Issues of Design and Methodology in Long-Term Followup Studies

by Thomas H. McGlashan, William T. Carpenter, Jr., and John J. Bartko

Abstract

Beginning with Vaillant, followup studies of schizophrenia have devoted increasing attention to issues of design and methodology. The major advances and elements are described here and include: (1) conceptual framework, (2) design, (3) sample representativeness, (4) sample description, (5) data source and quality, (6) measures, (7) reliability, (8) diagnosis, (9) comparison groups, (10) assessment independence, (11) outcome, (12) missing subjects/bias testing, and (13) statistical techniques.

Followup studies of schizophrenia are plentiful. Zubin et al. (1961) reported on the existence of 800 such investigations by 1961. These reports have broadened the perspectives of investigators and practitioners, but it is difficult to place the results in a general context. Most studies have inadequate sample description, do not sample from a well-defined or representative population, lack comparative cohorts, and leave open key questions about the quality and reliability of observations. Generalizability is tentative at best. The studies of Vaillant (1962, 1964a, 1964b, 1978), with their careful attention to key prognostic factors, are an early example of advances in design methodology. Since then, followup studies have shown progressive attention to design and methods, resulting in greater specificity and meaning of the data. But, even today, studies fail to approximate the ideal closely. Key methodological issues are described in this communication.

The major elements of followup methodology are: (1) conceptual framework, (2) design, (3) sample representativeness, (4) sample description, (5) data source and quality, (6) measures, (7) reliability, (8) diagnosis, (9) comparison groups, (10) assessment independence, (11) outcome, (12) missing subjects/bias testing, and (13) statistical techniques.

Conceptual Framework

The first generation of long-term followup studies based on broad description has proved to be generously informative. These studies have furnished the outline of the syndrome and have provided the basis of hypothesis testing for the next generation of studies. Since prospective studies are expensive and time consuming, it is highly desirable to organize the effort to answer crucial questions. A clear appreciation of theory, concepts, and hypotheses allows the investigator to design the study to accomplish the most important aims and to minimize the effect of artifact and bias. Design decisions flow directly from hypotheses. If the investigator postulates that lack of affective disturbance at index episode predicts progressive enlargement of cerebral ventricles, a host of design issues come into focus. How fast is this progression? This will determine how early in illness the cohorts are established and the duration of followup. If affect is the key predictor, then the comparative cohorts must be similar in such crucial areas as duration of illness, insidious versus acute onset, and premorbid impairment. A priori decisions as to primary dependent variables are required to

Reprint requests should be sent to Dr. T.H. McGlashan, Chestnut Lodge Research Institute, 500 W. Montgomery Ave., Rockville, MD 20850.
ensure that the design will provide reliable and valid ascertainment. Where to impose blindness or independence of data collection depends on the hypothesis being tested. How many subjects are studied and many other crucial details must be based on the purpose of the study, or else run the risk of major flaws in design for each descriptive analysis.

**Design**

Followup studies are usually identified as prospective or retrospective in design. In the former, patients are selected as followup subjects at baseline (in the present) and then assessed one or more times in the future. The data can be collected under predetermined conditions allowing the investigator to maximize their quality and completeness. In the latter design, patients are identified as followup subjects, and their current functioning is assessed retrospectively, usually on the basis of archival material collected for other purposes. This strategy allows for very long followup intervals and forms the design for many of the studies reviewed in this issue. At a minimum, investigators need to specify rules for subject selection, characterize the larger pool from which the sample is drawn, and state the limits on generalizing study findings.

**Sample Representativeness**

The discipline of epidemiology has highly refined procedures for sampling and ascertaining the representativeness of resultant samples. Long-term followup studies, however, rarely fulfill the optimal criteria from this vantage point. It is important, therefore, that the investigator understand optimal sampling and provide a clear, critical evaluation of the limitations of the procedures used. Too often, highly influential conclusions have been derived from skewed patient populations. The long-held belief in a uniformly poor outcome in schizophrenia was based on samples drawn from chronic and deteriorating subgroups. Broader based sampling techniques have clearly demonstrated considerable variation in course and outcome, as is documented in several articles in the present issue. At a minimum, investigators need to specify rules for subject selection, characterize the larger pool from which the sample is drawn, and state the limits on generalizing study findings.

**Sample Description**

Baseline demographic and predictor sample characteristics frequently determine a major share of outcome variance. Furthermore, they are usually more powerfully correlated with course than are treatment interventions, especially over the long term (Stephens and Astrup 1965). Such characteristics also define the extent and limits of sample generalizability and the utility of the sample for cross-study comparisons. Therefore, adequate description is vital. How much description is enough may be difficult to judge, but investigators should err on the side of overinclusion. Age, sex, ethnic origin, socioeconomic class, age of onset, educational level, premorbid functional capabilities, past occupational functioning, and other such defining characteristics are essential to enable the generalizability and replicability of the sample to be evaluated. Both longitudinal descriptors (e.g., course of illness up to baseline) and cross-sectional phenomenology (e.g., signs and symptoms at baseline) are needed.

**Data Source and Quality**

Clinical research is data based, and the strength of that foundation relates to its completeness. In prospective designs, problems usually involve subject attrition, making it important to test for missing subsample biases (see below). In retrospective designs, the archival records, upon which the baseline ratings depend, are virtually always incomplete and of variable quality. This is so pervasive a problem that it is a prudent preinvestigational exercise to ascertain whether sufficient data exist to support the study at all.

**Measures**

What to measure should follow from the hypotheses under scrutiny instead of relying on measures of convention or convenience. If the hypothesis is that a certain type of course is associated with the development of anhedonia, then a clinical assessment instrument that enables the investigator to differentiate anhedonia associated with schizophrenia from anhedonia secondary to dysphoric mood or anhedonia secondary to drug side effects is needed. If the investigator seeks to determine predictors of anhedonia, then key measures will be specified by theory.

Is it also critical to have data on birth and pregnancy complications, neurological signs, or childhood asociality? This depends on the purpose of the study. If the investigator plans hypothesis-generating analyses, then the areas of interest must be defined and covered with
assessment instruments. However, the selection of multiple, broad-based assessment procedures with the intent of discerning the crucial relationships in a large data set is fraught with problems such as chance findings and the treatment of statistically significant correlations as valid findings before they are replicated in hypothesis-testing designs. Another problem with relying on an "all-purpose" broad evaluation is that key variables may be unwittingly omitted.

Reliability

Present-day research standards require interrater reliability to be established on primary data. Because of the time and training involved, reliability assessments are often omitted in followup studies. This is particularly the case when a large number of variables are assessed without preselection of the most important. A statistical analysis of rater agreement against chance should be tested and reported for all study variables that require substantial inference (e.g., premorbid social functioning) or that involve multiple decisions (e.g., diagnosis) or that are based on nonsystematic data sources (e.g., medical records). The methods of reliability testing selected depend on the type of data selected to test study hypotheses (Bartko and Carpenter 1976). The use of rating systems of demonstrated reliability is a good start, but it does not ensure their reliable application in the proposed study.

Diagnosis

Schizophrenia is a clinical syndrome, and neither specific disease entities nor unique diagnostic markers have yet been defined.

While specific sets of diagnostic criteria need to be chosen and explicated, it is also important that investigators ensure that patients' overall psychopathological presentations conform to the diagnostic class concept being used. Much outcome variance may be predetermined by diagnostic criteria, and whether narrow or broad, cross-sectional, or longitudinal perspectives are preferable will depend on the study goals. The ability to use several criterion sets in the same sample has the advantage of adjusting diagnosis to fit study questions. The simultaneous application of multiple sets of criteria (the polydiagnostic approach) (Berner et al. 1983, 1984; Brockington 1985) may maximize confidence in the diagnosis and facilitate replication of the cohort by other investigators, but may also skew the sample to a nonrepresentative core. Furthermore, many studies may be interested in subgroups of schizophrenia that could be eliminated by the requirement that multiple criterion sets be met.

It is also apparent, as will be elaborated, that samples should be characterized in their full array of psychopathology—not just "primary" or "study" diagnosis. Schizophrenic patients can also be depressed, addicted, agoraphobic, or schizotypal. Upon closer inspection, comorbidity is almost always present, and it can have powerful effects on many fronts: pathogenesis, treatment response, and long-term course, to name a few.

Comparative Groups

Whether nonschizophrenic comparative cohorts are required depends entirely on the study aims and hypotheses. However, most questions require some comparative basis for evaluating real effects of key variables. Very often the comparison is within schizophrenia. If followup data are used to ascertain the impact of neuroleptic dosage reduction strategies on deficit symptoms or incidence of tardive dyskinesia, then the cohort must be divided accordingly to reduced and nonreduced medication strategy. Considerable perspective on the consequence of schizophrenia can be gained by the use of comparison groups, either patients with other specified psychiatric syndromes and/or nonpsychiatric controls. The Iowa 500 (Tsuang et al. 1979) included both, and thus provided a richer perspective on the meaning of outcome data from patients with schizophrenia.

Independence of Assessments

In the hypothesis-testing framework, which seems timely for new longitudinal studies, it is often essential to design the study to ensure that knowledge of independent variables does not bias judgment of dependent measures (or vice versa). Computed tomographic measurements, for example, need to be done without knowledge of data used to predict ventricular enlargement. Approaches to minimizing bias include: (a) separate observers for key independent and dependent variables, (b) raters blind to study hypotheses, and (c) raters blind to classification status. Circumstances of any particular study will determine how to design assessment independence.

Outcome

Advances in the followup field usually mean more work for the investigator. Nowhere is this truer
than in the assessment of outcome. Unanchored global adjectives ("good, fair, poor") no longer suffice. They are, in fact, quite meaningless alone. The long-term course of outcome assessment has paralleled that of diagnostic assessment. Today, in studies approaching outcome, several psychometric issues require attention. Reliability has been mentioned. The others include scope, mechanics, accuracy/validity, and specificity/generalizability.

**Scope.** Outcome is multidimensional. It consists of several semi-independent domains (e.g., symptoms, social functioning, work functioning, treatment, and global functioning) which are only moderately intercorrelated (Strauss and Carpenter 1972, 1974), and which remain so even over the long term (Carpenter et al. 1987). Outcome also consists not only of "results" but also of "processes." Therefore, descriptions of course should document phenomena likely to induce change—for example, further treatments or important life events. Alternatively, these variables may be controlled in the study design. Failing to account for the impact of such variables can seriously undermine findings in comparative group studies, since they will often be disproportionately represented in the groups. Finally, outcome exists in time and may vary with inherent rhythms of natural history (McGlashan 1984b). Accordingly, followup interval length(s) and the ages of subjects at outcome should be selected in relation to the hypotheses.

**Mechanics.** The way followup information is collected can be important. Where do the data come from: the subject, significant others, mental health records, written correspondence, death certificates? Who does the collecting: mental health professionals, nonmedical assessors, trained versus untrained personnel? How are the data collected: telephone versus face-to-face encounter, structured versus unstructured interview, at the patient's home or at the study center? All of these may affect the quality and completeness of data. Suspected biases linked to the mechanics of data collection can often be estimated through testing—for example, comparing outcome profiles of the group of patients interviewed on the telephone versus the group interviewed in person (McGlashan 1984a; Carpenter et al. 1987).

**Accuracy/Validity.** The accuracy/validity of outcome data can be enhanced by sampling information from multiple perspectives. These include eschewing sole reliance on self-reports of outcome (McGlashan 1984a, 1984b) or checking a subject's report against that of a knowledgeable significant other (Harding et al. 1987a, 1987b). Outcome may also be described both longitudinally and cross-sectionally. While correlated (McGlashan 1984b), each perspective adds to the completeness of the followup picture, and marked inconsistency between perspectives raises questions about validity. Validity is most difficult to estimate, but a literature or pilot work cross-referencing the rating instrument with other criteria is often possible. Here we can only encourage the investigator to attempt to select external validation criteria for key clinical assessments. If the crucial dependent variable is based on clinical observation, what nonclinical variable would be highly related? If a clinician judges the status of cognitive function, could a laboratory-based test of cognition be used to "validate" the clinical measure?

**Specificity/Generalizability.** The more detailed and specific the descriptions of outcome, the more outcome can be generalized and compared across studies. Methods for maximizing specificity include using structured interviews, defining core outcome variables with explicit operational criteria, and incorporating relevant existing scales with psychometric credentials as part of the assessment battery. Rating scale anchor points should be clear and methods of combining variables (e.g., into global ratings) explicit (Tsuang and Winokur 1975). It is also useful to illustrate key rating scale anchor points (especially on global scales) with representative detailed case descriptions (McGlashan 1984b). Using identical scales with identical anchor points across studies does not guarantee identical calibration. Such variation cannot be eliminated, but it can be specified or made explicit by presenting random or representative raw clinical cases and how they were scored.

**Missing Data/Bias Testing**

Long-term longitudinal studies with large data collection procedures will necessarily result in missing subjects and missing data on available subjects. Subject attrition over time can lead to systematic bias, and missing data may reflect pathology-related factors. There are a number of methodological procedures that help the investigator deal with these problems. Bias testing procedures can assist in ascertaining the nature and extent of bias. For example, testing whether the missing
subjects are different from remaining subjects on baseline data will indicate whether the outcome sample is being skewed toward better or worse prognosis by attrition. In some circumstances it is better to estimate a data point than to treat it as missing. For instance, if monthly Brief Psychiatric Rating Scale scores are used to reflect symptom status over a 5-year period in a cohort of schizophrenic patients, the investigator may find that ratings are missing for 20 percent of the months. If missing data are treated as no symptoms recorded, then the course of illness will appear more favorable than if each data point is estimated. Estimations could be made by using an average of surrounding data points for each missing rating. In other circumstances, it is preferable to treat missing data as absence of the phenomenon. If the study is to determine predictors of visual hallucinations occurring late in the course of illness, it may be prudent to treat all missing data on predictors or hallucinations as absence. The study statistician will bring expertise on various approaches to missing subjects and missing data, but the clinician-investigator will have to make critical decisions based on concepts and hypotheses.

Analysis

The process of data analysis and interpretation requires intimate familiarity with statistical methods and clinical concepts. These are often treated as separate tasks. This split is always to the detriment of the study. Elsewhere in this issue we adumbrate the most commonly encountered statistical methods for longitudinal studies. We recommend that investigators strive to achieve the closest fit of concepts, design, data analysis, and interpretation. An intimate interaction between clinician and statistician from start to finish is crucial. The fit between concept and design, as well as continuity of collaboration, affords the greatest statistical and interpretive power to the research.

Conclusion

There are many cogent design and procedural issues that merit explicit attention by investigators. We provide a brief discussion of 13 such issues because the literature is replete with serious omissions in these areas. Investigators now are in a position to design followup studies to deal successfully with most of these issues or to estimate their impact when factors cannot be adequately dealt with in the study design. The field deserves the implementation of high scientific standards in future studies.

References


Berner, P.; Katschnig, H.; Simhandi, C.; and Kieffer, M.D. Diagnosis of schizophrenia. *Integrative Psychiatry*, 2:3–9, 1984.


Strauss, J. S., and Carpenter, W. T., Jr. The prediction of outcome in


The Authors

Thomas H. McGlashan, M.D., is Director of Research, Chestnut Lodge Research Institute, Rockville, MD. William T. Carpenter, Jr., M.D., is Professor of Psychiatry, University of Maryland School of Medicine, and Director of the Maryland Psychiatric Research Center, Catonsville, MD. John J. Bartko, Ph.D., is a member of the Theoretical Statistics and Mathematics Branch, Division of Biometry and Applied Sciences, National Institute of Mental Health, Bethesda, MD.

Announcement

The Medical College of Pennsylvania at Eastern Pennsylvania Psychiatric Institute, in cooperation with the Pennsylvania Office of Mental Health, is sponsoring a conference entitled "The 4th Annual Pennsylvania Conference on Schizophrenia: Families as Partners—Research and Treatment."

The Conference will be held in Philadelphia, March 30-31, 1989.

A highlight of the Conference will be the presentation of the annual Arthur Noyes Award to an individual who has been selected for his/her outstanding contribution toward the understanding, research, and/or treatment of schizophrenia.

For further information, please contact:

Conference Coordinator
Division of Continuing Mental Health Education
Medical College of Pennsylvania
3200 Henry Avenue
Philadelphia, PA 19129
(215) 842-4095