2. Measurement in health care

SUMMARY

Evaluation of health care is dependent on the collection of reliable and valid data, which is itself reliant on good measurement techniques. The general principles of measurement are outlined and approaches to the measurement of various aspects of health care are discussed including measurement of resources, patient utilization, and outcome in terms of mortality, morbidity, and patient response to the health care provided.

INTRODUCTION

Evaluation research draws upon the social sciences for its methods of inquiry. The social sciences, like other sciences employ the scientific method, which is a 'complex interchange between theory and observation, their means of communication being measurement'.[1] Patrick and Elinson propose that measuring is a procedure that provides the means of relating a concept(s) to a set of controlled observations that should provide ordered knowledge about the concept(s). It therefore advances theory through the accumulation of empirical evidence and as a consequence it is imperative that the data the measurement process provide are good and reproducible.

MEASUREMENT PRINCIPLES

Depending on the evaluation study, different types of information about the health-care service(s) under study and about the results of that service for an individual or a population are needed. Regardless of the information required several general principles of measurement apply.

Measurement consist of rules for assigning a value (numerical or nominal) to objects or events in such a way as to represent quantities, qualities, or categories of an attribute.[2] The use of the word 'rules' implies that the method for assigning a value must be explicitly stated and unambiguous. The use of rules is important to the issue of standardization. One fundamental aspect of standardization implies that similar results should be obtained by different people using the same measuring instrument. It should be stressed that what is being measured is some particular characteristic of an object, and not the object itself. A child's height, weight, or blood pressure is measured, not the child itself. Height, weight, and blood pressure are all attributes of the child.

The assignment of rules should result in only one possible value for a given

Assume that blood pressure is the patient characteristic being studied in a population considered at risk for hypertension. Blood-pressure readings will be obtained by the general practitioner at certain time intervals during a two-year follow-up intervention programme. It should be assured that the measurement instrument is reliable within each practitioner's office, that the physician is consistent in his readings, and if the patients are permitted to change practitioners, that the instruments and practitioners are comparable in their readings. If not, one would not know whether to attribute any changes to the programme, or to random errors in measurement.

Time is an important source of variation in health-care evaluation and therefore it is imperative to assess random variation in order to measure true variation. An early study of reliability in X-ray readings was published by Yerushalmy in 1953.[6] Variation in blood-pressure measurements can be due to the instrument, the observer, or may truly be different. Clark[7] in a study on hypertension illustrates how blood-pressure measurements vary under different conditions. Fleiss *et al.*[8] studied the diagnostic agreement among psychiatrists.

Unfortunately, the reliability of the measured results is not frequently assessed in medical studies, as is illustrated by a recent literature review on the reliability of clinical methods, data, and judgements.[9] Reid and Holland[10] stress some of the underlying principles of measurement in the health-care field and provide various examples of potential problems of reliability in assigning a diagnosis, or a value in a clinical examination. The degree of reliability is assessed by the data-appropriate correlation or agreement coefficient. The strength of the association required depends upon the ultimate use of the study results.

A commonly used measure of reliability and validity is the proportion of perfect agreement between observers. The contingency coefficient C,[11] is another popular measure. A simple proportion of agreement does not take into account the degree of agreement which could be expected by chance alone. The contingency coefficient C, measures association and not agreement *per se*, that is, if two reviewers are in perfect disagreement they will have a very high level of association. A valid measure of agreement should incorporate a correction for the degree of chance agreement, it should measure agreement, and should be amenable to a test of statistical significance of the degree of agreement.[8] Kappa, a coefficient of inter-observer agreement that can be used for nominal and ordinal scales provides these three factors:

1. It measures agreement corrected for that which is expected purely by chance.
2. It is scaled from -1 to $+1$, where negative values indicate worse than chance agreement, 0 indicates exactly chance and positive values indicate better than chance

3. It has a well-defined standard error which permits statistical assessment of the significance of the observed degree of agreement.[12]

Validity

Reliability is a *necessary* but not sufficient condition for validity. Validity, sometimes referred to as accuracy in the epidemiologic literature, is essential for drawing conclusions and is the extent to which a particular measure reflects what it is supposed to measure. Reliability is mainly concerned with random errors, while validity is mainly concerned with systematic error. An instrument can be reliable, but not valid for its intended purpose. The sources of error may be the same as for reliability, but may also be due to the inappropriate use of the instrument or to erroneous underlying theory. A measure may be valid for the purpose for which it was developed, but not necessarily valid for a related but not equivalent purpose. Thus it should be stressed that an instrument is not validated in the abstract, but through some practical use to which it will be put.[2]

There are several different types of validity, and our interest in one or several types depends upon how the measure will be used, the time and cost for different types of validation, etc. Three types of validity* are especially pertinent to health-care measurement: face validity, criterion validity and content validity.

Face validity (sometimes referred to as consensual validity) reflects general acceptance that an instrument indeed measures that which it claims to measure. The sphygmomanometer for measuring blood pressure is accepted on its face validity.

Criterion validity is the degree of comparability between one measure and another that is considered to be 'more valid'. Two forms of criterion validity are concurrent and predictive validity. Concurrent validity reflects the correlation of two measures at the same point in time. Cerebral angiography can be used for concurrent validation in the evaluation of the brain scanner. Predictive validity is an indication of the extent to which a measure is predictive of some future event or characteristic. An aptitude test developed for medical-school applicants should be validated for its predictive validity of medical-school performance. It would be more difficult to be predictive of 'good physician performance' because a measure of a 'good physician' does not exist, whereas grades based on the acquisition of medical knowledge are a usual part of the educational process.

Sensitivity and specificity are two important concepts that are related to criterion validity. Sensitivity is a measure of a test's ability to detect those individuals affected by a health problem, whereas specificity is a measure of a

* Construct validity is more petinent to the social sciences, although it is also relevant to health care. A construct represents a hypothesis or theory—it is abstract rather than concrete. Anxiety, intelligence, psychosis, and aggressiveness are examples of construct. The interested reader is referred to Nunnally[2] and Kaplan[13] for in depth discussion of construct validation.

type of scale. Thus, irrespective of the type of scale used, the resultant values or classes are mutually exclusive. The most common types of scales that are used in measurement and evaluative research are: nominal, ordinal, interval, and ratio scales.

Measurement scales

Nominal scales (or measures) are labels or classes of objects or events. A hospital admission number is a unique label each value used once only. The International Classification of Disease (ICD)[3] is a system of classification for medical diagnoses and problems. For one disease attribute there is one code only; but for one patient, there may be several codes representing several problems. Other examples are 'male/female', 'absent/present', etc. whose values are then coded numerically to facilitate data manipulation. Nominal scales can be used only for determining how often an event occurs. It is one example of qualitative data. Although there are those who do not consider the nominal scale as a measure, it is widely used in the health-care sciences and should therefore be understood in terms of its advantages and limitations.

Ordinal scales also provide data that describe classes of objects or events, however, the important difference between this and the nominal scale, is that ordinal classification is based on a continuum of 'most' to 'least' with respect to some characteristic. There is no way of assessing that characteristic in absolute terms and there is no way of determining how far apart the classes are from each other. In medicine it is common usage to employ the ordinal-scale classification of severe, moderate, or mild for the severity of a disease in a patient. The degree of satisfaction with an outpatient clinic could be subjectively provided by the categories 'very dissatisfied', 'dissatisfied', 'satisfied', or 'very satisfied'. It is not imperative that 'dissatisfied' be equidistant between 'very dissatisfied' and 'satisfied'. Usually, only the frequency of occurrence of certain categories can be calculated. When data are collected in this form, scaling techniques can be used to convert ordinal scales to 'higher level' scales, though this can be potentially problematic. The Likert method[4,5] is a widely used technique for aggregating ordinal ratings.

Interval scales measure classes of objects or events that are rank ordered with respect to some characteristic (as is the case for ordinal scales); the difference or distance between the objects is known but there is no absolute zero point only a relative zero point. Body temperature as measured by a Fahrenheit or Celsius thermometer is based on an interval scale. Interval variables can be added or substracted, thus providing the opportunity to apply basic statistics such as means and standard deviations. They cannot be multiplied or divided; for instance, the temperature cannot have doubled in a certain time period, it can have increased by x amount.

Ratio scales differ from interval scales in that they represent variables that have an absolute zero point or a specifically defined point of origin on the scale. This allows numbers to be meaningfully multiplied or divided. Examples include number of children in a family, length of time for a surgical procedure, length of stay in a hospital, a person's pulse rate, visual, and auditory accuity, etc.

It is important to recognize the differences between these types of measurement because by observing their respective limitations and using the appropriate analytical methods, sound conclusions can be reached. If not, potential errors in reasoning and inferences can occur due to inappropriate use of the data. For example, if the ordinal scale 'cured, improved, unchanged, worse' is used to evaluate a new treatment, and the values one to four are assigned to these categories, it would be inappropriate to calculate the mean for each treatment group unless there was some data conversion to equal intervals. Calculating the distribution of values for each treatment would be appropriate and would allow conclusions to be drawn.

Reliability and validity

In order to determine whether or not a measurement is useful, two questions must be asked:

1. Is the measurement reliable?
2. Is the measurement valid?

Regardless of the type of measurement, the issues of reliability and validity must be taken into consideration. If they are not, the errors in the data potentially threaten the conclusions of the study. This chapter discusses the validity of a measure and Chapter 4 will discuss the validity of a study.

Reliability

Reliability is the extent to which the same measure will consistently provide the same results—it is synonymous with the reproducibility or repeatability of the measurement instrument. Reliability is concerned mainly with chance or random errors that can be attributed to the subjects under study, the observers, the situations, the instruments, and/or the processing. Some of the basic ways to assess reliability are:

1. **Inter-rater reliability**—will two or more observers assign the same value to the characteristic being measured at the same point in time?
2. **Intra-rater reliability**—will the same observer assign the same value to the measured characteristic at different points in time, assuming the characteristic being observed has not changed?
3. **Split-half reliability**—for questionnaires or surveys, is there internal consistency between responses to items considered to be measuring the same concept (or short-form versus long-form) at the same point in time?
4. **Test re-test reliability**—does the same test given at different points in time provide the same result provided that change has not occurred?

test's ability to identify correctly those individuals who do not have the health problem being studied. Sensitivity is the ratio of true-positives compared to the total diseased group, and specificity is the ratio of true-negatives compared to the total non-diseased group. In later chapters Roberts and Hjelm discuss these concepts in greater detail.

Content validity depends upon the adequacy with which a specified content is sampled.[2] Content validity should be ensured in the development of the instrument, since it rests mainly on appeal to reason with regard to the adequacy with which the content has been sampled. Content validity is especially pertinent when developing course examinations or qualifying examinations for professional credentials. An examination at the end of a medical-school course in introductory clinical medicine should be designed comprehensively to sample the material taught during the course. If not, the examination is not a valid measure of the course content that the students were expected to learn.

The importance of these validation assessments should not be underestimated. When they are, the results of evaluation are likely to be compromised.

DATA COLLECTION

Numerous methods are available for collecting data which can be used for evaluating health care. The method(s) selected as most appropriate depends upon the context (cost, time scale, design, objectives) of the study. There are two basic sources of data: pre-collected data and original data.

Pre-collected data

These are data collected for purposes other than that of the evaluative study. Almost all countries have some system for collecting general population data on: geographical, personal and household, and economic characteristics; vital statistics; and morbidity statistics. Other examples of established data sources of value in evaluating health care are, patient medical records, hospital data on patient utilization, and other administrative information on staffing levels and resource utilization, etc.

The advantage of such data is that costs are low. However, the transformation into a format convenient for analysis can increase costs and the information, may be incomplete, inaccurate or inappropriate for the stated objective of the evaluation study.

Original data

These are data collected specifically for the purpose of an evaluation study. Original data may be collected on a continuous or *ad hoc* basis. Disease

registers are examples of continuous data collection conceived for study purposes. The Health Interview Survey[14] collects data on a continuous but intermittent basis. A community health knowledge survey would be an example of *ad hoc* data collection.

Collection of data can be by observation (a non-obtrusive measure), clinical or laboratory examination, questionnaire or interview. Questionnaires may be close-ended where simple yes/no answers are expected or open-ended. They can be self-administered or completed with the assistance of an interviewer. Interviews can be conducted at a medical facility, at work or at the interviewers home, or conducted by telephone.

Each method has its advantages and disadvantages and any new technique or new application of an established test or method must be pilot tested for reliability and validity before being used in a full-scale survey.

Original data collection is frequently more expensive than using established data sources and it is usual initially to examine existing data. If this is not sufficient then a search for pertinent, existing, tested measures (questionnaires, measuring instruments) should be carried out. Only if these efforts are unsuccessful should new instruments be developed and tested. Usually a combination of methods will be called for and the practical solution will be a function of the context, the value placed on the various drawbacks and constraints, and the sensitivity of the subsequent decisions on uses to which the information provided by the evaluation study will be applied.

MEASUREMENTS IN HEALTH CARE

The health-care evaluator is concerned with determining the extent to which a health-care programme or a medical intervention or a reorganization has achieved its objectives. These goals can be defined by changes in the health of individuals and/or populations, by patient or community or professional satisfaction, etc. These are all examples of a response to the intervention being evaluated—these are dependent variables.

In order to describe a health-care service, the evaluator also needs to describe and measure the inputs and process of that service. These are the independent variables of the study. Furthermore, potential intervening factors that can affect the relationship between the inputs, the process and the results should be taken into consideration and measured if considered relevant.

Multiple evaluation

The classic method of evaluation has been to choose one action, one effect or set of effects and concentrate on their relationship. Examples in clinical evaluation are too numerous to require citing. Amongst pioneer 'sociological' evaluations in medicine, the work of Wing and Brown[15] stands out. Singling out mental health as the dependent variable, they examined its sociological causation in a series of papers published since the beginning of the sixties. In the course of these studies they can be said to have evaluated three mental hospital regimes according to the prevalence of sociological conditions affecting schizophrenia. This is evaluation in its simplest sense: a comparison of conditions affecting one stipulated end.

Recent models of evaluation in the health service have been based on a more complex two-stipulated ends system, in which the effectiveness of services is traded off against cost.[16] This is the basic shape governing the economic public preference model.

Ultimately it is the research team who determine the dimensions of the evaluation and the items to be studied. They do this by exploring all possible sources of suggestion which are: (i) the data needed for decision making; (ii) the general causal theory of their disciplines; (iii) empirical findings reported in the literature; (iv) the experience of any person connected with events similar to the one which the team is evaluating. The need is to avoid arbitrary selection.

Measurement of resources

In order to evaluate the efficiency of the delivery of a health-care service it is necessary to know what resources are available. This subject will be dealt with rather consciesly since it is relatively simple and dependent on local circumstances.

Resources include not only money but also buildings, equipment, personnel, etc. It is essential that all resources are clearly defined in practical terms. Even a simple concept like the 'hospital bed' can have different meanings when measuring resources. It is therefore necessary to use standard definitions such as those provided in the WHO glossary.[17] The different kinds of personnel involved should be expressed in whole time equivalents together with a precise explanation of the meaning. For example the work of two part-time nurses might be equivalent to the work of one full-time nurse.

Resources can generally be measured in numerical terms. However ratios and indices are usually more useful than absolute numbers. All resources, if possible, should therefore be expressed in relation to the population; beds per capita, manpower per capita, etc. Where possible resources should also be related to specific age groups; for example 'long-term beds' for the elderly. There is, however, no universally agreed definition for international use, for 'long-term beds' or even for 'elderly'. It is arguable that there should be at least three categories, 65–74 years, 75–84 years, and >78 years, since these groups have different demands on the health-care system.

Other useful indices for measuring resource availability include:

1. Outpatient consultations (number and duration per capita)
2. Different staff category ratios (nurses: doctors, anaesthetists: doctors, etc.).

3. Staff per hospital bed.
4. Distribution of patients by general practitioner.

Complete data on health manpower supply should include sex, qualifications, activity status, age and/or year of graduation, retirement, and data on health manpower training, e.g. number of students in different schools, qualifications of different teaching personnel.[18,19]

Measurement of utilization

At the present time, most individual facilities, as well as local regional and national government agencies, maintain statistics on the activity or utilization of facilities. Examples include: number of patients admitted to hospital, lengths of stay by diagnostic group, number of visits to the outpatient department or to general practitioners, number and type of surgical interventions, number of individuals screened during a prevention programme, and costs for treating a given diagnosis. Data of this type can be collected on a routine sampling basis, which is the case for Britain's Hospital In-Patient Enquiry (HIPE), or may be collected for all encounters with a given health care practitioner or institution, as in the Professional Activities Study and Medical Audit Program which is organized in the United States by the American Commission on Professional and Hospital Activities.

Whichever system is adopted the shortcomings of these data sources are similar. Utilization data usually count events and do not permit any assessment of the number of patients that generate those events, though record linkage could overcome this shortcoming. If events are identified by a unique patient/citizen number, studies reflecting more accurate utilization data (i.e. better knowledge of true patient population) might then be carried out.

Bed utilization statistics

Simple administrative indices of bed utilization will be considered first. Although their use and interpretation are apparently simple, data of this kind are far from universally homogeneous. The conventional indices of utilization are discussed in detail in a number of publications.[19,20] The most frequently employed indices are computed as follows:

1. Mean duration of stay

$$= \frac{\text{Total occupied bed-days in a period}}{\text{Discharges and deaths in the same period}}.$$

2. Turnover interval

$$= \frac{\text{Number of available bed-days} - \text{number of occupied bed-days}}{\text{Discharges and deaths}}.$$

The turnover interval is the mean interval during which each bed is empty after the discharge of one patient and before the admission of the next.

3. Bed turnover rate (discharges per available beds)

$$= \frac{\text{Discharges and deaths in a period}}{\text{Average available beds in the same period}}.$$

This rate expresses the mean number of patients that use a given bed during a period, usually a year.

4. Bed occupancy rate (percentage bed occupation)

$$= \frac{\text{Occupied bed-days in a period}}{\text{Number of available beds in the same period}} \times 100.$$

Usually the number of occupied bed-days is obtained by the total number of inpatients present in the department at midnight. However, in units characterized by a long stay (e.g. nursing homes or geriatric departments) it is better to calculate a mean duration of stay based upon a one-day census of the patients in the department. In addition to the average length of stay of deceased or discharged patients this would provide data about how long the patients present on a certain day of the year have been in the department.

The indices listed are inter-related and it is usually sufficient to look at only two of these indices to obtain all the relevant information. The question is which two? The Cogstat Report[20] recommended adopting the turnover interval and the mean duration of stay.

The turnover interval provides a direct measure of the wastage in hospital bed utilization and the length of the turnover interval is perhaps the only bed utilization index for which uniform standards are possible. There is less variation for the turnover interval than for the duration of stay. It is rarely possible to justify a turnover interval exceeding three days.[19] The turnover interval is certainly the index more easily modified by administrative intervention. A long turnover interval may reflect a low level of demand or inadequate admission procedures. It should be ascertained whether a short turnover interval is due simply to an increase in the mean duration of stay. Trends in mean duration of stay and in turnover interval must therefore be studied together. Obviously if *both* are decreasing or at least one is decreasing more than the other increases it is possible to observe the desired objective, i.e. an increase in bed turnover.

The length of stay reflects the medical decision taken during patient's stay. It can also reflect the patient's social problems, the inadequacy of domiciliary services (for example when patients are not discharged because they cannot be cared for at home) or the lack of appropriate facilities, such as operating facilities, radiology, or pathology. Generally the length of stay distribution is not normal, but skewed to the right. Hence the median length of stay would be a better measure of duration of stay than the average. The frequency distribution would be a still better measure but this is possible only when using the lengths of stay of individual patients and practically only if data are analysed by computer.

Bed occupancy rate is probably the most commonly used index today. It is important to appreciate that the same bed occupancy rate may be obtained with many different lengths of stay (Fig. 2.1). This index can be an overall guide to the use of inpatient resources, but is not useful by itself for comparing specialities or departments.

The relationship between mean duration of stay and bed occupancy is shown graphically in Fig. 2.1.[21] As can be seen, an 80 per cent bed occupancy may be achieved when the mean length of stay is four days and the turnover interval is one day or when mean length of stay is as long as 28 days if the mean turnover interval is seven days. It is implicit from what has been said that when the number of admissions decreases (as in paediatric departments today) there is a tendency to keep the bed-occupancy rate high by increasing the length of stay. Alternatively a high length of stay and a high occupancy rate (around 100 per cent) may reflect the fact that the department is overworked. Between these extremes there is probably a range of occupancy rates corresponding to

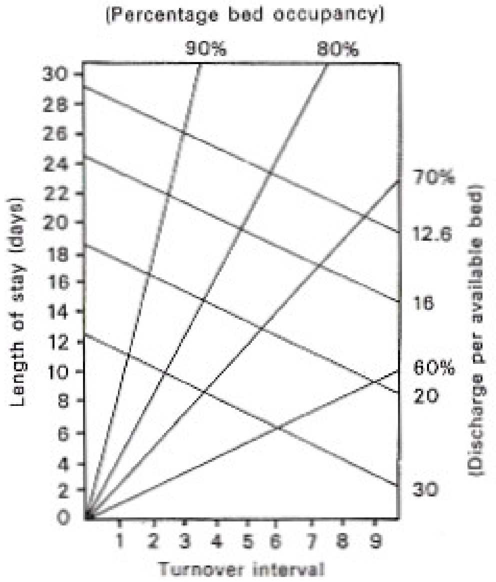

Fig. 2.1. Length of stay in relation to turnover interval plotted in days. Lines radiating from 0 indicate the percentage bed occupancy. Lines from left to right indicate discharges per available bed (bed turnover rate). Source: Tyrell (1975).[21]

an optimal mean length of stay.[22] This will differ from department to department and will not, in any case, have an absolute value since it is dependent on attitudes, organizational patterns, and the existence of pre-admission and post-discharge services.

All the indices described depend on the method by which the number of beds are determined, especially in countries where beds assigned to the various specialties are not physically separated into different departments. The number of beds available to a specialty includes those beds borrowed from other specialties but does not include beds on loan and as may be inferred, this solution is not completely satisfactory.

Waiting list

The importance of careful recording and examination of waiting lists in order to evaluate the relationship between resources and demand cannot be overstressed. The lack of data on waiting lists for non-urgent conditions is usually an indication of administrative sloppiness and/or the overprovision of services. On the other hand the mere existence of a waiting list is not enough to be a useful source of evaluation data. The waiting list should be properly updated and not manipulated.

The interpretation of waiting list data is not as simple as it may seem. Patients who *have* to wait for control of their clinical conditions should be excluded from waiting lists. It is more meaningful to consider the time each patient has to wait rather than the number of patients on the list.[21]

When waiting list admissions are available the mean waiting time can be calculated as follows:

Mean waiting time (e.g. week)

$$= \frac{\text{Number of patients on waiting list} \times 52}{\text{Number of patients admitted per year}}.$$

For outpatient consultations it may be sufficient to indicate the mean length of the delay. However the median length of time on a waiting list and the proportion of patients who have been waiting specified lengths of time, for example more than one month (i.e. the frequency distribution of waiting times) are often more revealing than the mean waiting time.

Waiting lists for elective surgery and admission to long-term health-care facilities are generally considered easier to interpret. Consequently they are particularly useful for comparative purposes.

Individual discharge data

All countries in the EEC (if not on a national basis, at least at a regional or hospital level) have introduced some kind of form to be completed on individual hospital patients at or after discharge. All these statistical summaries include at least data about diagnosis, age, sex, residence, and

surgical operations performed. It would seem a flagrant waste not to use these data to try to improve the evaluation of the through-put in hospital departments. The enormous variations in length of stay distribution by individual diagnosis, the percentage of patients operated on in surgical departments and the median waiting period between admission and surgical operation have been repeatedly shown.

A more general measure of utilization for comparing entire departments has been hindered because information on the principal diagnosis at discharge is not sufficient to judge the amount of work carried out for individual patients. Obviously this would depend also on the clinical severity, on the type of admission (first or follow-up, urgent or not), and on the type of clinical procedures commonly available in the department. The latest edition of the International Classification of Disease (IX)[3] itself is regretably inadequate for this particular purpose, as are also the proposed diagnostic groupings.

Interesting attempts to overcome this difficulty, usually known as the 'case mix' variability problem, have been made. For example Fetter *et al.*,[23] grouped the principal diagnoses into 323 homogeneous classes, and further subdivided patients according to the presence or absence on the discharge sheet of additional diagnoses and major or minor surgical operation. Usually discharge diagnoses are only weighted against some kind of average length of stay (regional or national), and this solution is far from satisfactory.

When data on the residence of patients are available, it is possible to estimate the 'true' catchment population as distinguished from the administrative one. This gives a useful indication of the service ability to satisfy and to stimulate demand. The most common procedure for computing the 'true' catchment population is that proposed by Bailey:[24]

Catchment population (by age and sex groups, and possibly also by diagnosis)

$$= \sum_{k} \frac{\text{No. of discharges in the service of interest (by age and sex group)}}{\text{total number of discharges}},$$

where the summation is over the k administrative unit from which patients come to the service of interest and the catchment age and sex groups are summated. It has been shown that Bailey's method is not appropriate in densely populated urban areas.

Outpatients activity

It is desirable to separate new patients from those returning for follow-up care and treatment. To evaluate the efficiency of the work of outpatient clinics the following checks should be kept regularly:[21]

1. **Number of new patient requests**—if this is larger than the number of new patients seen, the total number of patients attending outpatient clinics and the waiting time will rise.

2. **Number of new and number of old patients seen**—usually in terms of their ratio and their relationship with the staff working time. Time trends can be studied, and comparisons with other homogeneous clinics made.

3. **Waiting lists and waiting time for new patient appointments**—problems and aproroaches apply as discussed for inpatient waiting lists.

Furthermore the following checks should be done periodically:

1. **Waiting time in the clinic**—if an appointment system is in operation.

2. **Average consultation time for new and old outpatients**—to see if the booking rate is satisfactory. The variability between new and old outpatient attendances per 1000 population can be used to evaluate the need for and the quality of care.

General practice

Traditional indicators of activity in general practice are rates of referral to specialists and the amount and cost of drug prescription. If these two indicators are elevated this would suggest an excess transfer of patient responsibility and over provision of treatment. Data on drug prescription is usually obtained from pharmaceutical services. A periodic comparison between general practitioners and against national and international averages is probably useful and sufficient. The rate of referral to specialist can usually be obtained only through *ad hoc* studies.

Operational research

The methods of operational research applied to health-care utilization (use of models, network analysis, linear programming techniques, queuing theory, and simulation) are clearly and concisely described by Grundy and Reinke.[25]

Measurement of outcome

Outcome as discussed earlier is one of the most important indicators of the effectiveness of health care. The simplest measure of effectiveness of health care is *relative effectiveness* which is the ratio of the outcome in individuals exposed to a health-care measure and in a group given a different treatment or no treatment at all.[26] A more sophisticated approach to relative effectiveness is described by Peto *et al.*[27,28] *Attributable effectiveness* refers to the difference, rather than the ratio for outcome between the two groups. Attributable effectiveness when weighted for (actual or potential) frequency of application of the measure in the population gives a third index, the *population attributable effectiveness*. This index measures the beneficial impact in absolute terms, on the total population.

Since evaluation of effectiveness depends on pre-specified outcome criteria it is important that these should be measurable and appropriate. Inappropriate outcome criteria may be classified into three groups:

1. **Irrelevant or insensitive criteria**. For example, many psychotropic drugs may have a substantial impact on the quality of life, which is impossible to evaluate with the use of mortality or fatality indices.

2. **Restricted criteria**. For example, most of the standard epidemiological indices do not allow the evaluation of important psychosocial aspects of health care including professional, family, and community responses.

3. **Biased criteria**. For example, it would be inappropriate to compare the survival of screening-detected cases of cancer with the survival of cases detected after the appearance of the first symptoms.[29] Even without any beneficial effect of screening the survival in the first group would be better because of the additive effect of the lead-time bias (earlier diagnosis implies longer 'survival' even without postponement of the time of death) and the length-time bias (slow-growing tumours are over-represented in any prevalence study, including screening examinations).

Mortality and morbidity

General mortality rates, as well as diagnosis specific mortality rates have been used for a long time to assess the health of populations, the effectiveness of medical treatment, and the effectiveness of prevention such as vaccination programmes. Mortality statistics are relatively useful in the assessment of needs for health services. However, they may not be sensitive enough for assessing health interventions in the industrialized countries, and are of disputable value in areas of the world with high death rates but questionable data sources. Case fatality rates are good indicators of quality of care in comparative studies, as well as in effectiveness studies of alternative forms of treatment. Perinatal and infant mortality rates have long been used as health indicators, and are especially useful for identifying the subgroups of the maternal population at risk. Crude mortality rates have limited value for evaluators. Mortality data become more interesting once they are standardized, usually at least by age and sex (where pertinent) or by birth weight, parity, and mother's age.[30] Standardized mortality rates were used to compare geographical variations in morbidity in a review of the distribution of resources for health-care services within England.[31] Life expectancy or life tables are often used to evaluate two or more treatments. These are useful for testing the hypothesis that differences in effectiveness remain constant during follow-up.[32]

Although mortality statistics as an outcome measure may be appropriate for certain interventions, often they are not sensitive enough for evaluative purposes. For the past three decades researchers have been working on the development of other outcome indicators or indexes. The term indicator refers to the measure of a specific dimension of health, such as infant mortality rate, case fatality rate, accident rate, disease incidence and prevalence, etc. An index is a composite, usually weighted, of several indicators, and is sometimes referred to as a complex indicator. The problem with health-status indicators or indexes is their capacity to reflect the health state of a group. Should one aggregate by summing the health state of the individuals in the group or is the health of the group a quantity other than the sum of the constituents? The aggregation of indicators or indexes across a group complicates the measurement process in that different people assign different levels of importance to various diseases or handicaps.

Measuring total mortality is easy enough; however, cause specific mortality is more problematic since it is dependent upon who has identified the cause of death and the methods used for identification. The measurement of morbidity, and hence impairment, is far more complicated and difficult. Morbidity and disability are not always clearly defined, and the handicap they engender is affected by the social and emotional context. Furthermore, it is not only the occurrence of a morbid or handicapping condition that is of importance. The duration, intensity, and severity of and the stigmatization attached to the condition also add to the measurement difficulties.

The Apgar score is an indicator of the health status of the neonate that is made up of five criteria, each of which can take on the value 0, 1, or 2. The score is calculated by adding the values of the five criteria, and it can vary from 0 to 10 (perfect state at birth). It is measured at one and five minutes after birth. Roumeau-Rouquette *et al.*[33] discussed it's value as a risk indicator, i.e. an indicator that has predictive validity. The Apgar score has been validated for its predictive value of neonatal death, i.e. death within seven days from birth. They suggest that an Apgar score of eight is the optimal discriminating level (in terms of sensitivity and specificity) of this risk indicator for neonatal death.

A health index developed for medical care evaluation is the Sickness Impact Profile (SIP)[34] which measures the functional status of patients recovering (or recovered) from an illness. It is a 136-item questionnaire (a reduced version is being developed and validated) designed to determine physical and social/emotional status. It was initially designed for use on a general hospital population. It has recently been used to evaluate the impact of trained paramedics versus emergency technicians providing cardiopulmonary resuscitation to victims of cardiac arrest. This is part of the assessment of an experimental suburban paramedic programme.

Another functional status type of index has been developed by Bush *et al.*[35] An early version was used to evaluate a phenylketonuria screening programme in a community. For each form of the disease a person was assigned a probability of being at that disease or function level at some future point until death (Fig. 2.2). For each point in future time the probability of being at that level is multiplied by the relative value (that reflects a preference rating) assigned to that function level resulting in a weighted probability. A cure or plot is generated for the prognosis with and without treatment. Each set of points along the curve represents the expected level of well-being over all time periods for a group with the defined disease form. The difference between the two estimates is the amount of function level, that can be attributed to the treatment.

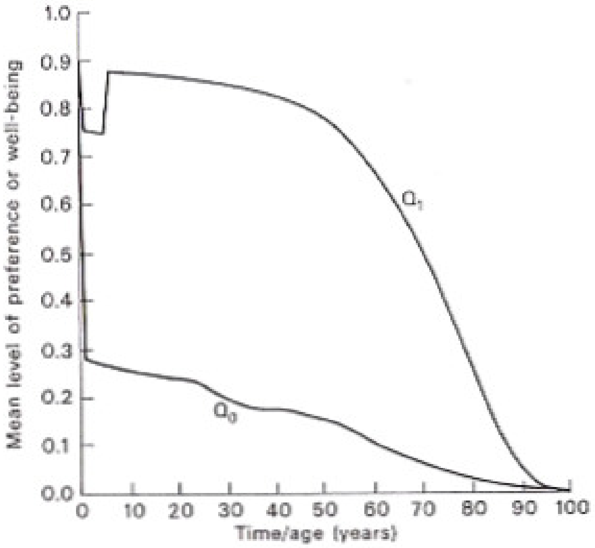

Fig. 2.2. Mean level of well-being over time for classic PKU, with and without treatment, using consultant's value set. Upper line (Q_1) represents treated cases and depicts the initial dysfunction imposed by diet during first six years followed by a gradual decrease in mean function imposed by general mortality rate. Lower line (Q_0) represents untreated cases and shows the lower levels of function and higher mortality rates experienced by the severely retarded. Source: Bush *et al.* (1973).[35]

In another section, the drawbacks of using only years of life gained as a measure of effectiveness are discussed. Bush's index is one approximation of a measure of quality of life that is incorporated with duration of life. This is important because often there is a trade off between longevity and quality. The concept of quality adjusted life years is therefore more acceptable. The approach measures the number of years with full health that are valued as equivalent to the actual years of life as experienced if ill or impaired.[36]

This measure of health was used by Weinstein and Stanson[36] to evaluate a national hypertension policy for the United States. For example, a year spent disabled following a stroke might be assessed as 0.4 quality adjusted years, two years as 0.8, while a chronic but mild respiratory problem might be valued at 0.99. As this scale is a subjective one, thus generating different values among different people, the societal range of weights should be used to reflect the spectrum of individual values. The sensitivity of a decision to this range of values should also be analysed.

Time until full recovery or duration of disability was used by Contandriopoulos *et al.*[37] to evaluate one-day surgery against hospitalized inpatient surgery for randomly allocated patients in predetermined diagnostic groups. It is a fairly easy measure to obtain for many patients, i.e. people who are employed and students. However, it is more difficult to determine for the elderly or housewives since the moment they resume full activity is not as clearly identifiable.

The issue becomes more problematic for people suffering from chronic of long-term illnesses. Assessing the effectiveness of a new service needs a measure highly sensitive to small changes. The Activities of Daily Living (ADL) Index grades patients on their dependency level based on six activities: transferring, bathing, dressing, toileting, eating/feeding, and continence. The Patient Classification for Long Term Care, which measures comparable indicators, but determines the type of long-term care needed by the patient, includes data on services rendered and frequency of social visits. This instrument is useful for comparing the outcomes for patients in different facilities or treatment settings and has been used for assessing the type of facility a chronic care patient should be in.[38]

Measurement of patient's responses

An early study dealing with patient's perceptions was by Cartwright.[39] This study set out to obtain a complete perspective on the patient's view of the hospital service and as a result many questions in her standardized interview schedule were non-evaluative. The evaluative questions tended to allow only two possible outcomes, a favourable or unfavourable evaluation: e.g. 'In general do you feel satisfied or dissatisfied with the medical treatment you received while you were in hospital?' Such questions would provide poor discrimination in any comparative study as very gross differences would be necessary for one hospital to have a statistically significant higher proportion of satisfied patients. Answers to such questions can also be criticized as they are likely to be influenced by many factors other than the patients' actual satisfaction, for example, the general willingness of subjects to admit dissatisfaction.

The type of information collected by Cartwright would also be inadequate because it gives no indication of the relative importance of the values investigated. McGhee[40] used an unstructured interview and this provides some idea of the importance of various elements of care by the number of patients mentioning them. Thus communications and food were mentioned by all patients, but only 26 per cent of patients made any comment on aftercare.

McGhee graded patient response into 'satisfied', 'satisfied with reservations', 'dissatisfied with reservations', 'dissatisfied', and 'no response'. Such a series is obviously more useful in making a discrimination between hospitals than simple dichotomous answers. Discrimination can be improved further by using attitude-scaling techniques such as those which have been used in the evaluation of psychiatric care[41, 42] and of patients' attitudes to doctors.[43]

These techniques, by combining the scores from numerous measures with small discriminative ranges, give an overall score which not only has a wider discriminative range, but also produces a score with greater repeatability and greater validity.

In a recently reported randomized controlled trial of two hospital regimens female surgical patients were asked to rate 17 different aspects of their hospital care: food, ward routine, sleep, toilet facilities, information, medical treatment, other patients, embarrassment, privacy, nursing staff, doctors, other staff, the member of staff they liked best, admission, discharge, and transfer. This list was constructed during pilot work, starting from items which patients described as important.[44]

> Three bipolar scales (good–bad, successful–unsuccessful, fair–unfair) each of seven points were used: these have been shown to measure an evaluative dimension over a wide range of attitude objects and over a number of different subject groups in several factor-analytic studies. For each aspect of care an evaluative score (ranging from 3 to 21) was obtained, plus an assessment of a seven-point scale for the importance of the item to the patient. Since importance scores were universally high. It appears that these 17 items were salient for the patients.
>
> In addition after completing the evaluative ratings, each patient was given the list of 17 items and asked to select the three most important to her. Patients also rated their whole hospital stay from admission to discharge on eight, seven-point bipolar evaluative scales giving scores between 8 and 56. In the outcome there were only small differences in patient satisfaction between the two regimens.

In the randomized controlled trial just referred to, anxiety was measured by the State–Trait–Anxiety Inventory,[45] and the State version of this well validated and extensively used questionnaire was employed to measure transient fluctuation in anxiety before, during, and after hospitalization and surgery. Significant differences were shown; patients continuing to be treated after operation in the specialist surgical units manifesting higher levels of anxiety that were maintained up to and after discharge.

This same study attempted to measure moods—anger, happiness, fear, depression, psychological well-being, and lethargy together with distress associated with hospitalization. This last response was measured by the Hospital Anxiety Scale[46] and Hospital Adjustment Inventory.[47] These two questionnaires investigate patients' worries about various aspects of hospitalization. They gave meaningful results in the original trials and the Hospital Adjustment Inventory gave similar results to a separate measure of psychological distress in surgical patients in a preliminary study.[48]

CONCLUSION

Measurement is an intuitively simple concept, but in fact it has the potential for creating many errors in the evaluation process. This chapter has presented some of the more important principles of measurement and has discussed the various ways of measuring outcomes, different aspects of health care such as;

morbidity, mortality, health and satisfaction, utilization of services, and resources such as facilities and manpower. Examples of these indicators and their calculation have been used to illustrate their utility or inadequacy.

REFERENCES

1. Patrick, D. L. and Elinson, J. Methods of sociomedical research. In *Handbook of medical sociology*, 3rd edn (ed. H. Freeman, S. Levine and L. Reeder) Prentice-Hall, Englewood Cliffs, New Jersey (1978). p. 437
2. Nunnally, J. C. *Psychometric theory*, 2nd edn. McGraw-Hill, New York (1978).
3. World Health Organization. *Manual of the international classification of diseases, injuries, and causes of death, based on the 9th revision conference in 1975*. WHO, Geneva (1977).
4. Likert, R. A technique for the measurement of attitudes. *Archs. Psychol.* **140**, 132 (1932).
5. Brook, R. H. *et al.* Overview of adult health status measures fielded in Rand's health insurance study. *Med. Care* Suppl. 7, **17**, 1 (1979).
6. Yerushalmy, J. Reliability of chest roentgenography and its clinical implications. *Dis. Chest* **24**, 133 (1953).
7. Clark, G. E. *et al.* Studies in hypertension I–IV. *J. chron. Dis.* **4**, 231, 469, 477, 490 (1956).
8. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol. Bull.* **76**, 378 (1971).
9. Koran, L. M. The reliability of clinical methods data and judgements—Part I and II. New Engl. J. Med. **294**, 642, 695 (1975).
10. Reid, D. and Holland, W. W. Measurement in health care studies. In *Health care and epidemiology* (ed. W. W. Holland and L. Karhausen) p. 8. Henry Kimpton, London (1978).
11. Siegel, S. *Nonparametric statistics*. McGraw-Hill, New York (1956).
12. Cohen, J. A coefficient of agreement for nominal scales. *Educat. Psychol. Measurement* **20**, 37 (1960).
13. Kaplan, R., Bush, J., and Berry, C. Health status: types of validity and the index of wellbeing. *Hlth Serv. Res.* **11**, 478 (1976).
14. US Department of Health, Education and Welfare, National Center for Health Statistics. *Health Interview Survey Procedure, 1957–1974*. National Center for Health Statistics, Rockville, Md (1975).
15. Wing, J. K. and Brown, G. W. *Institutionalism and schizophrenia*. Cambridge University Press (1970).
16. Cochrane, A. L. *Effectiveness and efficiency: random reflections on the health services*. Oxford University Press for Nuffield Provincial Hospitals Trust, London (1971).
17. Hogarth, J. *Glossary of health care terminology*. Public Health in Europe No. 4. WHO Regional Office for Europe, Copenhagen (1978).
18. Mejia, A. and Fullop, T. Health manpower planning: an overview. In *Health manpower planning* (ed. T. H. Hall and A. Mejia) p. 9. WHO, Geneva (1978).
19. Llewelyn-Davies, A. and Macauley, H. M. *Hospital planning and administration*. Monographs Series No. 54. WHO, Geneva (1966).
20. Morris, D. *et al. Cogstat*. King Edwards' Hospital Fund, London (1974).
21. Tyrrell, M. *Using numbers for effective health services management*. Heinemann, London (1975).

22. Zanetti, M. *et al.* Criteri di valutazione dell'efficienza di un sistema ospedaliero. *Gli ospedali della vita* **3**, 27 (1976).
23. Fetter, R. B. *et al.* Case mix definition by diagnosis-related groups *Med. Care* Suppl. 2, **18**, 1 (1980).
24. Bailey, N. T. *Mathematics, statistics and health systems*. Wiley, New York (1977).
25. Grundy, F. and Reinke, W. A. *Health practice research and formalized managerial methods*. Public Health Papers No. 51. WHO, Geneva (1973).
26. Roberts, C. J. *Epidemiology for clinicians*, p. 41. Pitman, London (1977).
27. Peto, R. *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. *Br. J. Cancer* **34**, 585 (1976).
28. Peto, R. *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br. J. Cancer* **35**, 1 (1977).
29. Cole P. and Morrison, A. S. Basic issues in population screening for cancer. *J. natn. Cancer Inst.* **64**, 1263 (1980).
30. Eduard, L. and Alberman, E. Changing maternal age, parity and cause of fetal wastage. *Revue epidemiol. Santé publ.* in press.
31. Department of Health and Social Security. *Sharing resources for health in England: Report of the Resource Allocation Working Party*. HMSO, London (1976).
32. Ipsen, J. Use of vital statistics. In *Health care and epidemiology* (ed. W. W. Holland and L. Karhausen). p. 32 Henry Kimpton, London (1978).
33. Rumeau-Rouquette, C., Breart, G. and Padier, R. *Methods en epidemiologie*. Flammarion Medecine–Science, Paris (1981).
34. Bergner, M. *et al.* The sickness impact profile: conceptual formulation and methodology for the development of a health status measurement. *Int. J. Hlth Serv.* **6**, 393 (1976).
35. Bush, J. W., Chen, W. W., and Patrick, D. L. Health status index in cost effectiveness analysis of PKU program. In *Health status indexes* (ed R. L. Berg) p. 172 HRET, Chicago, (1973).
36. Weinstein, M. C. and Stason, W. B. Foundations of cost-effectiveness analysis for health and medical practices. *New Engl. J. Med.* **296**, 716 (1977).
37. Contandriopoulos, A. P., Pineault, R., Lance, J. M., Lemieux, F. S. and Levasseur, M. An evaluation strategy for one day surgery programs. In *Evaluation of efficacy of medical action* (ed. A. Alparovitch, F. T.de Dombal, and F. Grémy) p. 421 North-Holland, Amsterdam (1979).
38. Jones, E. W., Densen, P. M., and McNitt, B. J. An approach to the assessment of longterm care. In *Measurement of levels of health* (ed. W. W. Holland, J. Ipsen and J. Kostrzewski) p. 299. WHO, Copenhagen (1977).
39. Cartwright, A. *Human relations and hospital care* Routledge and Kegan Paul, London (1964).
40. McGhee, A. *The patient's attitude to nursing care*. E and S. Livingstone, London (1961).
41. Rice, L. E., Pett, S. L., Berger, O. G., Sewall, L. G., and Lembau, P. V. The Ward Evaluation Scale. *J. clin. Psychol.* **19**, 251 (1963).
42. Caine, T. M. and Smail, D. J. *The treatment of mental illness*. University of London Press (1969).
43. Hulka, B. S., Zyzanski, S. J., Cassel, J. C., and Thompson, S. J. Scale for the measurement of attitudes towards physicians and primary medical care. *Med. Care* **8**, 429 (1970).
44. Johnston, M., Lee-Jones, M., and Bennett, A. E. A randomised controlled trial of post-operative care in community hospitals *J. Epidemiol. commun. Med.* in press.

45. Spielberger, C. D., Gorsuch, R. L., and LusRene, R. E. *Stai manual*. Consulting Psychologists Press, Paola Alto, California (1970).

46. Lucente, F. E. and Fleck, S. A study of hospitalization anxiety in 408 medical and surgical patients. *Psychosomat. Med.* **34**, 304 (1972).

47. De Wolfe, A. S., Barrell, R. P., and Cummings, J. W. Patient variables in emotional response to hospitalization for physical illness. *J. consult. Psychol.* **30**, 68 (1966).

48. Johnston, M. Aspects of inpatient expence. In *Community hospitals: progress in development and evaluation* (ed. A. E. Bennett). Oxford Regional Hospital Board, Oxford (1974).