Presented at the Operations Research and Mental Health Services Research Conference, National Institute of Mental Health, Bethesda, Maryland, March 6-7, 1980.

STOCHASTIC MODELS IN MENTAL HEALTH EVALUATION RESEARCH

by

Richard H. Shachtman *
Eric J. Feuer

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1282
April 1980
### CONTENTS

<table>
<thead>
<tr>
<th>I. Introduction</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>II. Experimental and Quasi-Experimental Design in Evaluation</td>
<td>3</td>
</tr>
<tr>
<td>III. Stochastic Models</td>
<td>6</td>
</tr>
<tr>
<td>IV. Stochastic Modeling Perspectives in Evaluation</td>
<td>16</td>
</tr>
<tr>
<td>V. Implementation of Stochastic Models in Evaluation</td>
<td>20</td>
</tr>
<tr>
<td>A. Initial Consensus on Objectives</td>
<td>20</td>
</tr>
<tr>
<td>B. Variable Definitions and Reporting Requirements</td>
<td>22</td>
</tr>
<tr>
<td>C. Design/Statistical Planning</td>
<td>24</td>
</tr>
<tr>
<td>1. Pilots and Sample Size</td>
<td>24</td>
</tr>
<tr>
<td>2. Dependent Variable Induced Design Considerations</td>
<td>25</td>
</tr>
<tr>
<td>D. The Data Collection Process</td>
<td>27</td>
</tr>
<tr>
<td>E. Analysis and Modeling Building: Researchable Problems</td>
<td>29</td>
</tr>
<tr>
<td>1. Homogeneous Subgroups</td>
<td>29</td>
</tr>
<tr>
<td>a. The Problem</td>
<td>29</td>
</tr>
<tr>
<td>b. The Literature</td>
<td>30</td>
</tr>
<tr>
<td>c. Stable Baseline Transition Probabilities</td>
<td>32</td>
</tr>
<tr>
<td>d. Two Strategies for Developing Stable Baseline Transition Matrices</td>
<td>33</td>
</tr>
<tr>
<td>2. Partial Path Observations and Aggregate Data</td>
<td>35</td>
</tr>
<tr>
<td>3. Development of State Definition Structure</td>
<td>38</td>
</tr>
<tr>
<td>4. Higher Order Markov Dependence</td>
<td>39</td>
</tr>
<tr>
<td>5. Validation</td>
<td>39</td>
</tr>
<tr>
<td>VI. Recommendations</td>
<td>42</td>
</tr>
<tr>
<td>VII. References</td>
<td>45</td>
</tr>
<tr>
<td>VIII. Discussion at Conference</td>
<td>48</td>
</tr>
</tbody>
</table>
I. INTRODUCTION

The purpose of this paper is to indicate desiderata for the application of a certain class of operations research (OR) models, known as stochastic models, to mental health program evaluation (MHPE). Specifically, we will critique how a stochastic model might be used for MHPE including several important, and researchable, problems in implementing such models. It is not our intention to ignore or demean other OR approaches such as network theory, simulation, systems dynamics; we concentrate on one which we believe is both applicable and encompasses broad aspects of MHPE. Moreover, as so often happens in OR and systems analysis approaches, we believe that the very process of preparing to create such models will lead to successful evaluation approaches regardless of whether the output includes a formal model.

This paper is directed, generally, at a non-operations research reader. We ask this person's forbearance as we set the framework for our approach so that s/he may critique the potential usefulness and pitfalls of stochastic modeling applications to MHPE. In Figure 1 we indicate the triangle of communication necessary for the actual implementation of stochastic modeling techniques in a specific health service agency. Our discussion concentrates on the arcs emanating from (I).

| I Statisticians/ | II Clinical Staff - |
| Operations Research | Psychiatrists, |
| Analysts, Epidemiologists, | Psychologists, etc. |
| Economists | |

*Figure 1. Communication for Modeling in MHPE*
We encourage and depend on communication with clinical staffs (II) to help solve the problems of defining appropriate measures for program performance and patient functioning and mental health administrators (III) to help set specific program objectives, evaluate feasibility of measures and solve the problems of data acquisition and storage.

This paper's concentration on criteria and problems in applying stochastic models is intended to build, quickly, from a general framework to special problems for MHPE. We briefly review points on experiments and quasi-experiments (Section II). Subsequent to outlining the general approach to stochastic model building in Section III, we proceed with brief points on evaluative measures and a few stochastic models for MHPE in the literature (Section IV) before proceeding to the longer Section V on the implementation of stochastic models. Recommendations are listed in Section VI.
II. EXPERIMENTAL AND QUASI-EXPERIMENTAL DESIGN IN EVALUATION

The fundamentals of classical experimental design (Cochran and Cox (7)) laid down by Sir Ronald Fisher in the early parts of this century were formulated to solve problems in agricultural experimentation and industrial quality control. In the early 1950's, the discovery of the psychotrophic action of certain drugs in the care of mental disorders opened the doorway for clinical trials and later evaluation studies in mental health (Cole and Gerard (8)). The classical designs were naturally and quickly accepted for these applications; however, many investigators feel that these models of experimental design, which worked well in agricultural experimentation and industrial quality control, do not meet the research needs of clinical psychology and psychiatry. Instead it is through the "development and application of more dynamic statistical systems and design concepts" (Chassan (4)) that these requirements can be fulfilled.

The standard classical design used in evaluation studies is the analysis of covariance (ANOCOVA) with randomized treatment assignment. This design uses a post-treatment score, such as a clinical assessment on a standardized scale, as a dependent variable to test if two or more treatments differ significantly after adjusting for the pre-treatment score and other possible confounding covariates. The basic problem with this design in mental health applications is that two fixed point observations (pre- and post-treatment scores) are insufficient to measure the dynamics of a person's, or group's, mental processes. The key is that we do not simply want to measure the change in a fixed score or value representing the mental condition of a person, but the CHANGING PATTERNS OF CHANGE. To accomplish this we must use sequences of observations instead of single fixed point observations.

In quasi-experimental designs /1/ ANOCOVA is still commonly used although there must be guards against threats to internal validity because of the lack
of randomization into treatment groups. Researchers often use this type of design when evaluating competing programs since ethical considerations and informed consent requirements may preclude randomization into programs not equally applicable to all subjects. This lack of randomization into programs may cause biases resulting from (1) differential selection of respondents for the comparison groups and (2) the interaction effects of selection biases and the experimental variable (post-score) causing what is termed the initial group difference problem (Campbell and Stanley (3)). These biases break an assumption of ANCOVA because we are not comparing groups which were initially comparable.

Sometimes matching is used to pair comparable individuals on the pre-score, as well as other relevant covariables, and individuals who cannot be adequately paired are removed from the analysis. Because subjects within a particular program interact, however, subjects excluded from the analysis can still affect the outcome measurements of those remaining in it. For example, suppose we compare two classroom programs for the mentally retarded in which the subjects could not be randomized because the two school programs are aimed at slightly different, although overlapping, initial ability levels. The experiment's purpose is to compare improvement levels in the two programs. In order to make the groups comparable, we match the students with the highest ability level in the lower group with those with the lowest initial ability level in the higher group. (See Figure 2). It is not inconceivable that the lower ability levels of the unmatched students in the first group would slow the progress of those matched in this group, while the higher ability levels of the unmatched students in the second group might hasten the progress of the others in their group. That is, observations are not independent and the nature of the patient mix affects the outcome. The net effect of these biases might cause us to incorrectly conclude that it is the actual teaching program
which is causing the matched students in group 2 to improve more than the matched students in group 1, while it is actually the differential patient mix.

![Figure 2. Comparison of Two Groups of Mentally Retarded Students with Overlapping Ability Levels](image)

Thus the mental health researcher has the task of developing analytic tools which take into consideration the problems of confounding effects, the non-independence of observations, and, perhaps most importantly, the need to reflect change patterns.

/1/ Quasi-experimental design is used "in many natural social settings in which the research person can introduce something like experimental design into his scheduling of data collection procedures (e.g., the when and to whom of measurement), even though he lacks the full control over the scheduling of experimental stimuli (the when and to whom of exposure and the ability to randomize exposures) which makes a true experiment possible." (3)
III. STOCHASTIC MODELS

A stochastic model is a mathematical, and sometimes computer-based, representation of sequences of probabilistic events over time. That is, it is an abstraction of empirical phenomena which is useful for description, predicting outcomes, making statistical inferences and testing the effects of changes in, and to, the process. Thus, such models may be used to compare and evaluate alternative interventions and decisions. Different types of stochastic models use different definitions of events occurring during the process and the time interval during which they occur. One basic dichotomy is between models that use continuous or discrete events and continuous or discrete time parameterizations. One class of stochastic models is known as Markov chains (MCs) – for these models one assumes that the probability of the next transition (between states or to the same state) depends only on the current state and not on prior ones (Feller (11)). A more general class – known as semi-Markov processes – allows more flexibility in the modeling of holding times in states. (See step (8) below.) We give an MC model example for anxiety and depression patterns at the end of this section and review several MC examples for MHPE in subsequent sections.

As a prologue, we suggest nine basic steps in constructing a stochastic model, aimed towards the derivation of Markov or semi-Markov processes. (Shachtman and Hogue (28); Kastner, Shachtman and Guade (18).) Although these points may be evident to the experienced OR analyst, a surprising number of papers do not follow the stages or utilize the statistical tests available for testing the concomitant assumptions. This section provides a brief definitional statement and overview of these steps while Section V considers the special problems inherent in their application to mental health evaluation.

The steps in constructing a stochastic model are
1. Determine homogeneous subgroups of the sampled population for baseline and comparison analyses.
2. Characterize the type and extent of events missing from individual and group observations and other data problems.

3. Develop the structure of state definitions for particular subgroups.

4. Select a discrete or continuous time parameter and either an appropriate time interval or intensity function.

5. Estimate transition probabilities and standard errors.

6. Test for time-homogeneity (stationarity).

7. Test the order of the Markov property.

8. Determine the length of stay (holding time distribution) for each state.

9. Develop validation possibilities.

1) DETERMINE HOMOGENEOUS SUBGROUPS OF THE SAMPLED POPULATION FOR BASELINE AND COMPARISON ANALYSES

In order to perform a stochastic analysis, we must split the sampled group into separate subgroups with respect to characteristics related to the dependent (event) variable. This is extremely important to the mental health modeling process and will be discussed in detail later.

2) CHARACTERIZE THE TYPE AND EXTENT OF EVENTS MISSING FROM INDIVIDUAL AND GROUP OBSERVATIONS AND OTHER DATA PROBLEMS

The steps we outline here, while in a specific order, may require feedback from and iteration with other steps. These requirements obtain more frequently when there are either data bits missing from each path observation (e.g., missing Community Mental Health Center (CMHC) visit information or outside Community Support Program information on a patient) or the data comes only in some aggregate form. The latter may occur if, for example, we know only the proportion of patients to visit a state. We elaborate on this in Section V.
3) DEVELOP THE STRUCTURE OF STATE DEFINITIONS FOR PARTICULAR SUBGROUPS

The "state space" measures the dependent (event) variable over time. Some stochastic models use a discrete state space to characterize events such as the location of a psychiatric patient (hospitalized - unhospitalized, etc., Marshall and Goldhammer (22)) or a discretized scale of psychological well-being (Rodda, Miller, and Bruhn (26)). Others might employ a continuous state space to measure a cardinal scale of psychological well-being (Whitmore and Neufelot (36)). While the discrete state space is appropriate for many health service applications, the exact nature of the states used is intimately tied to other considerations of the modeling process. (See Section V.E.3.)

4) SELECT A DISCRETE OR CONTINUOUS TIME PARAMETER AND EITHER AN APPROPRIATE TIME INTERVAL OR INTENSITY FUNCTION.

The continuous time parameterization is the most natural choice for interpretation although the modeling process and statistical estimation simplify if we restrict transitions among states to discrete intervals. The discrete time reference frame chosen must be short enough to allow for transitions to be made and long enough so that the Markov property, see point (7), is not severely violated. For acute psychological states, a natural break might be a day, week, or month depending on the nature of the psychological variable. In mental deficiencies with long latent times or remission/relapse sequences, time intervals may be a year or more. Most models allow consecutive stays in a state. Many practical situations dictate a particular time frame chosen on the basis of available data. The interpretation of results for discrete time stochastic processes, especially by program managers, will probably be easier than working with an intensity function necessary for a continuous time parameter stochastic process. Another reason for using a continuous time parameter, occasionally, is to save computation costs; in both mental
health and other health services examples known to the authors, this does not seem an insurmountable problem.

5) ESTIMATE TRANSITION PROBABILITIES AND STANDARD ERRORS

A Markov chain is a discrete event, discrete time, stochastic process. This model is often the most practical for simple modeling considerations although the discrete time semi-Markov process may be very useful; see point (B).

The definition of an MC given above (the probability of the next transition depends only on the present state) is called first order Markov dependence. We display this dependence in a model for a three state MC as a transition probability matrix.

<table>
<thead>
<tr>
<th>STATE</th>
<th>WILL ENTER NEXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>p(1,1) p(1,2) p(1,3)</td>
</tr>
<tr>
<td>2</td>
<td>p(2,1) p(2,2) p(2,3)</td>
</tr>
<tr>
<td>3</td>
<td>p(3,1) p(3,2) p(3,3)</td>
</tr>
</tbody>
</table>

The element \( p(I,J) \) presents the probability of going from state \( I \) to state \( J \), \( I, J = 1, 2, 3 \). To estimate each \( p(I,J) \) in the matrix, one first sums, over all observations, the number of times a subject goes from state \( I \) to state \( J \), \( n(I,J) \), divided by the number of times a subject leaves state \( I \), \( n(I) \). Then \( p(I,J) \) is defined to be \( n(I,J)/n(I) \).

Once we estimate a transition probability matrix, we may employ a wealth of analytical results for estimation of hypothesis testing of movement among the states of the MC. From a first order MC, analytical results permit the computation of the standard errors of the transition probabilities, the expected time to return to a state, the probability of going from state \( I \) to state \( J \) in time \( N \) (N-step transition probabilities), the probability
of being "absorbed" in a state from which there is no exit (e.g., death, discharge), the mean time to absorption, etc. Anderson and Goodman (1) and Kullback, Kupperman, and Ku (20) present procedures for testing, for example, whether the transition matrix differs significantly from a specified matrix, the homogeneity of several realizations of an MC and corresponding results. Thus, these results yield inferential techniques to capture changing patterns of change in the subject population.

There are, however, serious gaps in our analytical knowledge. For example, as far as we are aware, there is no closed form, direct analytic technique for calculating standard errors of N-step transition probabilities. We have developed a simulation technique for estimating these standard errors and applying them to health program data which may be used for mental health program evaluation (Schoenfelder, Shachtman, and Johnston (27)). Standard errors are at least as important here as they are in any type of statistical work, especially for establishing confidence limits around transition probabilities which have direct programmatic interpretations; unfortunately, they are rarely reported.

6) TEST TIME HOMOGENEITY (STATIONARITY)

The assumption of time homogeneity requires that as subjects mature, their transition probabilities do not change. This can be tested by splitting the path codes (i.e., sequence of states that a subject occupies) at various intervals and testing the homogeneity of the matrices estimated from the subgroups as indicated in step (5). For example, suppose subject path codes (sequence visited in time order) run for three years. We can test the time homogeneity by estimating a transition matrix based entirely on first year data, second year data, and third year data, and then testing if these transition matrices differ significantly.

This assumption may be violated by the very mental health processes
which we wish to evaluate. The effects and suggestions for handling them are discussed in Section V.C.2.

7) TEST THE ORDER OF THE MARKOV PROPERTY

First order Markov dependence may not hold for the empirical process under study. The latter may satisfy a more general Nth order Markov dependence. This occurs when the probability that one next enters a certain state depends on the sequence of the last N states. To reflect Nth order Markov dependence requires an $N + 1$ dimensional transition matrix. Obviously the data requirements, and possibly computational effort, for even a moderate $N$ preclude estimation of higher order matrices. Anderson and Goodman (1) and Kullback et al. (20) provide tests that determine if data come from a chain of specified order. Frequently, one may capture the order properties by expanding to a first order MC with many more states; some states in the latter process may reflect a sequence of states in the original $N$th order MC. Health service examples occur in Thomas (33) and in (28).

8) DETERMINE THE LENGTH OF STAY (HOLDING TIME DISTRIBUTION) FOR EACH STATE

An implicit assumption of first order Markov dependence (i.e., a chain is memoryless before the present state) is that the length of stay in a state has a geometric distribution. This distribution characterizes probabilities of waiting in a state. At each time period the process may stay in the state ("failure"), with a probability of $q$, or leave the state ("success"), with a probability of $p$, where $p + q = 1$. The probability of leaving after $n$ time periods is then $q^n p$, $n = 0, 1, \ldots$

If the assumption of a geometric length of stay distribution cannot be met, then the empirical process should be modeled as a semi-Markov process. In the latter case it is necessary to calculate frequency distributions of length of stay in each state called holding time distributions. Often,
for computational efficiency, the analyst fits an analytic distribution
to the frequency distribution for holding time and estimates parameters so
that all manipulation can be accomplished with fixed, well-known analytic
distributions. Frequently, mixtures of geometric distributions are used for
the discrete case and mixtures of exponential distributions are used for the
continuous case.

9) DEVELOP VALIDATION POSSIBILITIES

In developing a stochastic model, we need internal as well as external
validation criteria. Internal criteria include validation checks on elements
within the data set. For example, we might test how well the derived N-step
transition probabilities match the empirical ones; see Shachtman et al. (29)
for an example. External validation is done in terms of how well the model
estimates known quantities, such as well-established prevalence rates. This
topic is further discussed in Section V.E.5.

Many papers do not follow all of the above stages. One of these, however,
does provide a quick review of stochastic modeling with an MC as a concrete
example. In their article, Rodda et al. (26) attempted to model
the differential anxiety and depression patterns among 31 male patients with
coronary heart disease and 46 healthy male controls. For exemplary purposes
here, we consider the anxiety transition matrix for young subjects (i.e., under
fifty years of age). To measure anxiety, the authors used a 10 question sub-
set of the Bendig Manifest Anxiety Subscale of the Minnesota Multiphasic Per-
sonality Inventory (MMPI); see Dahlstrom and Welsh (9). Subjects answered
true or false to ten questions, such as "I get nervous if I sit for very
long" and "I am usually worrying about something." They received a composite
score from 0 to 10 and this scale was categorized to form three states: low
anxiety (0-3), medium anxiety (4-7), high anxiety (8-10). The subjects were
measured an average of 7 times each over a period of 15 months. From this path data on each individual, the following separate transition probability matrices were estimated for young controls and young patients.

<table>
<thead>
<tr>
<th>ANXIETY LEVEL</th>
<th>LOW</th>
<th>MEDIUM</th>
<th>HIGH</th>
<th>YOUNG CONTROLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW AT TIME N</td>
<td>0.946</td>
<td>0.053</td>
<td>0.000</td>
<td>LOW</td>
</tr>
<tr>
<td>MED</td>
<td>0.217</td>
<td>0.739</td>
<td>0.043</td>
<td>MED</td>
</tr>
<tr>
<td>HIGH</td>
<td>0.000</td>
<td>0.200</td>
<td>0.800</td>
<td>HIGH</td>
</tr>
</tbody>
</table>

The first observation, that the controls tended to remain at lower anxiety levels than patients, could probably be determined from fixed point observations and classical experimental designs. One can also see that controls have a much higher probability of remaining at a medium level of anxiety or moving to a lower level. Coronary patients, however, seem to have a higher probability of bouncing between high and low anxiety levels. This is a transitory phenomenon and is recognizable only through sequences of observations.

Among the advantages of these stochastic models over linear models like ANOCOVA is the ability of the MC to produce time-dependent results and other parameterizations (functions of the transition probabilities). In particular, assuming the stationarity of the transition probabilities over time, we can derive expected results which are comparable to those obtained from a classical design. If at time 0 we have 10 people in each of the three categories (pre-score measurements) then we can derive the expected number in each category after a reasonable amount of time — say 5 time intervals. Analytic results mentioned earlier, Section III (5), permit the computation of a 5-step (in general N-step) transition matrix directly derivable by taking the fifth (N-th) power of the one step transition matrices given above.
ANXIETY LEVEL AT TIME N + 5

<table>
<thead>
<tr>
<th>ANXIETY LEVEL</th>
<th>YOUNG CONTROL</th>
<th></th>
<th></th>
<th></th>
<th>YOUNG PATIENTS</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOW</td>
<td>MEDIUM</td>
<td>HIGH</td>
<td>LOW</td>
<td>MEDIUM</td>
<td>HIGH</td>
<td>LOW</td>
<td>MEDIUM</td>
</tr>
<tr>
<td>AT TIME N</td>
<td>0.837</td>
<td>0.145</td>
<td>0.013</td>
<td></td>
<td></td>
<td></td>
<td>0.451</td>
<td>0.287</td>
</tr>
<tr>
<td></td>
<td>0.593</td>
<td>0.321</td>
<td>0.081</td>
<td></td>
<td></td>
<td></td>
<td>0.192</td>
<td>0.274</td>
</tr>
<tr>
<td>HIGH</td>
<td>0.252</td>
<td>0.377</td>
<td>0.369</td>
<td></td>
<td></td>
<td></td>
<td>0.083</td>
<td>0.254</td>
</tr>
</tbody>
</table>

These matrices give the probability of being in state J at time 5 conditioned on starting in state I at time 0. For example, the expected number of patients in the medium state at time 5, conditional on the low starting state, is simply the number that started in L at time 0 multiplied by the probability of being in state M at time 5. That is, expected number = (10)p(L,M) = (10)(0.287) = 2.87 for coronary patients.

We calculate these expected numbers for all combinations of starting states and states at time 5 to yield:

Table 1 Expected Number in State at Time 5 Starting State

<table>
<thead>
<tr>
<th>State at Time 0</th>
<th>State at Time 5</th>
<th>Expected Number in State at Time 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW(10)</td>
<td>LOW</td>
<td>Young Control</td>
</tr>
<tr>
<td></td>
<td>MEDIUM</td>
<td>8.37</td>
</tr>
<tr>
<td></td>
<td>HIGH</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.13</td>
</tr>
<tr>
<td>MEDIUM(10)</td>
<td>LOW</td>
<td>5.93</td>
</tr>
<tr>
<td></td>
<td>MEDIUM</td>
<td>3.21</td>
</tr>
<tr>
<td></td>
<td>HIGH</td>
<td>0.81</td>
</tr>
<tr>
<td>HIGH(10)</td>
<td>LOW</td>
<td>2.52</td>
</tr>
<tr>
<td></td>
<td>MEDIUM</td>
<td>3.79</td>
</tr>
<tr>
<td></td>
<td>HIGH</td>
<td>6.19</td>
</tr>
</tbody>
</table>

The data in Table 1 are comparable to post-score measurements from the analysis of covariance with the measurements at time 0 representing pre-scores. From these data, a comparative set of interpretations can be derived for both the stochastic and classical analyses. An area for further research could
involve determining whether stochastic modeling based on several distinct transition matrices is superior to classical analysis.

This type of analysis suggests the applicability of Markov models when measuring transitory phenomena, although it doesn't begin to plumb the analytic alternatives for evaluations subsequent to estimation of the transition matrix.

/2/ Note that the row sums are not one. This is due to the computational roundoff in the original transition matrix which is compounded when taking successive powers of the matrix.
IV. STOCHASTIC MODELING PERSPECTIVES IN EVALUATION

In this section we point out two Mental Health Program Evaluation (MHPE) problems which stochastic modeling attempts to resolve, two types of MHPE measures and a brief introduction to four MHPE papers using stochastic models. Details on the latter appear in Section V.

Stochastic models for program evaluation in mental health are especially applicable because of the large amount of variability present in the psychodynamic processes of individuals. These models solve many of the problems inherent in the use of experimental and quasi-experimental designs. The model utilizes sequences of observations (rather than fixed point observations) in order to reflect patients' psychological transitions. Chassan (5) comments on how statistical models should capture the psychological processes of individuals: "not only is the patient to be described in terms of a multidimensional or multivariate probability distribution, but the study of change in the patient-state, or the evaluation of a patient's progress, or lack of progress, in the course of time in relation to any program of therapy, is then to be performed in theory by comparisons of individual variability. Such variability is the basis of defining the patient state in terms of statistical distributions, or probabilities, estimated from sequences of observations of a given patient, and consequently provides the framework for the application of statistical inference and experimental design to the data of the individual patient." We expand on Chassan's comment by emphasizing the aforementioned objective; we wish to make inferences about patterns of change, not just levels of change. We should achieve this by analyzing models for groups of "similar" patients.

In addition to variability, patient interaction may violate the independence of observations assumption in classical designs and cause problems such as the matched pairs analysis comparison mentioned in Section II. These problems are especially troublesome in hospital ward settings where patients
have close and continual contact. Using a stochastic view of modeling, patient interactions are part of the psychopathological process and should be included instead of excluded from the study. A transition probability matrix should be framed around the natural mental health unit, a ward, producing a complete statistical picture reflecting patterns of patient interaction. There are still analytic problems in this area because estimation of transition probabilities assumes independence of observations; however, this type of complete ward analysis could start to provide the parameters of patient interaction (i.e., patient diagnosis mix, patient severity mix, number of patients in the ward, etc.).

There are two types of dependent variables commonly used in evaluation studies in mental health. We characterize them as program performance measures and patient functioning measures. Program performance measures focus on the utilization patterns of the patient in the mental health system. They may reflect inter-institution movements (for example, Reuter and Young's (25) study uses such states as: walk-in-unit, evaluation unit, general clinical division (hospital), individual psychotherapy, couples-family therapy, in-patient CMHC, outpatient CMHC, etc.) or intra-institution movements (for example, Eyman et al. (10) use states such as: in hospital-enrolled in school program, in hospital—not enrolled in school program). Program performance measures like the latter may also be referred to as patient location since they emphasize facility or type of service. Note that there are additional program performance measures including service intensity (number of hours, visits) and labor intensity (transportation services, variation due to length of time—e.g., extra time for outreach services).

On the other hand, patient functioning measures focus on the direct modeling of patient condition and behavior over time—the psychodynamics of the illness. Since mental illnesses vary so widely in their manifestations (i.e., the range includes schizophrenia and other psychoses such as anxiety,
depression, paranoia, retardation, organic brain disease, alcoholism), a wide range of patient functioning measures are applicable. Many standard measures are available, e.g., the MMPI, although patient specific measures, have recently become popular. The problem with the latter type of measure is that patient aggregation is difficult; yet aggregation to similar groups is necessary for any overall evaluation of patient functioning.

The choice of the dependent variable differentiates the evaluation studies using stochastic models. Trinkl (34) and Meredith (24) use patient functioning to model the effectiveness of different programs using an MC model. Trinkl uses a combined scale, rating independence in the living environment and independence in the working environment, to develop a value-weighed MC. In order to obtain benefit-cost ratios, he compares the subjective benefit of the transitions made in each of five programs to the costs, whereas Meredith uses an unweighted average of scores on eight self-help scales (ambulation, arm-hand use, toileting, expressive communication, receptive communication, feeding, bathing and dressing) which are conducted annually to obtain matrices for five different programs for the mentally retarded.

Eyman et al. (10) and an earlier article by Meredith (20) use patient location as the dependent variable to develop their models. Eyman et al. compare a matched set of mentally retarded children; one group attended a day school program as part of their hospitalization and the other group attended industrial therapy classes. The analysis followed the two groups for four years after initial hospitalization with respect to their location status (in hospital—dropped from school or industrial therapy class, in hospital—attending class, released from hospital on indefinite leave). From this data they develop separate transition matrices for the day school and industrial therapy groups. Meredith uses an MC model to show that a geriatric resocialization program (GRP) is more effective from a cost standpoint than a geriatric ward in placing patients outside the hospital.
Location states for this model include the GRP, the ward, placement outside the hospital, and death.
V. IMPLEMENTATION OF STOCHASTIC MODELS IN EVALUATION

In this section we lay the groundwork necessary for identifying and describing problem areas in the use of stochastic models for MHPE. Subsequent research efforts should have the solution of these problems as a high priority objective. We employ the basic framework of Section III and use the aforementioned papers, other health services research, and our research to illustrate the perspective, problems and results.

The process of constructing a mathematical model as part of an OR study, whether for program evaluation or other use, involves a lengthy series of stages and feedback, testing, implementation and evaluation before the modeling efforts are complete. It is not within the scope of the present paper to elaborate on each of these subprocesses and interpretations of results. We have selected several stages taken from a version of a Management Science/Operations Research Problem-Solving Process by Shachtman (30). We comment only on those items which we believe have immediate implications as problem areas for MHPE modeling.

A. Initial Consensus on Objectives
B. Variable Definition and Reporting Requirements
C. Design/Statistical Planning
D. The Data Collection Process
E. Analysis and Model Building: Researchable Problems

A. INITIAL CONSENSUS ON OBJECTIVES

This first point derives from one author's experience with an organization whose mission is to monitor and evaluate mental health program services in a large urban area; yet, the finding underlines a general evaluation criterion independent of either MHPE or OR modeling. Namely, we observed
the effect of competing, or at least non-conjugated, program objectives. In particular, there was a disparity between the target population of some of the programs and the client categories actually being served. Initial funding for the mental health services came from the state level which has structured objectives for its mission involving a spectrum of services. Somewhat different objectives emerged, however, from the operational experiences of city officials and, a third group, community level organizations responsible for direct contact with patients and clients. The differences became clear when we reviewed state eligibility requirements based primarily on the frequency and time proximity of lengths of stay in state institutions. City and community agencies soon discovered that locating, contracting, and convincing "eligibles" to participate in these programs was not an easy task. Yet, the agencies were able to find other classes of individuals who needed the services they were equipped to offer; this was especially true for outreach (outpatient) programs. In another type of institutionalized program, individuals soon discovered tricks for satisfying eligibility definitions sufficient for admission, even though they did not meet the spirit of the criteria.

In short, the analyst/evaluator is faced with a complex situation: should models be constructed whose outputs satisfy all the sets of objectives so that these outputs can be used to compare the competing objectives? Is the analyst/evaluator required to become the catalytic agent influencing officials from two or three different agencies to derive common program objectives? In many cases, if those responsible for both monitoring and evaluating program data do not take these objectives into consideration, the results of their output will be deemed to be inadequate or non-targeted even if they meet the criteria for one of the agencies.

Thus MHPE, along with OR analyst input, must deal with settling on program objectives and the conjugation of these objectives among several
groups.

B. VARIABLE DEFINITION AND REPORTING REQUIREMENTS

Model construction may not begin until program directors reach consensus on objectives, the choice of evaluation approach and the subsequent selection of variables keyed to objectives and evaluative measures. If program management pilots a monitoring/evaluation project over six months or a year, there may be time to redefine these variables, but data collection processes are sufficiently difficult to require much thought when mounting the program. The OR analyst should be an integral part of the process of objective setting and variable definition since s/he may have suggestions leading to better efficiency in the data collection design system—the necessary predecessor to efficacious analysis and model building.

Concomitant to these considerations is the choice among several types of evaluation and the specification of reporting requirements necessary to the design of the monitoring system. Program performance measures generally include items having to do with the amount of hours spent on different types of program visits (case management, clinic visits, on-site rehabilitation), social center visits, outreach efforts, sheltered workshops, day treatment, coping skills, competency and other services (homemaker, housekeeping, transportation) and they usually summarize numbers served, time spent, number of cases handled in different ways, etc.

On the other hand, patient functioning measures emerge so that there is an attempt to reflect outcomes for the patients. For some patient types, these reflect diagnoses, individual abilities to perform simple physical tasks, behavior problems, and other environmental measurements as well as occupational abilities. Physical health is an important factor. Note that these are only examples and may vary widely depending on patient and condition.

Later these variable definitions and reporting requirements lead to the
state space definitions for stochastic models; examples of resulting definitions are given in Section V.E.

Finally, we would like to mention a special issue in psychological measurement which appears to lend itself readily to the stochastic modeling approach. We may refer to it as the short term/long term change pattern. There are apparently two basic types of psychological instruments that can be utilized as patient functioning measures. Personality assessment scales (e.g., the MMPI) measure the average psychological level of a person over a certain time interval, which is usually not made explicit. The averaging is subjectively assessed by subjects through the use of questions containing phrases like "do you usually ...." or "when I encounter a certain situation, I ....". The second type might be called mood monitors (e.g. Mehrabian-Russell Semantic Differential); these measure the present psychological frame of mind of the person and, thus, are less subject to selective recall than personality assessment scales. Theoretically, the patterns of moods over a certain time interval determine individual personality traits. For example, quick swings of depressive and manic moods determine the manic depressive personality disorder.

Because of the small number of observations in classical designs, the subjective averaging is necessary. That is, the analyst does not want to record a measurement for a subject when s/he is in a particularly "rare" mood. This is especially true if the analyst has access to only one or two observations. However, if we are truly going to measure the changing patterns of change, the subjective averaging process will tend to lose information. Stochastic models, because of the numbers of observations on similar individuals, have the capacity to measure patient functioning at its most basic level—the mood. We suspect that most of the stochastic models using patient functioning as their dependent variable still use personality assessment scales. Frequent observations are required for short-term measurements; nevertheless,
the payoff in tracking specific psychodynamic movements may be worth the effort and might not be expensive if performed as an adjunct to services delivered on a regular basis. Mood measures, therefore, represent another area of potential research.

C. DESIGN/STATISTICAL PLANNING

1. Pilots and Sample Size

Once there is selection of an evaluation design (e.g., program performance, patient functioning, combination) and related variables (measures of effectiveness like those mentioned above), statistical planning should ensue, including protocols for data collection and sample size estimation. Theoretically, these might occur concurrently with the piloting of program services delivery. Frequently, however, startup to a full level of services is rapid, and quick decisions on data collection are made, often with emphasis on reporting requirements for the budget office. In these cases, programs may not be collecting the data most valuable for future evaluation.

The first task in selection is the definition of the appropriate populations for data collection and, where needed, sample size estimation. For reporting requirements, especially to the funding agency, a program often summarizes population statistics, such as number served, totals of types of services, etc. These may encompass large groups of individuals with distinct subsets of mental deficiencies. In the latter case, if program management is interested in estimates of certain prevalence or incidence measures (e.g., the proportion of clients exhibiting certain levels of anxiety or depression and changes in these patterns), statistics must be computed from samples. The latter provokes a need for sample size estimates. With the resulting sampled data in hand, the analyst may estimate prevalence/incidence of patients in a broad target population suffering from the deficiency of interest.
2. Dependent Variable Induced Design Considerations

To illustrate certain concerns with design, we use stochastic modeling examples, although the design considerations are of general concern.

Clearly, the choice of design used depends on the choice of dependent variable under investigation, especially as it relates to a type of evaluation. The studies that use patient condition as a dependent variable, (34, 24), develop transition matrices representing the course of a patient's stay in a particular program. These transition matrices for distinct programs are then compared. One assumption of Markov modeling is the time homogeneity (stationarity) of the transition probabilities – an assumption that the transition probabilities do not change over time. The purpose of a successful program, however, is to change the transition probabilities over time. For example, many "normal" individuals have brief periods of depression. A transition matrix may reflect this by representing a simple pattern of flow among levels of depression, with corresponding transition probabilities. For a group with a higher propensity for longer, or "deeper," periods of depression, the effect of a successful program may be to modify the transition probabilities so as to result in a pattern with shorter, or "shallower," depression occurrences. The MC at the end of Section III, although it deals with anxiety, provides an example. In that case the coronary patients exhibited a propensity for "deeper" periods (patterns) of anxiety. The key consideration here is that programmatic efforts, especially direct services, contain interventions (one-on-one meetings, group conferences, prescription of drugs) which affect these very probabilities. Thus, one is not necessarily able to obtain a static picture of patient outcomes so that comparison with alternative means of delivering services are observable.

An alternative design establishes a baseline transition matrix which describes and characterizes patient functioning before, or in the early stages of, the program; a program effect matrix is developed from
data in the latter parts of the program or just post-program completion. The latter design is especially applicable when using a patient functioning variable which tends to have a cyclic, or seesaw, property over time – one example is depression. We measure the program effect only in terms of the changing patterns of the cyclic process. Sometimes expert subjective estimates make up the corresponding transition probabilities when empirical data are not available, e.g., (34). For the coronary patient example, we may have to test the effect of an intervention to reduce anxiety by modifying some of the transition probabilities and looking at patterns of change in long term stable behavior or other measures from powers of the modified matrix. For example, suppose we subjectively assume an increase in the probability of staying in the low anxiety state, \( p(1,1) \), for the coronary patients; we may assume that \( p(1,1) \) is 0.9 rather than 0.812. Then we may compare the probability of being in that state after five time periods, the \( p(1,1) \) element from the fifth power of the modified transition matrix, to the corresponding element from the fifth power of the original transition matrix, to examine the effect of the subjective assumption. It is with these types of dependent variables, with which we wish to measure the changing patterns of a transitory phenomenon, that the clear superiority of stochastic rather than single fixed point pre- and post-program measurements is apparent.

Employing program performance measures as dependent variables in a program evaluation alters the focus of a study. These variables may effectively measure costs and utilization, but remain, at best, proxy variables for patient benefit. For example, although Meredith (23) computes only the long-term relative costs of the ward versus the GRP, he assumes implicitly that sooner deinstitutionalization of the GRP patient is beneficial; i.e., if someone is discharged from the hospital, s/he must be better. This is a dangerous assumption especially since the modern trend towards early
deinstitutionalization might make premature release feasible. There must be measurement of conclusions reached concerning patient benefit using a patient functioning variable unless the validity of patient location as a proxy variable is well-established. This is not to say that evaluation studies should categorically avoid patient location. In many instances, especially in the care of the severely ill, a program's purpose involves the maintenance of patient status rather than improvement; the gains can come through lower costs and better utilization of services.

Early evaluations of psychopharmacological agents used patient condition indicators as dependent measures—studies involving specific drugs, dose response, patient type, effect on the patient. With the increase in deinstitutionalization, emergence of community mental health centers and concern with other aspects of community support programs, the focus necessarily turned to global evaluation of interlocking programs. For these applications, patient location measures have become easier to obtain and are often considered to be more reliable.

We may question whether global evaluation, especially if it implies the use of only patient location measures, is the most appropriate choice for evaluating the impact of community support and other mental health programs. With the increased emphasis of consumer presence on CMHC boards, informed consent for patients or clients and public clamor for accountability, there may be a requirement for patient condition evaluation for specific subcohorts in addition to other measures.

D. THE DATA COLLECTION PROCESS

In this section, we briefly touch on a very involved issue—the data management of large data bases. This includes protocols for coding, collection, conversion to machine-readable form, automatic and manual editing, further testing and creation, and maintenance of backup systems.
See Jampol, Shachtman and Burns (16) for further details on a large study. Many mental health programs are of sufficient size to justify computerization, but the importance of this stage of preparation for analysis is under-emphasized.

The OR analyst should be well aware of the intricacies of data definition, its accessibility, its reliability with respect to content as well as with respect to coding in the collection process, the quality of the edit system and any quality control systems meant to check on data processing efforts. Our experience with large databases, even when derived from a carefully controlled institution-based study (345 hospitals with detailed patient data), has taught us that lack of care in considering these issues can severely affect the quality of model building for evaluation efforts.

An underlying problem with many evaluation analyses is that data collection proceeds with other intentions (usually for patient records or monitoring purposes) and the evaluation analysis is designed to "fit" the data. This problem could be acute in the collection of path data for stochastic processes. Mental health service systems usually collect extensive data upon entry and departure along with data on the various services used during the stay. Outpatient clinics probably keep less encounter records because there is no continuum of patient contact. An analyst may often derive path coded patient location or patient utilization data from these records, but systematic patient condition data collected at periodic intervals may be a rare commodity. Record linkages between various health service systems are even rarer. There are, however, several progressive new record keeping systems in the process of development. Honigfeld and Klein (15) developed a data system which produces "fever charts" showing the patients' week to week progress with respect to relevant patient condition variables. Nonetheless, clinicians were not enthusiastic about utilizing these standard forms. Instead, a chart review technique could be developed similar to
medical chart review techniques designed to extract information from hospital records for epidemiologic studies. This technique, along with additional protocols for chart entries, would insure sufficient data bits for such "fever charts." We have some experience with a targeted chart review technique which yields inference on infections acquired in the hospital; validation analyses suggest that it is successful. (See Haley et al. (14) and Whaley, Guade and Haley (35).) In another example, Bloom (2) has developed a patient tracking rather than service component oriented data base.

Missing linkages and event information also pose research problems for the estimation process in stochastic processes; see Section V.E.2.

E. ANALYSIS AND MODEL BUILDING: RESEARCHABLE PROBLEMS

We discuss some of the stages of stochastic model building, with concurrent criteria, from the task list given in Section III. Our objective is to point out several areas that merit caution in construction and application of the models, while noting that they also provoke researchable problems. Some initial results for these problems are available.

1. Homogeneous Subgroups
   a. The Problem

As early as 1959, Greenhouse and Kramer (13) identified the need to obtain a homogeneous group of patients for valid evaluation studies in mental health. Today, especially with the effects of deinstitutionalization, the problems of obtaining a fairly homogeneous group with similar mental deficiencies, at one point in time, is even more acute. One assumption for MC analyses is that the transition matrix for a typically selected individual in the study (theoretically obtained by following that individual over an infinite amount of time) should approximate the transition matrix obtained from a conglomerate of people measured for finite amounts of time; this is known as person-to-person homogeneity. Clearly, the probabilities of a patient moving from
state to state, for example those representing functional levels, may vary with patient age, previous history of institutionalization, characteristics of the outside support system, and other factors which are not necessarily reflected in the state structure of the model being used. We refer to this as the "heterogeneity" problem.

If we cannot solve the heterogeneity problem in a reasonable manner, then the possibility of generalizing results from these sorts of models is severely hampered. The key to this problem is that when one develops separate transition matrices for comparing two programs statistically, significant differences between the two transition matrices (e.g., representing two distinct programs) may be more a function of an imbalance of relevant or covarying characteristics rather than the actual difference between the two programs. Therefore, at a minimum, the analyst must develop suitable measures of what program managers are willing to consider as homogeneous populations.

b. The Literature

The reviewed articles, which apply stochastic models to program evaluation, attempt to deal with the heterogeneity problem. Meredith (23) is cognizant that each patient's probabilities of movement through the GRP may depend upon his/her individual diagnosis, the GRPs overriding consideration for selection, and the patient's amenability to resocialization. He points out that the last characteristic automatically results in the selection of a group of nonsenile patients who can generally be described as "burnt out" (a medical term referring to the deterioration of the original illness over time and the consequent substitution of institutionalization symptoms). Due to this fact, the patients all show a surprising similarity; thus he treats them as one homogeneous class with identical transition probabilities "regardless of whether an individual patient was originally, for example, a schizophrenic, female, or an alcoholic male." He makes no attempt, however, to prove this
point by estimating the transition matrices for groups of patients with different diagnoses.

Eyman et al. (10) compare school and non-school groups. They admit that there is no way of determining which changes in the school group are due to the program itself, over and above the factors involved in selecting patients for the program. They match by age and IQ to reduce the initial group difference problem. However, because patients within a group or ward interact with each other (thus affecting each other's mental state), a different patient mix can cause invalid results even when comparing homogeneous subgroups from each of the programs – recall our example in Section II.

Meredith attempts to obtain homogeneous subjects by selecting individuals who are profoundly and severely retarded (IQ strictly less than 35), non-physically handicapped, and under the age of 15. Because the hospital's treatment programs are not meant to apply to the full ability range of children, there is a severe initial group difference problem.

Trinkl (34) notes the inability to obtain enough data for estimating transition matrices if these matrices were estimated separately for each homogeneous subgroup. He states that "it could be argued that a finer classification of states reflecting health status, degree of handicapping condition, age, and other socioeconomic conditions, is important to obtain well-defined transition probabilities and, consequently, an immensely larger number of states is appropriate. If this approach were taken, conventional cost-effectiveness could be applied to evaluate specific provisioning of services for homogeneously defined groups of mentally retarded adults, but it would have been exceedingly difficult to evaluate alternative system approaches." Trinkl obtains separate subjective estimates of the transition probabilities from several expert groups for each of the alternative programs; this technique could be extended to deal with heterogeneous subcohorts.
c. Stable Baseline Transition Probabilities

The problems presented seem to point to the necessity of establishing stable baseline transition matrices against which we can measure change. In establishing a stable baseline we must know what group of relevant characteristics or symptoms defines a particular mental illness. This brings up the definitional problem in mental health so vividly illustrated by the study in which video-taped sessions with patients shown to English and American psychiatrists yielded widely differing diagnoses. (UK: Kendall, Cooper, Gourlay, Copeland; US: Sharpe, Gurland (19)).

Old theories of psychotherapy based on traumatic experiences in one's past dictated a deterministic interaction with events in the present and future. Modern psychotherapy is based more on the influences of day-to-day experiences and, according to this model, the way in which one will react to experiences in the future can be effectively modeled by a die loaded in directions influenced by the past - our transition probabilities. In the modern terminology one does not say that a patient is compulsive or has phobia, but that there is a certain probability that someone will act compulsively, or have a phobic reaction given a certain situation. The probability of a certain reaction depends on the psychological state one is presently in, as well as the particular situation encountered. These probabilities determine the severity and nature of the illness (4).

According to these concepts, different transition matrices could be formed for different sets of potentially relevant characteristics until transition matrices were found that were stable over different studies and situations. These transition matrices would abstract defining characteristics of a particular mental illness and could provide stable baselines for later evaluation studies.
d. Two Strategies for Developing Stable Baseline Transition Matrices

There are two basic design strategies for developing stable transition matrices. We may conduct an "intensive design" by "following" a few (or even one) subjects over a long period; the matrices may evolve from real or subjectively estimated data. Those who feel that mental processes are so variable as to preclude large scale studies or global evaluation advocate intensive designs. One advantage of such studies is the possibility of tailoring dependent psychological measures to the small group of people in the study. Also, because very few subjects are under study, the sample can be made homogeneous with respect to many relevant characteristics.

We might raise the objection that although any results obtained are certainly valid, they can be extrapolated only to the subpopulation which consists of people with identical characteristics to the people in the study - a very small population. Strictly speaking, this is true, but Chassan (6) believes that decisions on how to treat a given patient, based on previous experience with other patients, do not necessarily demand that our given patient's relevant variables all be identical to the preceding patients'. We base decisions on similarities rather than complete identities. If patient characteristics are completely specified, the results of one study may lead logically to the formulation of its successor. Thus the patient parameters under study can gradually be broadended to include a wide range of patients.

The "extensive design" appears to be more applicable to evaluation studies: we conduct such studies by trying to choose relevant variables to form homogeneous subgroups from a large sample of the population.

The first method of achieving this goal is to form separate transition matrices for each level of possibly relevant discretized covariates, combining matrices that are not shown to be statistically different. Rodda et al. (26) use the Bendig Manifest Anxiety subscale and Welsh Pure Depression subscales
of the MMPI (9) to develop stable transition matrices for 31 male patients with coronary heart disease and 46 healthy male controls. They match each patient with a healthy control on age, sex, race, education, occupation and height and weight. Previous research has shown age to be a factor, so separate matrices were developed for young (under 50) controls, young patients, old controls, and old patients on each of the two subscales; see the example at the end of Section III.

There are several problems with this approach; one is the obvious problem with the quantity of required data to which we have previously alluded. Another is what might be termed the "identifiability" of the groups; see (33), for example. In this problem, one is faced with determining how to stage conditions or diseases. Can we use a single state or are multiple states necessary? Do we need multiple chains? Several approaches compromise between trying to reflect all of the variables under consideration in separate models (chains) for separate groups versus combining groups into a single model. See Shachtman et al. (29) who construct separate subchains for aborters of a first pregnancy and deliverers of a first pregnancy but combine states for reproductive behavior past the first delivery.

Other techniques for overcoming the heterogeneity problem are directed at adjusting for covariates. One is Spillerman's (32) which is based on multiple regression. Reuter et al. (25) discuss this and point out that Spillerman's work is simply a generalization of the separate cohort approach to the extreme where sub-cohorts contain individuals. It is clear that in this case only a single transition is predictable since the prediction of subsequent transitions requires knowledge of the new distribution of characteristics in each state at each time period. We would have to have these predictions for each of the individuals involved or, in a slightly more general case, for each of the number of groups. The question is how might we convert this approach to represent appropriate heterogeneity without
severe computational costs and data requirements. Reuter et al. go on to develop an alternative approach to the population heterogeneity problem.

A third method for achieving homogeneously defined transition probabilities and covariate adjusted transition probabilities is still in its early stages of theoretical development (see Fertig, Murthy, and Sposto (12)). This technique has potential as it would allow for continuous covariates and could be used more easily in a modeling sense than the standard technique of forming separate transition matrices for every level of discretized covariates.

2. Partial Path Observations and Aggregate Data

All studies must deal with the problem of missing observations; for example, there may be variables in a hospital study concerned with a rare disease and unusual therapy so that there are few observations for those particular variables. This type of missing information is a function of the scarcity of occurrence of condition/therapy pairs. Another type surfaces when there may be many potential observations (e.g., the numbers of a relatively common surgical procedure) but, for some reason, a large proportion of hospitals have no observations of the procedure. Are they truly missing, or does our study "miss" them because of design inadequacies?

The nature of this problem takes on a different form when we consider (1) stochastic models and (2) certain event occurrences which we would observe for mental health patients, especially for patient functioning measures. To illustrate this problem, consider the following programmatic concerns and an artificial example. With the increased trend in deinstitutionalization, the recent increase of community support programs and the linking of other social support services to clinic or institution based services, many more variables emerge as we construct event definitions in a mental health patient path. (Some researchers refer to these as micropaths; see Lee, Judge, and Zellner (21).) A path, theoretically, would contain the psychological status
and program encounters of every patient at every time period being measured. In some community support programs, we may be able to track a patient from an institution to certain clinic visits, but, for many patients, we may "miss" information on homelife/friends stimuli, nature of activities, occupational contacts and influences and other institutional supports not directly connected with facility-based mental health care services. Thus, for example, a CMHC worker may derive some of this information from a careful interview and, possibly, link other institutional information, but may find very little information on some of the variables for some patients. We refer to this type of missing data as the partial path observation problem. This problem appears to be more significant for MHPE than in our experience with hospital based studies (18) and clinic/home based studies in population (28,29).

Therefore, procedures for estimating patient status and progress, as well as the effects of intervention, are absolutely necessary. This is all the more poignant when we wish to measure time-dependencies and pattern changes. Clearly no model can help supply information when there is no data, or expert opinion, on which to build the model. Frequently, however, stochastic models have the advantage, as with parametric models for statistical inference, that with a little data of decent quality one may estimate some parameters and test different patient flows and outcomes with respect to different interventions. This is clearly demonstrated in Trinkl (34) and other papers which are able to employ standard techniques of sensitivity analysis on transition probabilities even when the original transition probabilities did not come from a data base. (Sensitivity analysis on parameters means that the analyst varies an input parameter, such as a relapse rate, and measures the effect on an output measure, such as the eventual prevalence of institutionalized patients.)

We illustrate our concept of partial path observations, and the need for estimation techniques, by using an artificial example. Consider a set of
states which include patient status through events occurring to patients who are institution-based, clinic-based and derived from intensive interviews for events which cannot be observed in either type of facility. Suppose that we have numbered these states, which may include patient functioning measures, from S1 through SN. We illustrate below the paths of patients visiting some of these states over time for M patients. The star (asterisk) represents a missing event observation for patient 1 at time 2.

<table>
<thead>
<tr>
<th>Time Periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4 5 6 10 11 12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patients</th>
<th>1</th>
<th>S5</th>
<th>*</th>
<th>S5</th>
<th>S2</th>
<th>S2</th>
<th>.....</th>
<th>S3</th>
<th>S3</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>S5</td>
<td>S7</td>
<td>S5</td>
<td>S4</td>
<td>S4</td>
<td>S2</td>
<td>.....</td>
<td>S3</td>
<td>S3</td>
<td>S3</td>
</tr>
<tr>
<td>.....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>.....</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3 State Occupancies by Patients Over Time

The objective is not to impute a state occupancy for patient 1, e.g., by guessing that it is S5 because patient 1 is in S5 at times 1 and 3. However, we may be able to estimate additional transitions for a cohort of patients including patient 1 which "replace" his/her unknown transitions, (S5 -> *), (* -> S5) during periods 1 through 3. For example, patients 1 and 17, along with other patients, may be statistically similar and the * occupancy count may be replaced by S7.

Another type of incomplete observation may result from knowing only
the proportion of a cohort which visit given states at a given time. This is known as the aggregate data problem; Lee et al. (21) summarize several schemes for estimating Markov parameters from aggregate time series data. Shulman and Shachtman (31) suggest several schemes and an approach for aggregate data for semi-Markov processes.

Since MHPE analyses are even more likely to encounter the partial path observation and aggregate data problems, the need for useful statistical procedures here is evident.

3. Development of State Definition Structure

Redefinition of the state structure often occurs to more closely satisfy other Markov model properties, for example, time homogeneity and the Markov property. Sometimes one may also use redefinition to control for some of the heterogeneity problems. This device was employed in (28) to separate reproductive events for aborters and deliverers of first pregnancies as indicated above. We iterated state definitions until statistical tests showed that the assumptions were acceptable. Similar definitions might be made for mental health patients after an extended period of working with their first levels of diagnosis and treatment in the system. A combined set of states might be used for "maintenance" patients. For another health example dealing with the redefinition problem, see the MC formulation in (33) and the later semi-Markov version in Kao (17).

A separate issue in defining states is the determination of which events occurring for a mental health patient help to define a state. For example, if our objectives require patient functioning measures and we define states with levels of anxiety or depression, we may need interim measurements from notes made by the psychiatrist or a social worker. A definition, perhaps algorithmic, would have to be made to translate the assessments made by these professionals into states of anxiety or depression. Although we have not been involved in this process for mental health problems, one
author helped construct definitions of reproductive events and their eventual aggregation into state definitions for the aforementioned research with abortion and delivery of first pregnancies. If the number of patients flowing into a CMHC or other program is large, there would have to be an unambiguous set of definitions so that a computerized algorithm could translate patient events into patient states.

4. Higher Order Markov Dependence

Theoretically, using a higher order MC, or an expanded first order MC, solves the Nth-order problem referred to in Section III. However, from both the heavy data requirements and a computational viewpoint, this is usually an impractical solution. A compromise is to collapse some information into single states, for example allowing certain states to represent several combinations of the psychological status of the patient, while not trying to represent all attributes. An example of this occurs in (34); Trinkl derived four levels of living environment and five levels of work environment yielding, after deletion of three inappropriate states, 17 combination states. This tactic seemed to be successful for that particular application.

Sometimes allowing a second order property to be satisfied may be acceptable, as occurred in the reproductive events defined for the time to delivery estimates in (29). In such cases, one may be able to prove, statistically, that the first and second order properties are the same, justifying a first order MC. In general, a certain amount of cleverness in designing state spaces will overcome these difficulties.

5. Validation

This is a difficult topic and well deserves a paper itself. There are several approaches to validation which are independent of the modeling processes used; for example, we may construct a model on a randomly selected subset of the database and compare it to one we construct on a disjoint subset. With MC and semi-Markov models in particular, one may often compute
transition matrices for various subsets and compare the matrices statistically. Another type of validation, although it must be carefully interpreted, is to compute functions of the transition matrices and compare these functions, not only with each other, but with certain empirical calculations of incidence or prevalence rates (see 29, 27). We found no examples of this in literature which the authors reviewed for mental health analyses, let alone program evaluation. In fact, validation of models in health services problems is rare. Often availability of data is one of the problems. Also it may be that the structure of the chain is acceptable to the decision makers and there is no data with which to compare model outputs; this occurred in Trinkl's work.

One should be very cautious about the type of validation procedure used. For example, suppose we estimate two transition matrices, A and B, which describe movement for mental health patients. We may wish to statistically compare these matrices to see if the movements are the same. Furthermore, one may be interested in certain functions of these matrices, f(A) and f(B), to calculate whether some particular "interesting" set of states (pattern of mental deficiency) is visited; perhaps states in which the patients are "doing better." Figure 4 displays the possibilities with our own assessment of the intuitiveness of the combinations.

<table>
<thead>
<tr>
<th>Transition Matrices</th>
<th>f(A) = f(B) statistically</th>
<th>f(A) ≠ f(B) statistically</th>
</tr>
</thead>
<tbody>
<tr>
<td>A = B statistically</td>
<td>I Intuitively expected to occur</td>
<td>II Non-intuitive but can occur</td>
</tr>
<tr>
<td>A ≠ B statistically</td>
<td>III Non-intuitive but can occur</td>
<td>IV Intuitively expected to occur</td>
</tr>
</tbody>
</table>

Figure 4 Equivalence Combinations Between Transition Matrices and Their Functions

All combinations I, II, III, and IV can occur. An example of III for a non-
mental health problem is given in (29), in which the two transition matrices, A for the aborters and B for the deliverers are different; thus their movement behavior is different. Yet the functions of the transition probabilities which calculate time to the next delivery of a pregnancy are the same:

\[ f(A) = f(B) \] statistically. Moreover, the latter two calculations were also the same as the actual frequencies of these times observed in the two databases. In this particular case, we were able to validate the calculations of time to delivery while showing the basic movements were different. Hence, validation techniques can be of great usefulness in interpreting the results of these movements but must be applied very carefully.

There are several instances in the MC evaluation papers in which the mathematics of stochastic processes are correctly applied but validation through external criterion is required. The most flagrant example of a case to which this criticism applies is Meredith (24). He develops a transition matrix based on the data from two year measurements (i.e., the transition matrix was based on one transition only from the 1970-1971 data). From this transition matrix a 1970-74 four year transition matrix is computed by taking the fourth power of the transition matrix. Meredith states that this matrix “may be used not only for individual patients but also to ascertain the expected future distribution of the initial cohort.” This is a case of extreme extrapolation beyond the end points of the data. Certainly the transition matrix might change over the course of four years, especially when the initial matrix is only based on one transition. These results should be validated on the basis of either prior substantive knowledge concerning the time homogeneity of the chain or four year path code data collected on at least a subset of the individuals in the study.
VI. RECOMMENDATIONS

1. Objective Setting

The OR analyst, as an evaluation team member, should participate with program managers and funding agency officials in the consolidation of potential objectives; the group should reach a consensus on those for which monitoring and evaluation will be performed.

2. Development of Data Systems with Monitoring/Evaluation in Mind (whether or not model-building is involved)
   a. Data collection and management, although tedious, may require more time than that spent on modeling for MHPE; routine needs for data management, including editing, processing and redefinition, should not be overlooked.
   b. Patient functioning variables should be recorded as standardized measures at regular, periodic intervals.
   c. Patient tracking/record linkage across health service systems is difficult to achieve but, even if only available on subcohorts, is well worth performing. (I.e., a few, high quality observations lead to useful models; this is a statistical power concept for Markov models which is not unlike the concept of power for parametric statistical inference.)
   d. Computerized data base systems should be established, if not already in existence.

3. Evaluative Measures
   a. An area with great potential for research and implementation is the linking of a sufficient battery of program performance and patient functioning measures; these must exceed budgetary reporting requirements and allow the type of evaluation which we have emphasized here, namely the ability to detect changing patterns of change. This type of research will necessarily involve interdisciplinary teams including psychiatrists, psychologists, analysts
and program managers, with experience in the accessibility of key variable information, for the construction of the scales.

b. Given the advantages of stochastic modeling, more use of patient functioning in addition to program performance measures should occur.

c. Thus, we should study the possibility of using psychiatric records to obtain patient functioning levels on a more regular basis, perhaps by chart review techniques.

d. We should investigate the possibility of mood monitors in addition to personality scales.

4. The Heterogeneity Problem

This is a significant area for research and several directions may be feasible.

a. We should investigate the use of stable transition matrices to help define and differentiate mental illnesses.

b. We should support basic statistical research in the area of covariate adjustment in stochastic processes.

5. Incomplete Observations

Another key area is the research needed to overcome the partial path observation and aggregate data problems.

6. Parameterizations

As a special point, we note that although researchers have constructed stochastic models and performed simple analyses with the resulting transition matrices when these models are Markov chains or semi-Markov processes, they have not tapped the potential for further analyses. That is, one may construct many functions of the transition probabilities to meet specific objectives. We have no examples of this for mental health but have been able to create such functions for several other examples (28, 18). In addition to what we we can learn from performing sensitivity analysis on the transition matrices, the same sensitivity analysis may be performed on functions of the transition
matrices to yield even more information about the pattern change behavior for mental health patients.

7. **Subjective Estimates/Expert Opinion Input**

Finally, we should not undervalue the potential to apply subjective parameter estimates (e.g., transition probabilities), especially when dealing with the heterogeneity problem and other problems where direct observation is not possible or is too expensive. Trinkl (34) was able to use this approach with some success and other health services applications have been able to utilize these techniques.

The data requirements for MC and semi-Markov model building may not be so expensive. The assumptions of the models, plus the strengths indicated in points 2c, 3, 6 and 7 above suggest that these approaches are cost-effective for mental health evaluation research.
VII. REFERENCES


16. Jampol R, Shachtman RH, Burns RS. A data management system for the SENIC project. NN 74B of the SENIC Project, Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill, August 1979.


VIII. DISCUSSION AT CONFERENCE

During the discussion at the conference, Dr. Ciarlo argued that rather than short, quickly changing psychological states, the slower, longer-term transitions are more related to program goals. Recently, these include such states as: being productive or holding a job, being less withdrawn and more communicative, having positive supportive relations with other persons, participating in community life, being able to continuously care for oneself in terms of life's necessities and being free of the debilitating effects of drug or alcohol use. In these cases, he feels that the traditional pre-post assessments are as effective as the more complicated and expensive stochastic models.

Even if program goals consider only slower, longer-term transitions, we argue that they should also be concerned with long-term patterns of change measurable by MC and semi-Markov models. These need not be more expensive; cost is a function of which data are being collected - some data may be available through existing instruments and interviews. Moreover, the stochastic models may require a smaller number of more completely specified patient observations, which may not cost much more than a larger number of less completely specified observations.

Furthermore, we believe that in many cases long-term changes can be partitioned into short-term patterns of behavioral frequencies. For example, a classical analysis might yield data indicating that a certain program has made an individual generally more communicative and less withdrawn. This data would probably be obtained through observations made in a diagnostic interview before and after treatment. Alternatively, we could measure the frequency of certain specific withdrawal related behaviors which would allow us to pattern changes in these frequencies using an MC. In another case, we could model the patterns of unemployment and productive employment rather than just noting that a patient was becoming a more productive and stable employee. Certainly behavioral
frequencies are a more direct (and hence probably more reliable) measure of a person's functioning than a clinical interview. The eventual usefulness of such direct and frequent measurements, however effectively they may be utilized in stochastic models, needs to be ascertained.

Dr. Sondik suggested that too much emphasis may be placed on constructing the "perfect" model for determining causal mechanisms. He proposed that we should take a macro view and concern ourselves with an overall research agenda for MHPE implementing Dr. Joseph Holey's idea of "evaluability assessment." In this way, we may not lose sight of the underlying decision problem by over-emphasizing the search for the best design.

We feel this supplements our recommendations for more detailed evaluation models. There must be a balance between the overall evaluation approach, perhaps even at a purely conceptual level, and an operational evaluation approach - regardless of whether either approach uses models. Although, there are cases for which models may not be cost-effective, they may be employed for both types. For example, Dr. Sondik suggested that many federal agency decisions emerge from pragmatic evaluations - mount a program or not, terminate it or not, replicate it or not.

Coincidentally, we just completed a modeling effort, see /3/, to decide on exactly this type of question. In our study we built a decision analysis model, using both objective and subjective (Bayesian) estimates, to determine whether to perform a large complex medical survey. The type of model uses a decision tree with some corresponding probabilistic and economic parameterizations. This is a stochastic model, albeit not an MC. It has the property of encoding our knowledge and indicating to decision-makers where to "draw the line" in a decision.

Thus, we agree with Sondik's proposal that the Bayesian approach be integrated into the process and, further, that this might allow for a deterministic version of MCs - or flow models. A good example of a model
which is data-based, but could have been subjectively estimated, and yields a
flow model is given in (29). A "research agenda" use of this model is
suggested there, namely, to estimate sample size needs for cohort studies —
clearly a pragmatic decision for agencies which perform and fund research.

/3/ Shachtman RH, Decision analysis assessment of a national medical study.