of a routine, or for reasons other than the arthritis.</td>
</tr>
<tr>
<td>(2) Received concomitant treatment because cortisone was considered insufficient.</td>
</tr>
<tr>
<td>(3) Stopped cortisone because of inadequate benefit.</td>
</tr>
<tr>
<td>(4) Stopped cortisone because of toxicity.</td>
</tr>
<tr>
<td>(5) Lost because of progress of the arthritis.</td>
</tr>
<tr>
<td>(6) Lost from other causes and from unknown causes.</td>
</tr>
</tbody>
</table>

The body of the Table would contain numbers of patients for each side-head, and these could be expressed as percentages of the total. Group 5 would contain patients who died from the disease, and also other patients, such as those who became bedridden and unable to attend the clinic. Other losses (Group 6) will be discussed more fully later. This group might be subdivided in various ways, but it should contain those who went into remission and were lost before the end of 12 months.

Only in Group 1 would it be wise to record the status of patients, including those who had stopped cortisone because of remission, but were still under observation at 12 months. Group 2 is a lesser degree of Group 3, and to show the status of its patients would offer too great temptation for illicit comparison with Group 1.

In the final analysis the Grades, or other subdivisions, in Group 1 would each show a percentage of total patients, and the percentages in Groups 2 to 5 could, if desired, be pooled to show the proportion in whom cortisone had failed, partly or wholly.

Unless it is possible, objectively and without ambiguity or risk of bias, to classify patients as mainly on one treatment from the beginning, such a heading as “Cortisone-Treated” has little meaning. A clinician could more logically say:

In my clinic cortisone is used largely, but not exclusively or on all patients. What is the condition of all patients 12 months after treatment is started, each patient having been treated as appeared best at the time?*

* Implicitly included here as part of the treatment is the physician himself.

Methods of Assessing Patients after Treatment

Diseases differ greatly in method of assessing the post-treatment condition. Even a disease like cancer presents problems (Dorn, 1951; Hopkins, 1953; Lees and Lees, 1950; Park and Lees, 1951), and the difficulties in rheumatoid arthritis, already referred to, are in many respects greater. The American Rheumatism Association scheme (Steinbrocker and others, 1949) promotes a standardization and objectivity that were much needed (Steinbrocker, 1946), but it still allows too great opportunity for inter-observer differences, and it is undergoing revision.

There may be surprisingly large differences between clinicians' findings in commonplace observations, such as interrogation of persons regarding the history of dyspepsia, presence of cough, and occurrence of sputum (Cochrane and others, 1951; Fletcher, 1951). In chest-film assessments, inter- and intra-observer variations have been studied extensively (Birkelo and others, 1947; Fletcher, 1951; Garland, 1950), but less attention has been paid to other regions.

If, for example, a radiologist is assessing films of rheumatoid joints, what effect has his knowledge of the treatment (cortisone or not cortisone), and the temporal sequence of the films?

It might be replied that the particular radiologist has no stake in the effects of cortisone one way or the other, and that films out of order would cause mental confusion. This answer would hardly satisfy anyone who, starting with a perfectly open mind, has felt bias creeping on him in a series of histological assessments. In any inspectional assessment, even the effort to avoid bias can create bias, and blindfold assessment is the only safeguard. Strictly random order, where possible (e.g. in a set of films or slides), controls also the bias due to a “drift” or “trend” in the observer, which is commoner, even in measurements, than medical observers apparently realize.*

Blindfold clinical assessment is possible to only a very limited degree (e.g. by arrangement with colleagues who are ignorant of the treatment). Hence the value of such devices as rings to measure the swelling of finger joints.

Methods of Collecting Data

Case Records.—Hospital records are known to contain undependable clinical data and grossly misleading laboratory reports (Belk and Sunderman, 1947), but chiefs of service are perhaps less aware than their juniors of the risk of purely fictitious benefits.

* Randomization methods suitable to this and other problems are described by Mainland (1952, p. 267), but the method shown there for allocating two treatments to equal numbers of subjects can introduce bias, as can coin tossing. For a simple unbiased method see Herrera (1955).
entries (Mainland, 1952, p. 240). These errors may, indeed, not affect a patient’s welfare greatly, for a clinician who knows his patient well needs few notes. Most disconcerting, however, is the fact that, although they know how un dependable individual records may be, many clinicians are very ready to draw conclusions from a table of percentages derived from a mass of such records. However large the mass, systematic errors (all in the same direction) do not “average out”, and large random errors (of the ± type) cause serious lack of precision.

Clinical training and the pressure of routine work do not make for reliable record keeping, and experience in making clinical records, as well as in trying to use them, has led the author to believe that the only records trustworthy for research are:

1. those kept meticulously by a physician regarding his own patients because he wishes to do so,
2. those kept regarding a particular group of patients by a suitable and adequately instructed person, specially assigned to the task.

Patterson (1954) rightly regrets the appearance of a tendency to place “too much emphasis on the prospect of research work being made possible by elaborate systems of record keeping which would enable clinicians to get all sorts of statistical material on request”. As he says, “the fundamental duty of the records officer is to produce the record when it is wanted, the whole record, and at once”.

**Planning of Record Forms.**—If records are desired for anything more than their primary purpose (help in treating the patient), thorough planning is necessary. In such planning the first two requirements should not be: How can we avoid omitting data that may be wanted in some future research? and How should the inquiries be formulated for transference to punch cards?

A proposed routine record scheme for skin cancer patients covered four types of study:

1. aetiological, e.g. family history of tumours, parents’ eye colour, patient’s exposure to light;
2. therapeutic, e.g. cure rate, relation to dosage, five-year follow-up;
3. socio-psychological and economic, i.e. reasons for delay in obtaining treatment;
4. anatomico-pathological, i.e. location of various types of tumour.

Adequate exploration of all these topics would require a vast multitude of questions. For example, the question on the family history of tumours, showing four possible answers (Yes, No, Questionable, Name of Organ) should be either deleted or expanded until it created a family tree.

Regarding punch cards, it should be remembered that questions on any subject, real or imaginary, can be adapted to a card-sorting machine, and that it is not the function of a tabulating or business machine company to guard its customers against wrong questions or false data.

Hill (1951a) has declared that, in planning a questionnaire, for every question that a planner proposes to ask a respondent, he should ask himself at least three questions. A similar rule in the planning of a record sheet produces three sets of preliminary questions (on topics, on responses, and on the use of the answers), with three questions in each set.

**Questions on Topics for Survey**

1. What information do we desire soon?
   10 years hence, a skin cancer patient’s excretion of some urinary constituent, now unknown, may be of far greater interest than his family tree.

2. Of the topics that interest us, which can be profitably studied from routine case records?
   The most suitable topics are frequencies or averages, such as proportions of different diagnoses made in one clinic in a year; relative frequency of skin tumours in different parts of the body; percentage of patients with a certain disease improved after 12 months’ treatment.
   (For other types of question, see Causal Relationships.)

3. Among the topics amenable to case record study, what difficulties were met in the best work already done, and how far can we overcome them with our facilities?

**Questions on Responses.**—The proposer of any question for the record sheet should ask himself:

1. What are all the possible answers to this question, and are they shown alongside the question?

The Cortisone Study sheets (Figure, p. 327) purposely did not indicate all possible answers. Under “Subcutaneous Nodules”, for instance, they omitted the possibility that new nodules might appear, and that some nodules might enlarge while others diminished. In a final survey this would not be allowable, for it might lead to bias in various ways.

For example, one careful observer omitted statements on nodules regarding patients in whom, as subsequent inquiry revealed, some nodules had enlarged and others had decreased. Other observers, doubtless in the same predicament, had presumably chosen one or other of the printed answers, and the grounds of choice could not be ascertained.

Answers headed “Other” and instructions to “Amplify on reverse side” conduct greatly to bias, because responses vary from nothing to abundant detail. They are valuable, however, for anecdotal purposes and in a preliminary exploration to reveal the quality and difficulties of respondents.
USE OF CASE RECORDS IN STUDY OF CHRONIC DISEASE

(2) What are the various possible interpretations of the question by the person who has to answer it—patient, physician, technician, or other respondent?

In some instances no uniform interpretation is possible. For example, in 51 per cent. of the first 446 data sheets received in the Cortisone Study the "Year of Onset of Disease" was stated to be the same as the "Year of Onset of This Attack", sometimes many years before the start of the therapy under investigation. The distinction seems to be inapplicable in rheumatoid arthritis.

(3) What factors, other than misinterpretation of the question, may cause the answer to be untrue, or elicit the response "Don't know"? Besides gross errors in records, already mentioned, the following three factors should be considered:

(a) Lack of remembrance or of knowledge (which may be registered as a biased guess).

Patients' power to recall events and dates is much less than many physicians seem to believe, and independent checks are seldom possible. Some questions a conscientious respondent would find very difficult; for example, in a rehabilitation follow-up questionnaire: "How many hours a day do you stand?"

(b) Ulterior motives, such as hope of gain or desire to please.

Even the final verdicts on the outcome of treatment, on case records, or on certificates from certain sources, are greatly influenced by various motives. Therefore anyone, including an unsuspecting statistician, who plans to use data from large organizations, governmental or other, should, indirectly and discreetly, acquire an insight into the factors that may have been operating when the statements were made. (This would save some expensive retrospective and follow-up studies.)

(c) Emotional factors.

When fearful of a particular diagnosis or bewildered when it is made, no patient should be expected to answer objectively. Even in healthy people the questionnaires themselves may produce symptoms (British Medical Journal, 1954; Glaser, 1954)—apart from the anger which would be the most appropriate response to some lists of questions.

At an early stage, proposers of questions should test them on colleagues, non-medical acquaintances, and patients, including the less intelligent and less co-operative. Because of the unreliability of patient response, no conclusions were drawn from such information in the Cortisone Study.

Questions on the Use of Answers.—Visualizing the use of each answer, the proposer should ask:

(1) In what ways may the user interpret the answer? Will he know what it really means?

(2) If answers to some of the questions are to be combined, how is this to be done, and what will the composite answer tell one? For example, if an index of capacity for activities of daily life changes or remains the same in an individual, what inferences can be drawn?

(3) Exactly how is the answer to be used?—see skeleton Tables discussed under Purposes of the Survey. The proposer should prepare such Tables even if he does not expect to use the answers himself, for then he will be forced to ask what must be done with doubtful or ambiguous answers, blanks, and descriptive answers.

Mail Questionnaires.—The foregoing recommendations apply also to mail questionnaires, but here the questions should be very few. Hill (1951a, 1951c, 1953) half-facetiously suggested a maximum of five, but showed how, without exceeding that number per person, answers to more than five questions could be obtained.

Extracting Information from Records

When record sheets are properly planned for a particular survey and data are properly entered, extraction from the records consists essentially in copying. The Cortisone Study illustrated a much commoner set of circumstances. The interest in cortisone had resulted in more detailed case records than are usual in many diseases, but many of the questions asked in the Study were not directly answered by the original records.

Preliminary Study.—In such circumstances the preliminary study should follow the lines recommended for the planning of record sheets, but should also explore specimen records fully, especially the more difficult ones, and from all clinics if a co-operative study is projected, the purpose being:

(1) to test the proposed questions,
(2) to consider the usefulness of the data (apparent reliability, ambiguities, blanks),
(3) to formulate instructions for transcribers.

Instructions should take nothing for granted. The Cortisone Study revealed that in a pilot or final study they should state: the equivalence of ACTH and corticotrophin, trade names for gold preparations, the distinction between intercurrent and concomitant therapy, and the intention to include oral hydrocortisone (oral compound F) in the survey.

The analyst in a survey, whether clinician or statistician, should never need to seek interpretation of an answer in one section by referring to another section, perhaps to a third, and then to remarks volunteered on the reverse side.

Members of the Cortisone Study planning committee remarked that if many instructions were sent, the transcribers would not read them, and even the few instructions that were issued were often ignored. The analyst had to make many calls for elucidation, and her fear of undetectable errors grew ever stronger. If a clinic has
no transcriber with the ability, time, and willingness to follow necessary instructions minutely, its contribution to a survey will be useless.

If, however, more than a few instructions for the interpretation of the record sheet data are found necessary, this should raise the fear of bias through arbitrary interpretation of records that are unsuitable for anything except an anecdotal survey.

Risks of Interpretation.—The dangers attendant on using records not designed for the particular survey were present in two areas of the Cortisone Study that would have been vital in a final survey. The transcribers, not necessarily acquainted with the patients, were required:

1. to translate recorded data into the Stage, Class, and Grade of the American Rheumatism Association scheme,
2. to state the reason for permanent cessation of cortisone treatment, which might or might not entail an inference, not merely a transcription, from the original records.

Extensive checks would have been necessary to show the full effects, but it was noted that out of 28 patients whose cortisone was stated to have been stopped because of remission, eight were stated to have shown a Grade II improvement, whereas by definition Grade I = remission.

When trying to infer the reason for stopping cortisone, the transcriber would find in the records evidence that the patient was not doing well, or evidence of toxicity, or both. If both, he would not necessarily give to each factor the weight given by the physician who had stopped the drug; and other factors, such as financial circumstances, might not be recorded at all. Discrepancies between answers regarding the same patient reflected the difficulty.

When a physician has extracted and interpreted the information from his own records of private patients, the accuracy of interpretation will be greater, but even then the question could be asked: “Would he have given the same verdict (Class and Grade) for the 6-months observation if he had not known what was to happen later?”

One cannot be certain, either, that he would in all cases recall the degree of importance that he had given to the various factors when stopping cortisone. Indeed, even at the time of stopping it he would have found the weighting difficult in some patients.

It seems doubtful, therefore, if even a perfect survey could give a simple answer to such a question as: “In patients whose cortisone was not stopped because of toxicity, what was the condition at 12 months?”

A more reliable statement would be obtained by pooling data from all patients whose cortisone had been stopped because of toxicity, inadequate benefit, or both.

Pilot Study.—If a preliminary study does not show that the survey should be dropped or made purely anecdotal, a pilot study may be started, preferably on random samples of the records, in order to provide estimates of sample size required in the final study. The general technique is not difficult (Mainland, 1952, p. 271), but in applying it the help of a statistician well versed in sampling is usually desirable. To take every second, fifth, or tenth record is likely to give a false estimate of variability.

In any pilot or final survey, numerous cross-checks between transcribers are necessary to avoid serious differences in interpretations (Densen and others, 1952).

Time Factors

Essentially all the problems of chronic disease epidemiology revolve around the time factor, since every characteristic by which a group of people is commonly classified, except their sex and race, can change with time. (Sartwell and Merrell, 1952).

An investigator of a chronic disease should, during the planning, analysis, and presentation of results, ask such questions as:

1. Will this estimate (e.g. percentage of patients improved, average erythrocyte sedimentation rate, incidence of toxicity) apply equally to adults and children, young adults, the middle-aged, and the aged? If not, is the difference important in our use of the estimate?
2. Is it safe to pool several years’ experience of this therapy? (Types of patient selected, dosage schemes, and the initial optimism regarding a drug can change greatly in the first 5 years after its introduction.)
3. What allowances should be made for the duration of treatment, observation, and exposure to risk of disease, recurrence, or toxicity?

In the preliminary report on the Cortisone Study (American Rheumatism Association, Panel Discussion, 1953), in order to give at short notice a picture of the material collected, some percentages from the pooled data were presented without any allowance for the time factor, and the subsequent general discussion, with audience participation, showed that this was a mistake.

Estimates from such pooled data, even from a single clinic, would be applicable only to a population that had the same composition, in duration of treatment and observation, as the group studied—so many patients treated for 12 months, so many for 13 months, and so on. The patients actually studied would be a different population a month later. To obtain an estimate with useful meaning, one must show what has happened to the patients at some fixed time after starting treatment—the time being the same for all patients. (Comparison of patients treated in the early part of the survey...
period with those treated later will be discussed in the article on analysis.)

Another effect of the time factor was pointed out by Kellgren (1953) in the discussion of the preliminary report of the Cortisone Study. In a disease characterized by remissions and exacerbations, treatment does not begin in remission. Therefore, even if a therapy has no effect at all, at any given point in time after it has started, some of the patients should be expected to be in remission.

Numbers of Patients and of Observations

A primary question in any investigation is: "Have we enough material to provide useful estimates?"

In a clinical record survey the answer involves:

1. Losses from the sample;
2. Fictitious sample sizes;
3. Sample size and confidence limits.

(1) Losses from the Sample.—These include disappearance of patients, interruptions of treatment, losses of records, blanks in records, and failures to come for examination or to answer questionnaires. If the losses were strictly random they would merely reduce precision; but the only safe assumption, usually, is that the losses are not random.

Patients whose treatment has been stopped because of remission tend to disappear, and so do patients discouraged by lack of benefit. Blanks in records may indicate mere accidental failure to record a test or treatment, or a belief that such a test or treatment was unimportant, or lack of interest in the patient, which itself can affect the outcome. Specially "interesting" case records or X-ray films, removed from the series for study or teaching, may also be lost.

In a mining population of 600, all of whom were voluntarily X-rayed, those with pneumoconiosis mostly came up first, and those with tuberculosis came last (Cochrane in discussion of a report by Stewart, 1951). Lack of insistence on a full survey would have caused bias.

In questionnaires it should be assumed that non-respondents' answers would differ, to a greater or less extent, from the answers received, unless this can be disproved experimentally—a difficult task, for it involves measures to insure that answers are received from a strictly random, and sufficiently large, sample of the non-respondents. It was this danger of bias that led Hill (1951a) to declare:

I would therefore myself infinitely sooner have, say, a one in four [random] sample of the population, of a size thereby which enabled me to pursue relentlessly, and complete the records for, all or nearly all the persons in it, than have to interpret figures derived from survey of the "whole" population from which finally a quarter was missing.

Treatment of Losses.—Occasionally, blanks in the records can be safely filled in by interpolation, and very occasionally, after much critical thinking, by extrapolation.

Where losses in a therapeutic survey are due to death, accident, intercurrent disease, or change of residence, these patients can be excluded from the population, provided that the events are entirely independent of the disease and therapy. In rheumatoid arthritis pregnancy should also exclude patients, because in many women it has a beneficial effect.

In reporting a survey, three basic rules should be:

1. State how any exclusions modify the definition of the population.
2. Account for all losses, at whatever stage they occurred.
3. Show the effects of the various possible allocation of losses, as exemplified in the following paragraph.

Tolerance Limits.—The tolerance limit of about 5 per cent. (Doll, 1951) for loss in questionnaires (except when purely anecdotal) is applicable to other surveys. Investigators should, however, see for themselves in the preliminary study what the amount and effects of losses and blanks may be. Thus, fifty records in a certain group (e.g. females over age 40, initial condition: Stage II), classified according to condition (A or B) after 12 months' treatment, might show:


The possible estimates are:

From available information on 45 patients.—

A: 22 per cent.; B: 78 per cent.

If losses or blanks were As.—

A: 30 per cent.; B: 70 per cent.

If losses or blanks were Bs.—

A: 20 per cent.; B: 80 per cent.

If, instead of A and B, the estimates were average measurements (e.g. E.S.R.) the two suppositions would be:

1. that all losses or blanks contained low readings,
2. that all losses or blanks contained high readings.

Obviously, losses of fewer than 5 per cent. of the total patients in a survey may render useless the data in some sparsely peopled categories.

(2) Fictitious Sample Sizes.—Samples can be rendered fictitiously large in several ways, such as:

1. inclusion of supplementary cases,
(2) spurious replication,
(3) pooling of heterogeneous data, as in co-operative surveys.

Supplementary Cases.—Except for anecdotal purposes, it is seldom useful, always troublesome, and often misleading, to include a few subjects outside the main class under study. Five children would have to differ from adults enormously in response to treatment before they could differ "significantly" in the statistical sense, but to apply to children the estimates from adults (or from adults plus children) could be grossly misleading.

Spurious Replication.—In a study of anaesthetics in children it was proposed to use 343 administrations as denominator in estimating the percentage frequencies of various reactions, but only 148 children had been observed (one to fourteen administrations per child). The only safe assumptions, however, were that:

(1) a child would often tend to react in a similar way to the same anaesthetic on different occasions,
(2) a first administration might influence the reaction to later ones.

The "sampling unit" should have been the first administration in each child (although further study of multiple administrations could, of course, be made). Spurious replication ("counting the same thing over again") is not uncommon (Mainland, 1952, p. 32; Mainland and Herrera, 1954b).

Heterogeneous Data in Co-operative Surveys.—The main reason for a co-operative survey is to increase sample size. Such a survey usually requires expert help in planning and analysis; but as soon as it is thought of, one general outcome should be anticipated. If, after it is completed, averages from pooled data are presented, some of the investigators will accept them, but others will probably dispute them because the estimates do not agree with their own experience, and this is the safer attitude.

Even if the survey were perfect in all respects, and if groups of patients, one group from each clinic, could be matched according to all the categories in the records (sex, age, initial severity, chief and concomitant therapy, and other relevant features), there would still remain unrecorded but relevant inter-clinic differences in the environment and personalities of patients, diagnostic methods, therapy, and judgment and personality of physicians and other attendants. Major factors would be the physicians' differences in the selection of patients, on the basis of history, for the therapy under study, and their interpretation of standard criteria of status and progress.

Therefore, 222 rheumatoid arthritics from 23 clinics, in Stage III at the start of therapy, are not to be considered as a single sample but as 23 samples, each from a different population. The estimates (e.g. percentage improved or toxic after 12 months' treatment) will perhaps not differ between clinics more than is often attributed to chance, i.e. they may not be significantly different at the 5 per cent. level, and that will be discussed in the article on analysis, but it is generally safest to assume that the observed differences represent actual differences (heterogeneity).

Hence, even without further division into sub-populations, the above figures would provide 23 estimates with differing sample sizes, and the pooling of the data could take one of two forms:

(1) A Weighted Estimate.—If fifty of the total 222 patients showed improvement, the weighted estimate would be $50 \times \frac{100}{222} = 22.5$ per cent. This gives greatest weight to the clinics that contributed most patients, but the size ratio of the 23 samples does not represent anything general, such as a nation-wide ratio of clinic population types. The percentage represents an estimate of the value that would be approached if one could take larger and larger samples of the same types of patients from the same clinics and in the same ratio of sample sizes—not a very helpful piece of information.

(2) An Unweighted Estimate.—This would be obtained by finding for each clinic the percentage improved, totalling the 23 percentages, and dividing by 23. It is an estimate of the percentage that would be approached if one could obtain larger and larger samples, all equal in size, from the same clinics, and pool them (or, of course, find the average of their percentages).

If it could be assumed that the 23 clinics formed a random sample of a population composed of all rheumatic disease clinics (each clinic being a unit), the unweighted average would be an estimate of the average for this population. During the period covered by the Study, the 29 clinics that furnished data probably comprised the majority of the clinics in North America, but the 29 were not a strictly random sample; and still less were they a random sample of the much larger number of clinics now in existence.

The safest and most reliable treatment is to use with each clinic its own estimate. This seems to throw doubt on the usefulness of co-operative studies; but with reliable data they could be very useful, for they would enable one to compare, contrast and perhaps combine information, obtained from the data of each clinic separately, on trends of improvement or regression of patients.

Elaborate methods were not justified in the Cortisone Study, and unweighted averages, without estimates of error, were used in order to give a rough picture of the change in patients' status during the 2 years.
USE OF CASE RECORDS IN STUDY OF CHRONIC DISEASE

(3) Sample Size and Confidence Limits.—Even with no losses or blanks, a sample may be too small to give a useful estimate, because possible variation due to chance must always be allowed for. Thus, if of fifty adult female cortisone-treated arthritics, after 12 months' treatment, fifteen (30 per cent.) showed improvement and 35 (70 per cent.) showed retrogression or no improvement, the true (population) percentage of improved cases in the same type of patients, treated in the same way, might lie anywhere between 18 and 45 per cent, but would probably not lie outside this range. If an investigator uses these estimated limits, and corresponding estimates in other problems, he will be correct in 95 per cent. of such judgments regarding the effects of chance. That is, he has a 95 per cent. probability of being correct in any one judgment, and the limits are called "95 per cent. confidence limits".

Such estimates of error are necessary in all scientific work on either frequencies or measurements, but they are often omitted in medical research (Mainland, 1954c). For percentage frequencies in two-class (binomial) samples (e.g. improved, not improved) the limits can be easily found from Tables and Graphs (Mainland, 1952, pp. 53 and 311), and more detailed Tables are in preparation (Mainland, Herrera, and Sutcliffe). The commonly used approximation, derived from the standard error of the binomial, can be misleading in many samples met in clinical research.

Causal Relationships

Even a perfect survey could demonstrate only an association between phenomena, not a causal relationship, which a properly conducted experiment could prove with a specified probability. Sometimes the survey method has led to correct causal inferences of great practical value, the most famous example probably being Snow's tracing of a cholera epidemic to the Broad Street pump in London (Hill, 1953); but such successes are to be expected only if a survey approaches an experiment in design and execution, particularly with reference to controls (perhaps better called "contrasting samples") and to selection of samples. These two features can be considered in relation to four kinds of topic on which information is often desired from case records:

(1) Aetiological relationships or clues thereto, such as sex incidence, age incidence, and relationships to other diseases;
(2) Comparative therapeutic effects of different dosage schemes;
(3) Relationship between dosage and toxic effects;
(4) Interrelationship of phenomena within a disease (discussed later, under Fallacy due to Competing Selection Rates).

(1) Aetiological Relationships

Lack of suitable controls is the chief reason why case record surveys are seldom of use in studying the relationships of diseases to other phenomena, events, or attributes, including other diseases—for example, the relationship of skin cancer to eye colour, myocarditis to rheumatoid arthritis, and peptic ulcer to personality. It is hazardous to employ as controls patients with all other diseases or certain selected diseases in clinic or autopsy records. Even the most careful choice of controls does not eliminate the risk of a subtle bias (discussed under Fallacy due to Competing Selection Rates).

Much more easily avoided, but still frequently committed in medical writing, are blunders due to ambiguity in terms like "sex incidence" and "age incidence". It would, presumably, be unreasonable to doubt that rheumatoid arthritis, sufficiently severe to bring patients to physicians, is more frequent in women than in men; but figures such as 60 per cent. women and 40 per cent. men, obtained from rheumatoid arthritic patients, do not, by themselves, justify the inference at all. Underlying the inference are two assumptions:

(1) The existence of a control or standard of comparison—roughly equal numbers of men and women in the total population.
(2) The absence of selection factors, such as have affected dermatologists' records of the sex ratio in acne—the tendency of girls to be more concerned than boys about their appearance (Sulzberger, 1955).

As commonly used, "incidence" is equivalent to "percentage frequency" (or frequency per thousand, per hundred thousand, and so on), without reference to the epidemiologist's distinction between "incidence" and "prevalence". The figures 60 and 40 per cent. show the incidence of women and the incidence of men in a group of arthritics, not the incidence of arthritis in women or in men.

To obtain, for any particular place and time, the incidence (the epidemiologist's "prevalence") of a certain disease in women there would be needed, for the numerator, the number of women who had that disease (say, A). For the denominator there would be needed the same women (A) plus all the other women (B) who would have appeared in the numerator if they had had the disease; and likewise for the incidence in men.

The denominator (A + B) is the "population at risk", or, as sounds more appropriate in non-infectious diseases, the "source population"; and the source populations for case records are unknown, except in special groups such as the armed forces and some industrial companies.

The denominator used in the estimates, 60 per cent. women and 40 per cent. men, was the total
number of arthritics; this fault of "wrong denominator" has led to many invalid conclusions, such as an association between amenorrhea and sedentary work (Mainland, 1952, p. 33), and is very commonly found in relation to "age incidence".

somewhat fewer women who have not previously developed it).

In the left part of the Figure the slope might have been vastly different if, instead of concentrating on rheumatologists, the Cortisone Study had included paediatricians and general practitioners.

The dip in the central part of the Figure, however suggestive, and even if "statistically significant", tells nothing in the absence of the source populations.

The error of "wrong denominator" is committed twice over in such comparisons as the "age incidence" of gastric and duodenal ulcer criticized by Hogben (1945), where peaks occurred at different ages. It was quite unsafe to assume that the source populations and sample selection factors were the same for the two groups.

(2) Comparison of Dosage Schemes.—The amount and scheduling of a therapy are nearly always determined largely by the apparent needs of the patient throughout the treatment. Therefore the comparison of effects of different doses or schedules is argument in a circle—a fact that seems to need frequent reiteration (Hill, 1952).

The circle is less obvious, but still usually present,
in demonstrations of relationships between dosage and phenomena that were not used by the physician as indicators for dosage change. In rheumatoid arthritis such phenomena might be bone density or results of C-reactive protein tests. Unless the phenomena can be shown clearly to have no relationship to the dosage indicators, the demonstrations are fallacious.

(3) Dosage to Minimize Toxic Effects.—Although by no means comparable to an experiment, a survey of case records might give some useful information on dosage-toxicity relationships if the investigator could overcome certain difficulties, chiefly of definition:

(1) Definition of toxicity (side-effects).

(2) Period of observation—the same for all patients.

(3) Definition of dosage—size of each dose, intervals, total for a period.

In the Cortisone Study the doses were divided into less than, and more than, 50 mg. daily; but the lower dose is quite inadequately defined, for this class would include patients on very small and infrequent doses. The almost infinite variety of dosage procedure seen in the Cortisone Study illustrates the great difficulty of finding a satisfactory dividing line between "low" and "high" doses, even in an individual clinic.

(4) Exclusion of possible causes other than the treatment. To avoid difficulty and prejudice, it is probably best to count as actual or possible side-effects all events (of predefined types) that could have been due to the treatment.

(5) Pooling of all patients or division by severity of disease, concomitant therapy, and other relevant features. Subclasses may be too small for the necessary precision, but the whole group of patients may be too heterogeneous for useful estimates.

The conclusion of many attempts to establish a dose-toxicity relationship must be simply the obvious one that, usually, the less of a drug is given the lower is the risk of toxicity.

**Fallacy due to Competing Selection Rates**

(*Berkson’s Fallacy*)

At one time surgeons were prompted to remove the gall bladders from diabetic patients because of an unusually high frequency of cholecystitis in diabetics, and this inference seemed to be confirmed by using, as a control sample, persons who came to a clinic for eye testing. Berkson (1946) showed, however, that such an association, demonstrated by comparing in hospital patients the incidence of a Disease X in samples of two other diseases, A and B, can be entirely spurious, in the sense that it does not reflect an actual association outside the hospital sample. It can be due solely to the fact that different diseases lead to different percentages of diseased persons admitted to hospital.

**Mode of Operation.**—The way in which the bias is created is easily demonstrated by simple arithmetic (Mainland, 1953; Mainland and Herrera, 1954a), and the outcome of one such demonstration (one mode of operation of the bias) can be described thus:

In the population outside the hospital the percentage frequency of persons with Disease X was exactly the same in persons with Disease A and with Disease B. Disease A had a higher admission rate than Disease B, and in the hospital samples there was a higher percentage frequency of Xs among the Bs than among the As. In such cases the higher rate in the As seems to push the Xs into the Bs; and, to extend the terminology beyond hospital examples, one can speak of “competition between selection rates”. The difference in percentage frequencies, it should be noted, can be very large and highly significant in the statistical sense.

**Scope.**—Unless there is strictly random sampling from the population regarding which an inference is to be made, the bias, though unsuspected and often undetectable, can cause fallacious inferences from a survey of any phenomena by creating a spurious association, masking a real one, or reversing it. Thus it can throw doubt on inferences from autopsy records (Mainland, 1953), attempts to relate lung cancer and smoking (Berkson, 1955), studies of blood groups in disease, and physiological and anatomical surveys (Mainland, 1955). It can occur, of course, where more than two samples, and more than two classes of individuals, are analysed. It can affect measurements as well as frequencies, but this field has not been explored.

**Effect on Relationship within a Disease.**—The fallacy could vitiate any inferences regarding interrelationships of phenomena in a disease, either a status (when the patient is first seen) or a change in phenomena during therapy—for example (in rheumatoid arthritis), x-ray assessment, C-reactive protein, erythrocyte sedimentation rate, subcutaneous nodules, numbers of joints involved, functional activity, improvement, or resistance to therapy.

As a hypothetical example, let it be supposed that:
There is no association between the presence of subcutaneous nodules (X) and resistance of the disease to therapy;

(2) Patients whose disease is more resistant to treatment (As) are more frequently referred to a certain specialist than are patients with less resistance (Bs);

(3) Patients with abundant nodules have a greater tendency to seek the specialist's help than those with few nodules or none.

The scheme described above, using the same letters (A, B, and X), shows that the specialist will find in his sample of arthritics a greater frequency of patients with nodules in the less resistant cases than in the more resistant cases.

By contrast, let it now be supposed that patients with subcutaneous nodules are, in fact, commoner among those who are resistant to therapy. It can be shown, again by simple arithmetic, that this factor can neutralize competition between selection rates, so that the specialist will find approximately equal percentage frequencies of nodule-bearing subjects among the more resistant and less resistant patients. Or, if the association between nodules and resistance is very strong, he will find evidence of it in his patients, but diluted by the competition between selection rates.

Suggestions for Counteracting the Fallacy.—The implications of Berkson's fallacy raise the question: "Have non-experimental observations on patients any more than anecdotal value?" For clinical purposes they certainly have, when properly made. If in a certain clinic an association has been found between a feature, present in patients before treatment, and their improvement or lack of improvement under treatment, that association will be useful in prediction and as a guide to therapy, even although it has arisen through competition of selection factors—providing, of course, that the factors remain stable.

Regarding deeper exploration of relationships between and within diseases, suggestions such as the following can, so far, only be general:

(1) Investigators, including experimenters and practical statisticians, should explore the various ways in which the bias can occur, using simple arithmetical examples at first, until they come to think of it automatically in any problem.

(2) In each particular problem they should try to think of all possible ways in which the bias might occur, and see whether at least some of the possibilities cannot be excluded. For example, if a scheme involved the assumption that in a specialist's practice mild cases were more frequent than severe cases, the scheme could usually be ruled out. (Even this example partly breaks down when a specialist treats many mild cases among his medical colleagues.)

(3) When a survey has revealed an association between phenomena (or lack of an association that was rationally anticipated), repetition of exactly the same kind of exploration in other places, though useful, is not enough, because the same selection factors may be at work. Where possible, it is better to explore, in circumstances where the selection factors are likely to be different, other aspects of the relationship under investigation. Here and there it may be possible to reinforce survey by experimentation, and ultimately considerable confidence, psychological rather than numerical, may be felt in the interlocking evidence. The process is, in fact, a miniature replica of the one by which some of the major biological theories have been established.

Summary

With the aid of examples taken chiefly from the American Rheumatism Association's Co-operative Study of Cortisone in Rheumatoid Arthritis, the possible usefulness of surveys of case records is examined.

A controlled trial would be necessary to measure the benefit actually due to a treatment, but a properly conducted survey can assess improvement after treatment, provided that the records are suitable and lost cases are few. Because therapy in most chronic diseases is multiple, the assessment must often refer to all patients with a certain disease in a certain clinic or subgroup, rather than to patients treated by any one method.

Aetiology and the relationships, or differences, between diseases can seldom be satisfactorily studied from case records.

Unless a survey is very carefully planned and executed, it can have "anecdotal" value only.

Satisfactory co-operative studies are very difficult, and, even if a perfect survey were possible, interclinic differences in verdicts should be expected.

Detailed recommendations are made regarding the planning of case-record surveys and follow-up studies.

Common errors are discussed, such as the misinterpretation of "incidence", and the belief that dosage-response relationships can be measured even though doses are changed at will during treatment (a circular argument).

Any inference from a survey, such as an association (or lack of association) between phenomena in two diseases, or within a single disease, can be fallacious because of undetected bias due to competition between selection rates (Berkson's fallacy). Some general methods are suggested for partial counteraction of this bias.
I am greatly indebted to Miss Claire Ling for many of the examples quoted from the Cortisone Study as well as for valuable comments, to Mrs. Lee Herrera for meticulous and constructive criticism throughout the preparation of the article, and to members of the Cortisone Study Committee—Drs. William H. Kammerer, Currie McEwen, Charles Ragan and Morris Ziff—for their careful scrutiny of the manuscript and for very helpful suggestions.

REFERENCES

(1954b). "Ibid., 6, 152.
(1955b). "Ibid., 6, 184.

L'emploi des dossiers des malades dans l'étude de la thérapie et d'autres traits d'une maladie chronique

I. Projet de l'enquête

RÉSUMÉ

A l'aide d'examens venant surtout de l'enquête en Collaboration de l'American Rheumatism Association sur la Cortisone dans le Arthritis Rhumatismale, on examine l'utilité d'enquêtes employant des dossiers des malades. Une expérience contrôlée serait nécessaire pour mesurer le bénéfice réel d'un traitement, mais une enquête bien dirigée peut évaluer l'amélioration après le traitement, pourvu que les dossiers s'y prêtent et que peu de cas se perdent. Comme la thérapie dans la plupart des maladies chroniques est multiple, l'évaluation devra toujours tenir compte de tous les malades atteints d'une certaine maladie dans une certaine clinique ou dans un groupe, plutôt que de malades traités par une méthode quelconque.

Les dossiers des malades ne se prêtent que rarement à une enquête satisfaisante de l'étiologie, des rapports ou des différences entre des maladies.

A moins que le projet et l'exécution d'une enquête soient très soigneux, sa valeur ne sera qu'anecdotique.

Des recherches satisfaisantes en collaboration sont très difficiles et même si une enquête parfaite serait possible, il faudrait s'atteler à des différences entre les maladies intercliniques.

On fait des recommandations minutieuses concernant la manière de proter les enquêtes utilisant des dossiers des malades et des données de la surveillance post-hospitalière.

On discute les erreurs habituelles, telles que l'interprétation erronée de la "fréquence" et la conviction que le rapport entre la posologie et la réaction thérapeutique peut être mesuré, même en changeant la thérapie les doses au cours du traitement (un argument circulaire).

Toute déduction d'une enquête, telle que la présence, ou l'absence d'une association entre des manifestations dans deux maladies ou dans une seule maladie peut être fallacieuse en raison d'un parti pris (biais) méconnu ou à la concurrence entre les taux de sélection (sohpisme de Berkson). On suggère quelques méthodes générales pour neutraliser partiellement ce parti pris.

Empie de fichas de enfermos en el estudio de la terapia y de otros rasgos de una enfermedad crónica

I. Proyecto de una investigación

SUMARIO

Con la ayuda de ejemplos tomados principalmente de la Investigación Cooperativa de la American Rheumatism Association sobre la Cortisone en la Arthritis Reumatoide, se examina aquí la utilidad posible de investigaciones que usan fichas de enfermos.

Se necesitaría un experimento controlado para medir el beneficio efectivo de un tratamiento, pero una investigación propiamente dirigida puede evaluar la mejoría después de un tratamiento, provisto que las fichas se presten a ello y que pocos casos se pierdan. Como la terapia en la mayoría de las enfermedades crónicas es múltiple, la evaluación debe siempre tomar en cuenta todos los enfermos con enfermedad determinada y en una clínica o en un grupo determinado de preferencia a enfermos tratados con un método cualquiera.
Las fichas de enfermos permiten raramente un estudio satisfactorio de la etiología, de las relaciones o de las diferencias entre enfermedades.

A menos que el proyecto y la ejecución de una investigación sean muy cuidadosos, su valor no será más que anecdótico.

Un estudio cooperativo satisfactorio es cosa muy difícil y hasta con una investigación perfecta habría que anticipar diferencias interclínicas de dictamen.

Se hacen recomendaciones detalladas acerca de la manera de proyectar investigaciones con fichas de enfermos y de seguir la evolución posthospitalaria de éstos.

Se discuten los errores comunes, tales como la interpretación errónea de la “incidencia” y la creencia de que se puede medir la relación entre las dosis y la respuesta terapéutica aunque se cambie libremente las dosis durante el tratamiento (un argumento circular).

Cualquier inferencia de una investigación, tal como afirmación o negación de la existencia de una asociación entre fenómenos en dos enfermedades o en la misma enfermedad puede ser delusoria a causa de un factor de parcialidad (bias) escondido, debido a la competición entre las razones de selección (sosfismo de Berkson). Se sugieren algunos métodos generales para neutralizar en parte esta parcialidad.
Use of Case Records in the Study of Therapy and Other Features in Chronic Disease: I. Planning the Survey
Donald Mainland

*Ann Rheum Dis* 1955 14: 337-352
doi: 10.1136/ard.14.4.337

Updated information and services can be found at:

[http://ard.bmj.com/content/14/4/337.citation](http://ard.bmj.com/content/14/4/337.citation)

**Email alerting service**
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

**Notes**

To request permissions go to:
[http://group.bmj.com/group/rights-licensing/permissions](http://group.bmj.com/group/rights-licensing/permissions)

To order reprints go to:
[http://journals.bmj.com/cgi/reprintform](http://journals.bmj.com/cgi/reprintform)

To subscribe to BMJ go to:
[http://group.bmj.com/subscribe/](http://group.bmj.com/subscribe/)