DOCUMENT RESUME

ED 088 948 TM 003 513

AUTHOR Chen, Martin K.
TITLE Outcome Measures of Health Programs: What and How?
PUB DATE Apr 74
NOTE 12p.; Paper presented at the American Educational Research Association annual meeting (Chicago, Illinois, April 15-19, 1974)

EDRS PRICE MF-$0.75 HC-$1.50
DESCRIPTORS Evaluation; *Health; *Health Programs; *Measurement; *Models; Research Design; Statistical Analysis

ABSTRACT

With the proliferation of new health programs, such as Health Maintenance Organizations (HMO's) and Professional Service Review Organizations (PSRO's), the task of evaluating the impact of such programs on the health delivery systems and on the health of the American people becomes more urgent. Thus far no experimental or quasi-experimental designs have been found that are both feasible and satisfactory. Some quasi-experimental designs, particularly interrupted time series, have been suggested as a possible solution to the problem. The strengths and weaknesses of this and other designs, as well as the statistical problems associated with them, are discussed. (Author)

Session 9.03

OUTCOME MEASURES OF HEALTH PROGRAMS--WHAT AND HOW?

Martin K. Chen
National Center for Health Services Research and Development
Rockville, Maryland

There are two basic problems confronting policymakers in the organization and delivery of health services. They are (1) the measurement of health in quantitative terms, and (2) the establishment of a causal nexus between a health program and the measure of health as an outcome of the program. Until these problems are successfully resolved, decision-making in the arena of health care must by necessity be made on evidence other than hard data on health status. Some health administrators in their decision-making use such indicators as the rate of utilization of different types of services, cost of operation, and consumer satisfaction with services, etc. To be sure, all these indicators tell something about the quality of the program. In the final analysis, however, a health program cannot be said to have fulfilled the requirements of society unless indisputable evidence is generated that the program has improved the health status of the community it aims to serve. In other words, the most important outcome measure of a health program should be health status.

But what is health status? Is it mortality rate? Morbidity rate? Combination of mortality and morbidity? Freedom from physical and mental dysfunctions? Positive feeling of well-being? Predisposition to illness? Health is, of course, all of this and more. Depending on one's orientation, health has been defined in many ways that involve one or more of these aspects. The only definition that is designed to be comprehensive and encompass the totality of health as it is generally conceived is that of the World Health

Organization (WHO). (1) However, this definition that "health is a state of complete physical, mental and social well-being and not merely the absence of disease and illness" is considered by the WHO as being too imprecise to lend itself to objective measurement. (2)

In spite of the conceptual and definitional difficulties just described, many health indicators have been developed. Explicitly or implicitly, authors of these indicators have based their developmental work on operational definitions of health that differ in varying degrees from one another in orientation, emphasis, and conciseness of terms. To facilitate the study of health indicators extant and to be developed, Chen (3) has designed a classification model by which the indicators can be properly categorized for various purposes, including the establishment of a clearinghouse for health indicators.

SUGGESTED CLASSIFICATION MODEL

Based in part on the classification scheme of Baumann,(4) this classification model possesses the following characteristics:

(1) It is flexible enough to cover all of the essential dimensions on which the health indicators can be differentiated;

(2) The dimensions are non-overlapping and independent; that is to say, the scale of any one dimension is independent of the scales of all the other dimensions in the model;

(3) The model is exhaustive in the sense that all health indicators that have been developed and will be developed in the future can be fitted into the proper categories of the model; and

(4) The designation of each category can be uniquely determined for easy manual or computer storage and retrieval; in other words,

the categories of indicators generated by this model are mutually exclusive.

To simplify exposition, only three dimensions are used in this model. There is nothing sacrosanct about the number "three;" this number is used because a three-dimensional model can be represented geometrically for visual inspection, whereas anything above three dimensions involves hyperspace that is easily represented by algebra but not by Euclidean geometry. Furthermore, the steps along each dimension are flexible, since the dimensional scales are nominal and have no ordinal value. However, it must be remembered that the total number of categories is the product of the number of steps of all the dimensions and this number can be staggering if the number of dimensions and the number of steps along each dimension become too large. The law of parsimony demands that the minimum number of dimensions and steps adequate to do the job be used as the optimum.

THREE DIMENSIONS OF CLASSIFICATION MODEL

The three dimensions are utility, measurement and orientation. By utility is meant whether an indicator is applicable to an individual or a community or nation. The adoption of this two-category dimension is by design. It is assumed that an indicator that applies to a community is also applicable to a nation. Furthermore, the family or household is left out because any indicator that is applicable to an individual can be used with a family when it is aggregated in some fashion. These assumptions are used for simplification, and are not necessary for the validation of the model. For example, one could adopt a finer gradation by creating categories such as, the individual; the family; both individual and family; the community; the nation; and both community and nation. These six categories fairly exhaust the possibilities of the utility dimension.

The measurement dimension is composed of three steps or categories, based on observed data, based on self-reporting data, and based on both observed data and self-reporting data. By observed data is meant data obtained through observation of the subjects, not through personal or questionnaire interview of the subjects. On the other hand, all data obtained through the personal or questionnaire interview are considered self-reporting data.

It is recognized that a good interviewer obtains both self-reporting data and observational data, but these two types of data are easily differentiated and it still is possible to dichotomize the data into the two types for the purpose of the model. Where they cannot be easily differentiated, then the data are both self-reported and observed.

One important distinction must be made to prevent confusion. Scores on intelligence, aptitude, or achievement tests are observed data, not self-reported data. This is so because the performance of a subject on any of these types of tests is evaluated by some objective criterion or criteria. For example, if in an arithmetic test a subject gives the answer "four" to the question, "What is two and two," then we have the observed datum that he knows the answer to the question. Contrast the test with the questionnaire item, 'Do you know the answer to the question, "What is two and two"?' The datum in the form of a yes or no is self-reported because all we have is the subject's word for it. While this is an over-simplified example, it does accentuate the difference between genuine tests and so-called psychological "tests," such as the Minnesota Multiphasic Personality Inventory (MMPI) and the California Psychological Inventory (CPI), which

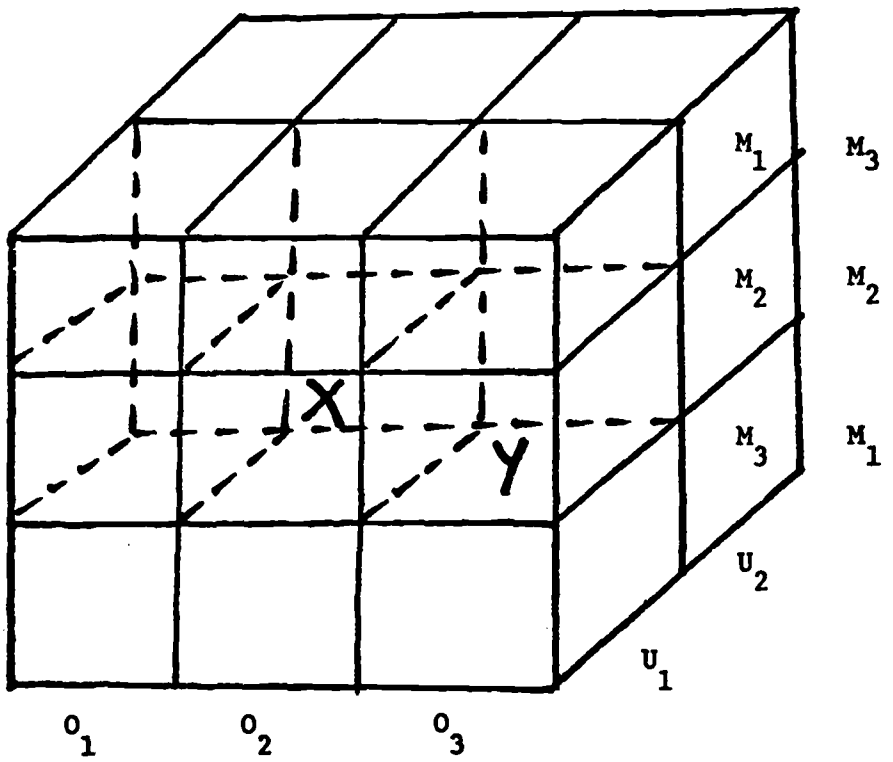
are no tests at all because they produce self-reported data rather than observed data.

The third dimension is orientation, using Baumann's classification scheme. Again, for simplification of exposition, only three categories are used: feeling state orientation, symptom orientation and performance orientation. To be exhaustive one would have to add all the possible combinations of the three categories, including feeling state and symptom orientation, feeling state and performance orientation, symptom and performance orientation, and feeling state, symptom and performance orientation. It is entirely possible that health indicators fitting all these categories will be constructed in the future, but the three categories should accommodate most of the indices extant.

With the three dimensions as described, the classification model can be geometrically represented as a cube, with M symbolizing the measurement dimension, U the utility dimension and O the orientation dimension. In the measurement dimension, M_1 or SR signifies self-reporting, M_2 or OB observations, and M_3 or SR-OB both self-reporting and observation. In the utility dimension, U_1 or IN signifies individual, and U_2 or NA community or nation. O_1 or FE represents feeling state orientation in the orientation dimension, O_2 or SY symptom orientation, and O_3 or PE performance orientation. This model is shown in Figure 1.

Figure 1

Schematic Representation of
Classification Model



To illustrate how this classification scheme is used, the cube OUM_{312}, marked Y, represents the cell into which fall all indicators that are performance oriented, that are used with individuals, and that are based on observed data. One of such indicators would be Katz' Index of Activities of Daily Living or ADL.(5) The Apgar Index for the Newborn (6), would fit in the X cube next to the Y cube, because it is symptom oriented, it is used with individuals and it is based on observed data. The total number of cells or categories is, in this particular case, 3 X 2 X 3 = 18.

The 18 categories are: OUM_{111}, OUM_{112}, OUM_{113}, OUM_{121}, OUM_{122}, OUM_{123}, OUM_{211}, OUM_{212}, OUM_{213}, OUM_{221}, OUM_{222}, OUM_{223}, OUM_{311}, OUM_{312}, OUM_{313}, OUM_{321}, OUM_{322}, and OUM_{323}. OUM_{111} would include indicators that are feeling state oriented, that apply to individuals, and that are based on self-reported data. OUM_{223} would comprise indices that are symptom oriented, that are applicable to a community or a nation, and that are based on both observed data and self-reported data. The other letter combinations can be similarly interpreted.

HEALTH STATUS AS PROGRAM OUTCOME

Assuming that a valid and reliable health status indicator has been developed, the problem of linking changes in health status with a health delivery system in a community still remains. This is so because any new health program implemented in a community constitutes only one type of input into an open system on which a multitude of known and unknown factors also impinge. To isolate the effect of the program from the effects of the confounding factors is an extremely difficult, if not impossible, task without experimental manipulations and/or controls. Since a new health program is usually unique in many aspects it is generally impossible to obtain a control comparable to the new program. Even if a comparable control were obtainable, it would not help much because consumers of services from the two programs could not be randomly assigned. Without randomization of consumers, it would be impossible to control systematic differences that might exist between the two groups and these systematic differences would then be confounded with differences in program effect.

The difficulties just described are, of course, not unique with health services research. They are confronted by all social scientists with an interest in the evaluation of social action programs. These difficulties

are discussed in considerable detail by Suchman.(7) Most researchers with some knowledge of experimental design and statistics are familiar with these problems, but in most cases they have to work in situations where they can do very little, if anything at all, about them.

On the other hand, there are some naive researchers who are undaunted by the complexities of measuring change in a social setting. These are the people who are not familiar with the classic, Problems in Measuring Change, (8) and take the cavalier view that a simple before-after design suffices in the evaluation of a health program. Statistically, they subtract the pre-scores from the post-scores and test the gain scores for significance. They are not aware of the fact that gain scores are notoriously low in reliability and that the use of gain scores, which Cox(9) terms an index of response, requires the assumption that the regression of the post-data on the pre-data is linear, with a regression slope that is unity. In most cases this assumption is not valid and considerable doubt is cast on the findings.

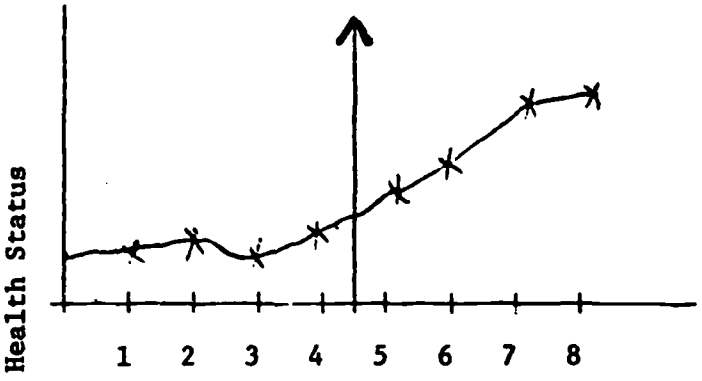
There are other ways of analyzing pre-post data, such as using the difference between the post-score and the regressed score, and using the pre-scores as the covariate in the analysis of the post-scores. While these procedures are an improvement over the simple analysis of gain score, the former are no more valid than the latter in establishing causality between a health program and the outcome measures. Whatever statistical procedures are used, a significant difference only means that a change has occurred; it does not automatically imply that the change is due to the health program.

STOCHASTIC MODELS IN OUASI-DESIGNS

Much more promising as a statistical procedure for measuring change over time is the application of stochastic processes to non-stationary time series data. Where a control group is not feasible, collecting data from the same group at regular intervals has the effect of using the same group as its own control. This design can be schematically represented as follows:

Figure 2

Hypothetical Representation of A
Non-Stationary Time Series



In this design the dependent variable is health status and the independent variable is the absence and presence of a new health program with the arrow symbolizing the division line. Collection of health status data is made at fixed intervals from the same sample eight times. Statistically, it is tempting to use the time periods, assigned some arbitrary values, such as 1, 2, 3 and 4, as the independent variable and regress health status on this variable, independently for the time periods before the introduction of the health program and for the time periods after the introduction of the health program, and then compare the two intercepts and the two slopes. This, however, would not be a legitimate procedure because of the problem of auto-correlation of the time periods.

For this type of data, the model and statistical techniques of Box and Tiao(10) appear the most appropriate. This is the integrated moving average model represented by the two equations:

$$z_1 = \alpha_1 \text{ and } z_t = L + \gamma \sum_{i=1}^{t-1} \alpha_{t-1} + \alpha_t \qquad (1)$$

for the n_1 observations before the introduction of a program and

$$z_t = L + \delta + \gamma \sum_{i=1}^{t-i} \alpha_{t-i} + \alpha_i \qquad (2)$$

for the n_2 observations following the introduction of the program, where:

z_t is the value of the dependent variable at time t,

L is a fixed but unknown location parameter,

γ is a parameter describing the degree of interdependence of the observed values of the dependent variable in the time series and takes the values $0 < \gamma < 2$, α_t is a random normal variate with mean 0 and variance σ^2, and δ is the change in level of the time series.

Inspection of (1) and (2) shows that for the pre series, it is composed essentially of random shocks, a proportion of which are accounted for by the non-independence of the time periods and assimilated into the level of the series. For the post series, a new parameter, δ, is introduced to account for change in level, presumably due to program effect. In either case the effects are linearly cumulative to and inclusive of the last time period.

While the logic of the model is extremely simple, the computations are not. This is so because the values of L and δ must be estimated from the data, using as the model in matrix notation: $Y = X\theta + e$, (3) where X is an N x 2 matrix of weights, θ a 2 x 1 vector with L and δ as the elements, and e an N x 1 vector of random elements with mean 0 and variance

σ^2. If the value of γ is known, the least squares estimate of δ is found by solving the least squares normal equations. In the case the value of γ is unknown, a Bayesian analysis using sample information about γ is performed to make inferences about δ. Then the estimated value of $\hat{\delta}$, which Box and Tiao have shown to have a $\underline{t}$ distribution with N - 2 degrees of freedom, is tested for statistical significance.

It should be noted that while these statistical procedures enable the experimenter to tell whether or not a real change in the level of the post series has occurred, they do not ipso facto establish causality between program and effect. It could very well be that simultaneously with the introduction of the program, one or more other events took place that had greater impact on the health status of the community than did the new program. One of these events could be the discovery of a new therapeutic procedure, a new "wonder drug," or a new diet for obese people. In the absence of a control series in parallel with the post series, which would have helped to rule out the effects of the extraneous events if they existed, one could accept causality with some confidence only if it were assumed that over the duration of the series no major events other than the new health program occurred to affect the pattern of the series. The validity of this assumption could checked by experts familiar with the health sciences and with the community where the health program was introduced.

References

1. World Health Organization, Constitution of World Health Organization, Annex I, in The First Ten Years of the World Health Organization. Geneva: World Health Organization, 1958.

2. World Health Organization, Measurement of Levels of Health. WHO Technical Report Series No. 137. Geneva: World Health Organization, 1957.

3. Chen, M. K., The measurement of health--issues, problems and approaches. Unpublished manuscript.

4. Bauman, B., Diversities in conceptions of health and physical fitness. Journal of Health and Human Behavior, 1961, 2, 39-46.

5. Katz, S. et al, Studies of illness in the aged: the index of ADL, a standardized measure of biological and psychosocial function. Journal of American Medical Association, 1963, 185, 914-919.

6. Apgar, V., A proposal for a new method for the evaluation of the newborn infant. Anesthesiology and Anatomy, 1953, 32, 260.

7. Suchman, E. A., Evaluation Research, Principles and Practices in Public Services and Social Action Programs. New York: Russell Sage Foundation, 1967.

8. Harris, C. W. (ed.), Problems in Measuring Change. Madison, Wis.: The University of Wisconsin Press, 1963.

9. Cox, D. R., Planning of Experiments. New York: Wiley, 1958.

10. Box, G. E. P. & Tiao, G. E., A change in level of non-stationary time series. Biometrika, 1965, 52, 181-192.