8-1995

A Model for Simulating Life Histories of the Elderly: Model Design and Implementation Plans

Douglas Wolf
Syracuse University

Jan Ondrich
Syracuse University

Kenneth G. Manton
Duke University

Eric Stallard
Duke University

Max A. Woodbury
Duke University

ISSN 1084-1695

Papers in Microsimulation Series Paper No. 1

A MODEL FOR SIMULATING LIFE HISTORIES OF THE ELDERLY: MODEL DESIGN AND IMPLEMENTATION PLANS

Douglas Wolf, Jan Ondrich, Kenneth G. Manton,
Eric Stallard, Max A. Woodbury, and Larry Corder*

Maxwell Center for Demography and Economics of Aging

Center for Policy Research
Maxwell School of Citizenship and Public Affairs
Syracuse University
Syracuse, New York 13244-1090

August 1995

This research was supported by grant number R01 AG11815 from the National Institute on Aging.

ABSTRACT

This paper provides a strategy for the development of a model of life-cycle change in functional status, economic well-being, and family composition, with particular attention to persons aged 65 and older. The overall goal is to use the model as the basis for individual-level projections of the later life cycle, that is, microsimulation. Specifically, the scope of the project includes:

1. Specification and estimation of equations for the dynamics of functional status, nursing home occupancy, income and death among those aged 65+, using data from the 1982, 1984, and 1989 National Long-Term Care Survey (NLTCS) linked to Medicare data for 1982-1993, based on extensions of the Grade of Membership (GoM) framework;

2. Developing equations for year-to-year income streams, determined jointly with changes of marital status, for all ages represented in the cohorts to be simulated;

3. Estimating parameters governing the dynamics of family composition (existence and characteristics of spouse, parent[s] and child[ren]);

4. Integrating the results of the above modeling efforts in a microsimulation computer program with the capacity to dynamically simulate life histories, focussing on the elderly population;

5. Validating the model by comparing its results to actual data where possible, analyzing uncertainty attached to the output from the microsimulation model, and conducting sensitivity analyses using alternative assumptions regarding trends in model parameters; and

6. Using microsimulation, producing disaggregated projections of the elderly population and its characteristics, for example cohort profiles of active life expectancy, or comparisons over time in the health, family structure and economic well-being of the oldest-old.

INTRODUCTION

The importance of understanding the dynamics of health and associated behaviors within the elderly population is widely acknowledged. Researchers and policy analysts also recognize the importance of simultaneously taking account of developments in the domains of functional status, family composition, and economic resources. These needs are felt with respect to data collection—for example, they are an important part of the rationale for longitudinal surveys of the elderly (e.g., the National Long-Term Care Survey, the Longitudinal Study of Aging, and the recently initiated Assets and Health Dynamics of the Oldest Old or "AHEAD" survey)—as well as for the development and application of quantitative models. However, no single data source is likely to address all these issues for all age, socio-economic, and health states. Consequently, development is necessary, in order to integrate data from different sources.

This project addresses issues in model development. The project's goal is to produce individual-level projections of functional status, kin network composition, income, and nursing home occupancy among the older population using microsimulation. The model underlying these projections will support scholarly inquiry of both substantive and methodological interest, and will inform current and future policy debates in the areas of income security, health service use, and long-term care policy for the elderly (65+) and oldest-old (85+) populations. The model will permit analysts to address questions such as: (1) what behavioral responses to changes in the functional capacity of the elderly occur; (2) what familial and economic resources are available to cope with functional declines in the older population; and (3) how is the "burden" of dealing with these declines distributed across the older population, within kin networks of the elderly, and, through public expenditures, on society as a whole?

To achieve these goals, this project will address six specific aims:

1. Specify and estimate equations for dynamics of functional status, nursing home occupancy, income and death among those 65+, using data from the 1982, 1984, and 1989 National Long-Term Care Survey (NLTCS) linked to Medicare data for 1982-1993. This analysis will build on and extend existing research based on the Grade of Membership (GoM) framework.

2. Develop equations for year-to-year income streams, determined jointly with changes of marital status, for all ages represented in the cohorts to be simulated (e.g., 48+ in 1988, as explained below).

3. Estimate model parameters governing the dynamics of family composition (existence and characteristics of spouse, parent[s] and child[ren]).

4. Integrate the results of the above modeling efforts in a microsimulation computer program with the capacity to dynamically simulate life histories, focussing on the elderly population;

5. Perform extensive validity tests comparing the model's projections to actual data where possible; analyze uncertainty attached to the output from the microsimulation model; and, conduct sensitivity analyses of model results using alternative assumptions regarding trends in model parameters, etc.

6. Using microsimulation, produce disaggregated projections of the elderly population and its characteristics, for example cohort profiles of active life expectancy, or comparisons over time in the health, family structure and economic well-being of the oldest-old; and, conduct policy-relevant analyses, such as estimating the size of the target population for proposed community-care entitlement programs under alternative eligibility criteria.

The paper documents our plans for achieving these goals. We discuss several substantive and technical issues in detail, and outline our general approach to the issues. It must be stressed, however, that the project is still in an early stage of development, and that numerous details of thee final model have yet to be specified.

BACKGROUND AND SIGNIFICANCE

Population and Individual Aging

Population aging has profound implications for public health spending and related public policy issues in the future. For example, Guralnik et al. (1988) project that if nursing home use rates are maintained, the number of institutionalized elderly aged 85+ would rise from 600,000 in

1991 to five million by 2040. Comparable increases in the use of acute care facilities and home care services can be anticipated (Soldo and Manton 1985a).

The policy implications of population aging underscores heightened interest in modeling the health, kin, and economic-status trajectories of *cohorts* or, at a more basic level, of *individuals*. Changes in the size or composition of the older population based on extrapolations of aggregate indicators do not necessarily translate into forecasts of needs or resources available to a population, due to the diversity of the older population at present and in the future—diversity (heterogeneity) along dimensions such as economic status; access to kin, constrained by the competing uses of the time and economic resources held by kin; and the severity and progression of disease and disability. A change in the mix of such factors may alter the risks of a variety of outcomes (Soldo and Manton 1985b). Changes in family composition, marriage and divorce patterns, life-cycle employment profiles, and timing of retirement all suggest that the economic and familial resource base for accommodating health care needs of the elderly will continue to change. Projecting any single factor mentioned above is complex. Yet their interdependency requires that they be addressed *jointly* to understand and project how resources are mobilized in response to changes in health status. The availability of family members as potential caregivers, for example, may, in combination with financial resources, affect discharge status and the length of a nursing home stay. Given the complexity of modeling each factor individually, and the additional complexity of considering them in combination, the need for a microsimulation model is clear.

Functional Status and Active Life Expectancy (ALE)

The growth of the elderly population has been discussed by many authors (e.g., Soldo and Manton 1985b). Less well described are health and functional changes associated with that growth, such as the amount of time expected to be lived in specific health and functional states after a given age. One measure frequently calculated is ALE, i.e., the average number of years a person can expect to live free of "serious" disability. This has been calculated for select populations and from NCHS area probability surveys. In studying ALE a variety of measures and thresholds of disability have been used. Katz et al. (1983) defined disability by limitations of Activities of Daily Living (ADL). Wilkins and Adams (1983) used disability measures from a Canadian labor force survey. The use of different measures raises questions on how to compare estimates from different studies. Wiener et al. (1990) found that, though estimates of impairment prevalence among the elderly appeared to vary across United States surveys because of differences in the definition of functional disability, when a common definition was used estimates for several of the surveys converged.

Much of the recent research on ALE uses simple measures to distinguish the "active" from the "inactive" population. Rogers et al. (1989), for example, classified individuals in the Longitudinal Survey of Aging data as "dependent" if they were (a) in a nursing home **or** (b) "needed assistance" with one or more of seven ADL activities. For those not in nursing homes, this simple binary indicator ignores much variation in the range of activities, and degree of difficulty, within the group classified as dependent. It also may be contaminated by measurement error, since a single erroneous code (or "noisy" response) may change one's classification.

ALE is one component of life expectancy. Others are "morbidity free," "autonomous," and "dependent" life expectancy. Duration-weighted measures of health and functioning were reviewed by a WHO scientific advisory group (WHO 1984) which emphasized that, since functional impairment is intrinsically multidimensional, a single ALE index is unlikely to be universally applicable. Problems in selecting a disability "threshold" to define the change in ALE with age increase at later ages, as the complexity of functional impairment increases. For the "oldest-old" (85+) population, defining ALE by a single disability state may not be meaningful because while many persons at extreme ages manifest some types and degrees of functional limitation, and though impairment is more prevalent, most persons retain *some* degree and type of functioning, e.g., a person may have intact cognitive function but have mobility impairments due to degenerative joint diseases. Alternatively, some elderly persons have cognitive impairment though little physical impairment. Thus, to accurately describe functional impairment in the elderly population, graded and multidimensional measures should be used (Manton et al. 1991c)

Functional Status and Policy-Relevant Forecasts. President Clinton's recently announced Health Security Plan (White House 1993) includes a home- and community-care entitlement program with eligibility "...without regard to income or age" (p. 170). One of the four proposed targeting mechanisms is the requirement that a person have problems with three or more of five ADL activities: eating, dressing, bathing, toileting, and transfer. The stated time line for implementation of this provision consists of phased funding beginning in FY 1996, with full funding by FY 2000.

Notwithstanding the stated intent not to make this a means-tested program, it seems likely that means-tested alternative proposals will receive attention in the coming health reform debates. And, even if a program is enacted along the lines proposed, for purposes of studying distributional implications of any plan it is desirable to examine the joint distribution of income and functional status in the future. It will further be desirable to characterize the family situation of potential beneficiaries of such a plan, to establish possible bounds on the "woodwork effect"—the fear that establishment of a community-care entitlement will stimulate demand for care among those who would otherwise receive care principally from family members. For these reasons, a model capable of making detailed forecasts of the population classified by age, functional status, family composition and income level would be of great value.

Familial and Economic Resources

Kin networks constitute familial resources that, in combination with economic assets, can be drawn upon by functionally impaired elderly in support of both community-based and institutional care arrangements. Recent research has pointed out that the nature of community care arrangements, including the mix of formal and informal service providers, depends in part on the details of kin network, and differ according to the presence of children by age, sex, and marital status (Wolf and Soldo 1990, 1991). These findings underscore the importance of representing individual members of kin networks when projecting LTC needs and use patterns.

Major changes in family patterns and in intergenerational resource flows involving the elderly have occurred in recent decades. For example, during the postwar years there has been a pronounced increase in the percentage of elderly widows living alone (Kobrin 1976; Michael et al. 1980; Wolf forthcoming) which has been shown to be associated, in part, with changes in family composition, specifically a drop in the number of offspring per elderly parent. However, this pattern has recently shown evidence of slowing or stopping, and may be followed by a reversal as higher-fertility postwar cohorts reach old age (Macunovich et al. 1995). The existence of such trends underscores the need for forecasting models able to accommodate a variety of assumptions about future paths of a wide range of demographic parameters, and, equally importantly, to account, at a low level of disaggregation, for interactions among trends.

Economic resources, in the form of both income flows and wealth stocks, interact in important ways with the family and health circumstances of the elderly. Rising real income has been identified as a partial explanation of the trend toward independent living among the elderly (Michael et al. 1980) and is an important determinant of LTC services in view of the high costs of home-based and institutional care (Rivlin and Wiener 1988; Coughlin et al. 1990). Asset holdings are currently of particular interest in view of assets tests used to establish Medicaid eligibility, and the consequent need for elderly to "spend down" assets to qualify for public funding of their nursing home expenses (a phenomenon about which we have still little systematic evidence).

The economic status of the elderly has changed dramatically in recent years. Recent research (summarized in Holtz-Eakin and Smeeding 1992) documents the increasing real income of the elderly (whether measured using mean or median income) both in absolute terms and relative to the nonelderly. Three important contributors to this trend are policy changes in Social Security, postwar growth in private pension converge (a trend that may have ended), and intercohort changes in lifetime work behavior and earnings. Yet there is also great diversity in the economic circumstances of the elderly, necessitating an examination of the full income distribution rather than simply its central tendencies. While intercohort differences are important (and will continue to be, as successive cohorts reach old age having had distinctive lifetime labor market careers), there are notable intertemporal within-cohort patterns of interest. For example, there remains subgroups of elderly with low income and asset levels who are economically at risk. Furthermore, some research shows a pattern of declining income (relative to a needs index) as cohorts age (Duncan and Smith 1988). Quinn (1987), has called attention to the fact that *averages* (of income, but of other attributes as well) can be very misleading, and points out the importance of examining the structure of the entire income distribution, cautioning researchers to "beware of the mean." Finally, it has been documented that family events, particularly the death of a spouse, can trigger dramatic changes in economic circumstances (Burkhauser and Duncan 1988; 1991).

Developments in Microsimulation

The term "microsimulation" is used in a number of contexts and is associated with a variety of activities. In some contexts it is a numerical analysis technique. For example, several formal demographic models have been "solved" using microsimulation, particularly models of the demography of kinship (Le Bras 1978; Wolf 1988). Similarly, Wolf (1986) used microsimulation to study the sample paths of general event-history (renewal) processes. More recently, microsimulation has been used in statistical estimation and testing. Microsimulation techniques underlie recently developed approaches to estimating the parameters of multinomial discrete-choice models; these approaches entail simulating choice probabilities in ML (Lerman and Manski 1981) or method of moment (McFadden 1989; Pakes and Pollard 1989) algorithms.

One line of relevant work is due to Orcutt (1957), which gave rise to a number of static and dynamic microsimulation efforts, some of which represented government tax and transfer programs in detail. In a recent assessment of microsimulation, entitled *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling* (Citro and Hanushek 1991), the NAS's Committee on National Statistics pointed out the value of such efforts and made recommendations for their improvement. Our project addresses several of these recommendations, including the use of a more sophisticated model of health dynamics, the importance of validity checking, and analyses of uncertainty.

For this project, microsimulation has several advantages:

- We are interested in predicting outcomes in several domains, including health and disability status, nursing home occupancy, and old-age income flows. Each is represented by one or more outcome variables. Together these variables represent a high-dimensional "state space" for which microsimulation is a feasible modeling option. The state space of outcomes is complex not only at a point in time: over time the number of potential sample paths through the state space is immense. An analytic approach to studying the state dynamics would be difficult for the outcomes we wish to study.

- Microsimulation offers flexibility in analyzing and displaying outcomes. The result of a microsimulation is a data base recording simulated trajectories of a sample of individuals, beginning with an initial array and ending in a stipulated target year (or, when all members of a cohort have died). The microdata can be analyzed in a variety of ways—cohort profiles can be tabulated, and a sequence of period or cross-sectional indicators can be derived. Simulated outputs can be transformed and/or aggregated for subpopulations of interest.

- Microsimulation provides a means for conducting detailed policy analyses, and for analyzing the distributions of outcomes, as well as for sensitivity testing of model results to alternative assumptions about input parameters.

- Microsimulation provides a natural basis for developing measures of the uncertainty of projections. Below we describe methods that recognize several sources of uncertainty. In part these methods employ resampling techniques that rely on microdata. The multiple-imputation theory used in the uncertainty analysis requires microdata.

DESIGN AND METHODS

Overview

As noted above, our ultimate goal is to be able to conduct microsimulation analyses of the joint trajectories of functional status, kinship and income. Central to dynamic microsimulation is forecasting the trajectories of variables at the **individual** level. For each variable the forecast must account for the *expected value* of the variable and explicitly represent *stochastic variation* around the expected value. This forces the microsimulator to pay close attention to the stochastic structure of the model; it further indicates (1) the importance of developing accurate algorithms for sampling from the multivariate distributions of stochastic elements, and (2) indicates a need for validity testing of not only the model's predictions of expected values, but also of the combined levels of expected value and stochastic variation. Included in the latter are comparisons of the distribution of forecast values against the distribution of actual values in data used to estimate the model.

The starting point for our project is model ***specification*** and ***estimation***. It is important to maintain the distinction between the **model**—by which we refer to the formally-expressed relationships among observable variables, and among their deterministic and stochastic components, accompanied by all the implicit and explicit assumptions built into these relationships—and the **computer program** written to allow microsimulation analysis of the model's long-run implications.

Ideally, the model would be one in which an array of observable variables, some inherently discrete, others inherently continuous, jointly evolve in continuous time. However, methodological and empirical limitations lead us to a more modest model specification. A key methodological limitation facing anyone attempting a rigorously-founded microsimulation analysis is the following: any variable that appears in the model must appear somewhere as an endogenous variable (unless it is either fixed—e.g., race—or varies deterministically—e.g., age). Thus, for example, if it is desired to include "pension wealth" as an explanatory variable in an equation predicting retirement (i.e., pension acceptance) then it is necessary to have a model of the acquisition of pension wealth. This, in turn, indicates the need for equations predicting job change, job tenure, pension coverage and plan characteristics on each job held, etc. Thus, to add what superficially appears to be just one or two variables to the state space of a model can lead to rapid growth in the model's complexity and (if one is serious about accounting for interdependence among the dimensions modeled) intractable estimation problems. Consequently, we have adopted a modest list of endogenous variables. A second methodological limitation is the need to represent individual-level stochastic elements ("error" terms, in regression terminology) in the forecasts, so that all the variability in observed outcomes is preserved. This further tends to rule out "fixed effect" estimators often used in econometric analyses, in favor of random effects specifications of the persistent components of unobservables.

Data limitations restrict the empirical form of our model in two important ways: first, rather than continuous observations of the outcomes of interest, we have a sequence of observations at discrete times; and second, since no single data source contains information on all outcomes of interest, we must make assumptions about conditional independence of model elements. The conditional independence assumptions pertain to interdependencies among outcomes to be simulated, and to correlations between random components of the outcomes. In some cases we must assume away possible correlations between the stochastic disturbances of equations governing the dynamics of outcomes.

The essence of the model is captured in the expression

$$X_t = f(X_0, X_1, \ldots X_{t-1}; \theta_0, \ldots \theta_{t-1}), \qquad (1)$$

in which the array X_t contains all variables of interest. As noted already, in our model the key variables in the X-array can be grouped under three headings, **functional status** (represented by GoM scores), variables describing each member of the **kin network**, and **income**. As in the earlier GoM models already described, one of the classes will correspond to nursing home residence, treated as a "pure type." Thus, institutional status (at each time point t) will be an endogenous variable in the model.

The empirical form can be viewed as the reduced form of a model in which the dynamics of health status, family composition, and income are governed partly by individual choices (e.g., of behavioral risk factors, which are unmeasured, or service use—which may or may not be measured—which depends in part on economic resources and family composition) and partly by stochastic elements (e.g., representing physiological and/or cognitive change).

Equation (1) identifies the two basic building blocks of our modeling efforts: *dynamic equations* represented by $f(\ldots)$ and the sequence of parameter sets $\theta_0, \ldots, \theta_{t-1}$; and the *initial conditions* represented by the array X_0. The dynamic equations represent not only the evolution of observed outcomes among survivors over time, but also the selective loss from the starting population caused by mortality interacting with the other outcome variables. Prior to discussing these two building blocks, we briefly discuss several general issues related to the temporal, cohort, and substantive scope of the model.

Cohort Coverage

The data base representing the ***initial cohort*** for our analysis will represent the population of adults (those aged 20+) in 1988. This reflects the use of merged files of data from the 1988 National Survey of Families and Households (NSFH) and the 1989 interview wave of the NLTCS. This sample will represent the **non**institutionalized population of people aged 20 to 64, but the **entire** population of those aged 65+, including the institutionalized.

However, to help contain the modeling effort, dynamic forecasts will conducted only for the cohorts aged 48+ at baseline. This choice partly reflects convenience: those aged 48+ in 1988 are aged 50+ in 1990, aged 60+ in 2000, and so on. This, in turn, will align the simulated population with published population projections from the U.S. Census Bureau and others, in which projected population figures are routinely presented in corresponding age groups. A more substantively important reason for this choice is that it effectively rules out the need to include fertility in the model. For women aged 48+ fertility can be safely disregarded: in the most recently published vital statistics (for 1990) the birth rate for women of all races, in the age group aged 45 to 49, is only 0.2 per 1,000 (NCHS, 1993). In other words, even if our starting population had 5,000 women in this age group (a number far in excess of what we will actually have), we would miss only one expected birth per year of simulation.

For men, the decision not to include fertility is more of a problem, since a substantial proportion of men age 48+ are married to younger women, and thus as a group have higher fertility rates associated with them. For example, the same NCHS source shows that among men aged 45 to 49 (of all races), 1990 rates indicate that there will be 7.5 births per 1,000 men; for men aged 50 to 54, the rate is 2.8 per 1,000, and so on. Our preliminary examination of NSFH data indicates that among married male respondents aged 48 to 54, 27.1 percent have a wife aged 44 or less; among married men aged 55 to 64, 8.1 percent had a wife of childbearing age. By failing to simulate fertility, we understate the number of children of the men (the male

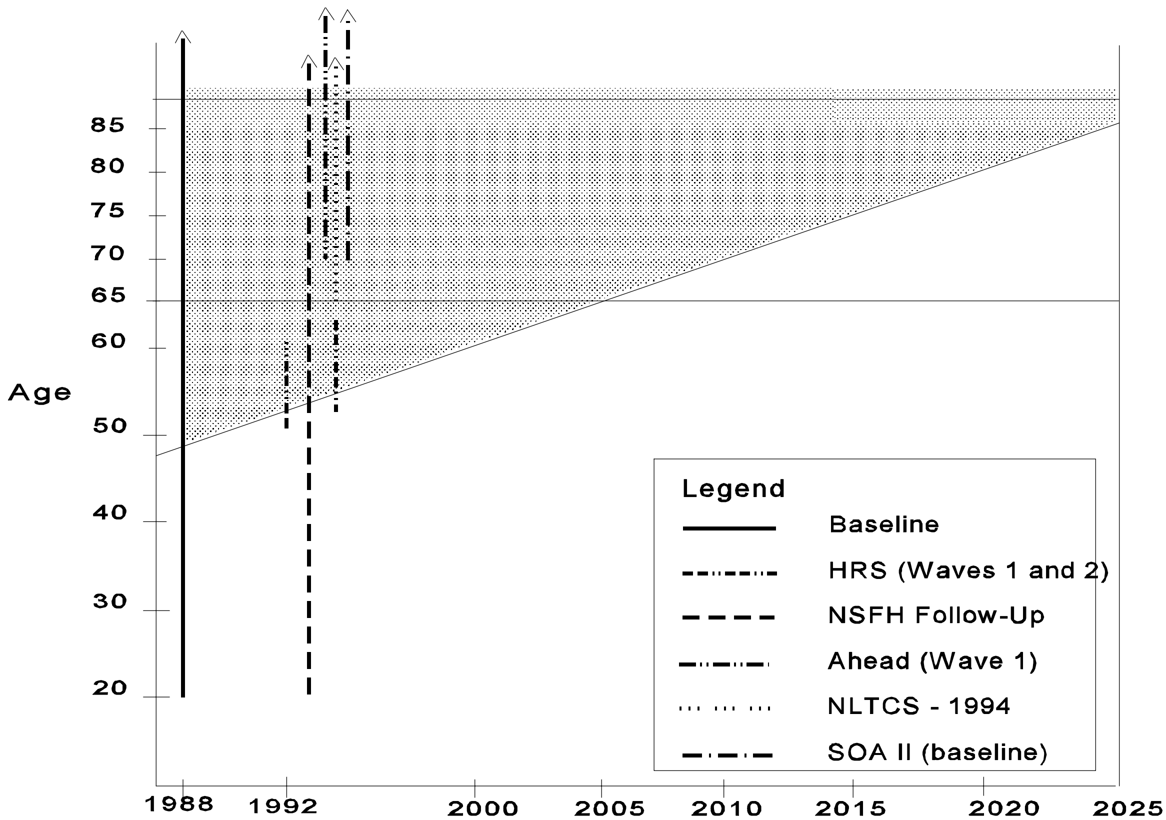
respondents) in our sample. Our solution to this problem (at least during the initial phase of the project) is to ignore the existence of children in the kin networks of the men in our simulated population. In other words, when tabulating output from the model, we will confine ourselves to components of the population for which we can make accurate statements. This will include all unmarried women age 65+, all married women age 65+, and married men age 65+ whose spouses are also age 65+. These groups represent the great majority of the older population (among the age 65+ population in 1990, unmarried women constitute 34.1 percent of the total, while married women are 24.2 percent; married men, a substantial percentage of which also have older wives, are 31.9 percent, while unmarried men are only 9.8 percent of the total; see U.S. Bureau of the Census 1992). We address the issues raised by including fertility in the model in more detail later.

Temporal Scope

Given the preceding choices for cohort coverage, the time period for which we can conduct interesting forecasts is limited to the years 1990-2025, at the end of which the youngest members of the cohort will have reached age 85. Thus, we can conduct *period* forecasts of the old (age 65+) population through 2005, and of the oldest-old (age 85+) population through 2025. We will also make *cohort* forecasts for cohorts represented in the dynamic analysis (i.e., those age 48+ in 1988), by running the microsimulation analysis until the cohorts are extinct (which will, of course, take place substantially after 2025).

These points concerning cohort coverage and temporal scope of the analysis are shown in Figure 1. The bold vertical line in Figure 1 represents the initial population (graphed against the vertical age axis, running from ages 20 to the maximum age found in the data base). The shaded area represents cohorts to be tracked over time, out to the year 2025 at which time they reach age 85+. Also shown are overlaps between our micro-forecasts and several existing/forthcoming survey data sources; this is discussed in the section on validation.

Figure 1: Cohort Coverage in Microsimulation; Overlap with National Sample Surveys



Variables Included in the Dynamic Model

Table 1 lists the endogenous variables included in the model. All variables are defined for an individual in our starting-population data base (which, as explained below, consists of respondents from the NLTCS and the NSFH). Many of these variables, however, refer to other people (e.g., the spouse, parents, and any children of these respondents). In the rest of this document we refer to sample individuals (i.e., those whose remaining lifetimes are to be simulated) as "respondents" and the family members attached to them (spouse, parents, children) as "kin."

Table 1: Variables Included in Model

1 if *Alive* in t; 0 otherwise
if Alive = 1:
  Age in t
  Sex (fixed)
  Education (fixed)
  Race/Ethnicity (fixed)
  Marital Status in t = 1 if never married (NM)
    = 2 if married (M)
    = 3 if widowed (W)
    = 4 if divorced (D)
  if Marital Status in t = 2:
    Spouse's Age
    Spouse's Education (fixed)
  Income (of R or R and Spouse) in t
  Retired in t (1 if receiving Social Security pension income)
  Functional Status in t (if Age in t $\geq$ 65)
    (Grade of Membership scores): $g_{1t}, \ldots g_{Kt}; g_{1t} + \ldots + g_{Kt} = 1$

  Number of children in t (K_t)
    if $K_t > 0$:
      Age of child 1 in t
      Sex of child 1
      Marital Status of Child 1 in t (1 if married; 0 otherwise)
      .
      .
      Age of child K in t
      Sex of child K
      Marital Status of Child K in t (1 if married; 0 otherwise)

For people under age 65, we forecast only the income and kin-network variables; this is because the NLTCS, from which we construct our measures of functional status, includes only respondents age 65+. Since our substantive focus is on the older population, the primary purpose of the equations for the income and kin-network outcomes among those under 65 is to project them forward in time to the point at which they reach age 65. The recently-available 1991 Medicare Current Beneficiary Survey provides a potential source of data with which to estimate a model of disability among the under-65 population. Extending our model in this way is a possible later addition to the proposed project.

Table 1 indicates that to forecast trajectories of the endogenous variables listed, we will need equations for mortality risks (for respondents, for spouses—through which the transition to "widowed" occurs—and for parents and children); for marital status changes (transitions from married to divorced and from divorced/widowed/single to married); for the evolution of GoM scores representing functional status, and for income. As previously stated (and as noted by reviewers), available data rule out a fully simultaneous model of all the outcomes listed. We do intend, however, to push available data as far as possible in introducing hypothesized interdependencies. Specifically, we plan initially to formulate a model in which: (1) year-to-year income flows, transitions to pensioner status, and the probability of divorce/marriage (as appropriate) are jointly determined (for all ages); (2) year-to-year income flows and the evolution of GoM scores for functional status are jointly determined (age 65+ only); and (3) own-mortality and spouse's mortality depend on the GoM scores (and hence implicitly on the linkage of functional status and income) (age 65+). The variables representing the continued survivorship of each parent and child in the kin network (and the variables representing the children's marital status) will, in our initial work, be assumed to evolve independently (of each other and of the respondent). Later we discuss ways to relax this assumption.

The <u>initial conditions</u> for the model (X_0 in equation 1) will be embedded in a data base containing initial values for the variables listed above. Some imputation of missing values, and/or of variables not measured, will be necessary to fully define initial conditions. For example, a substantial minority of the nonresident children enumerated in the NSFH have missing sex codes. Another example is the lack of information on living parents in the NLTCS. Imputation procedures and their consequences for uncertainty in model forecasts is discussed in a later section of the proposal. In addition, the initial-conditions database will include, for each respondent, the following <u>predetermined</u> variables: age, sex, race, educational status, and variables describing marital history. Respondents' kin are treated as "attributes" of the respondents in the model. For the spouse (when present initially) we will have age and education. For children and parents, we will have initial sex and age (except as noted above), and for the children, initial marital status.

We now discuss each of the components of the dynamic model. This is followed by further discussion of the creation of the initial-conditions data base.

Model Development

Our model of changes in functional status, income flows, and the time-path of the kin network is represented by a set of dynamic equations. However, the estimation of this system of equations is **preceded** by a Grade of Membership (GoM) "data reduction" step, and **followed** by the specification of ancillary equations necessary to control the dynamics of the GoM scores. The following discussion follows the logic of this three-step sequence.

GoM Representation of Functional Status. The GoM approach begins with what is essentially a data reduction step. At this step a large number of discrete indicators of observable symptoms or traits in one or more conceptual domains (e.g., health, functional status, productive capacity) is mapped onto a smaller number (K) of underlying classes. Using principles of fuzzy mathematics, each person is represented by K GoM "scores" each of which represents their degree of similarity to the corresponding class. These scores, in other words, represent grades of membership in the corresponding classes. Each class, in turn, represents a "pure type" which may or may not appear with appreciable frequency in observed data, that is, there may be a number of individuals whose observed symptoms correspond exactly to a given pure type.

The GoM approach is well suited to representing health and functional status. First, any underlying disease or physiological state may have associated with it several observable symptomatic manifestations. However, heterogeneity across individuals leads to variation in the particular symptomatic profiles of individuals with the same underlying disease states. At the same time, a given observable symptom may be the manifestation of any of a number of underlying diseases or conditions. Thus, even with perfect measurement of manifest symptoms our measures typically are insufficient to discretely classify individuals according to the conceptualized underlying health or functionality dimensions. Finally, there may be both measurement error and purely random factors leading to the presence or absence of particular symptoms at the time that measures are taken. The GoM fuzzy-set classification technique recognizes these uncertainties and indeterminacies in its treatment of observed indicators. Together these arguments provide a powerful rationale for the use of the GoM approach to representation of functional status in our project (for additional discussion see Woodbury et al. 1978 or Singer 1989).

GoM is a means of mapping responses corresponding to a large number (J) of observable indicators into a small set of scores representing an individual's degree of resemblance to each of K ideal ("pure type") response profiles at time t. ln the GoM framework response variables are coded as binary variables, x_{ijl}, with $x_{ijl} = 1$ if the ith person ($i = 1,\ldots,I$) had the lth response (of L_j) to the jth variable ($j = 1,\ldots,J$); $x_{ijl} = 0$ otherwise. The probability that $x_{ijl} = 1.0$ is represented using two coefficients. The first, λ_{kjl}, is the probability of the lth response for each of the $k = 1,\ldots,K$ fuzzy classes necessary to explain the variation of the x_{ijl}s. The second coefficient, g_{ik}, is an individual level weight used to combine the λ_{kjl}s to best predict the probability that $x_{ijl} = 1.0$. The g_{ik}s are estimated with convexity constraints, namely that $0 \leq g_{ik} \leq 1.0$ and $\Sigma_k g_{ik} = 1.0$. The model is written

$$Pr(x_{ijl} = 1.0) = \Sigma_k g_{ik} \lambda_{kjl} . \qquad (2)$$

Coefficients in (2) are estimated by maximum likelihood. Mathematically, GoM is a *fuzzy set* model, with gs being set membership functions. The restriction, $\Sigma_k g_{ik} = 1$, means that the model describes a fuzzy *partition* with K being the number of classes in the partition. If for some i, one of the gs equals 1, then that person's indicators exactly equal the corresponding "pure type" in the fuzzy partition. The K fuzzy classes represent characteristics that a person may express to a limited degree. Thus, in GoM a person may have partial membership in multiple fuzzy classes. Tolley and Manton (1991) show that the λ's, and the moments of the distribution of the g's of up to the Jth order, are consistently estimated. Likelihood ratio tests are used to make inferences about the model "order" (i.e., the value of K).

Using GoM also improves the precision of statistical estimation of disability parameters. That is, each disability measure is presumably reported with error. As the number of items used to construct a score increases, the precision of the score improves because the parameters are more precisely estimated. Also, with more measures the disability state space can be more finely partitioned (i.e., K can be increased). If discrete disability categories were used then persons near the threshold defining disability/non-disability states would be incorrectly classified more often than those in the interior of the classes. This does not happen in a fuzzy state model with sufficient measures to "densely" populate the space with finely estimated partitions.

As a starting point, we plan to use as response variables the 27-indicator set already extensively used by Manton and colleagues, including a published set of estimates of ALE (Manton and Stallard 1991; Manton et al. forthcoming). The 27 indicators used include nine binary indicators of ADL limitations, 10 binary indicators of IADL limitations, seven physical performance measures (e.g., difficulty climbing a flight of stairs, difficulty bending, reaching overhead, and so on) each of which have four response categories, and a binary indicator of vision impairment. These have been shown to map into a set of six GoM scores (i.e., a model in which there are six "pure types") which, in turn, exhibit strong external validity. A seventh "pure type" indicating nursing home residents completes the representation of the functionally impaired United States population.

The functional status modeling will use data from the National Long-Term Care Survey (NLTCS). The 1982, 1984, and 1989 NLTCS provide detailed nationally representative health data on the non-disabled, the community disabled and institutionalized United States Medicare eligible elderly population. Interview data are linked to Part A and B Medicare service use data, and to mortality data (with 100 percent follow-up of decedents), allowing for a nationally representative accounting of all types of Medicare service use for the entire sample, and a more accurate definition of functional characteristics for people with incomplete data. Another advantage of the NLTCS is that persons were followed prospectively from the community to institutions; hence a true nursing home admission cohort is identified in the 1984 and 1989 NLTCS.

The design of the NLTCS is complex, with separate accounting of sample members who did and did not screen in for purposes of the community-based detailed interview; an initially institutionalized population not interviewed in 1982, but traced and eligible for interview in 1984 (and again in 1989); a newly age-eligible component in 1984 (and in 1989); and, tracking of mortality and service use information from ancillary data sources. Further details concerning the 1982 and 1984 data, including counts of the various subcomponents of the sample, can be found in Manton (1988); an analogous discussion, extended to include the 1989 data, is contained in Manton et al. (1993b).

The NLTCS has been used in numerous analyses by Manton and colleagues and by others. Manton and Stallard (1991) use data from the 1982 and 1984 waves of the NLTCS to compute an estimate of ALE for the United States population, using the GoM technique to represent classes of increasingly dependent subjects based on a set of 27 observed functional indicators.

Important (and somewhat unexpected) findings have begun to emerge from analyses based on the existing data. In particular, in the first of two recent publications (Manton et al. 1993a) results were presented that document changes over time in the use of assistive devices among those with difficulties in ADL/IADL functions. Corresponding to the increased use of equipment is a decline in the use of personal assistance as the sole form of help. A subsequent publication (Manton et al. 1993b) documents **declines** in both the incidence and the prevalence of chronic disability among the community-based and the institutionalized population during the period 1984-89.

Modeling Income/Marital Status Trajectories. The equations representing dynamics in our three outcome domains comprise the heart of our analysis. Here we begin to lay out the series of equations to be estimated. As already mentioned, we are unable to adopt a specification in which all outcome domains are fully simultaneously determined. Moreover, we have a somewhat different model for those under, and those over, age 65.

Equations will be estimated to forecast year-to-year changes in income, pensioner status, and marital status for respondents 48+ in the starting population. The equation for annual total income streams can be viewed as the reduced form of an implicit structural model of wage growth, labor force participation, earnings, nonwage and transfer income. While a more complex representation of the process, distinguishing among these several interrelated components, is clearly a desirable feature of a forecasting model, we choose to restrict our attention to the simpler model in the proposed project to keep the project scope within manageable bounds.

A number of considerations guide the specification of the income equation, foremost among which is our wish to track year-to-year income dynamics accurately, to have a basis for projecting future income streams. In addition, the model is intended to: (1) track persistent individual unmeasured differences in income from year to year (Lillard and Willis 1978; Lillard and Weiss 1979); (2) reflect intercohort differences in income trajectories, including secular increases in labor force participation by women and more general phenomena related to the earnings prospects of different educational groups and cohorts (Levy 1988); (3) account for real income growth (or losses) associated with productivity change (Levy, 1988), and allow for the possibility that such real-income growth may not be shared equally across age groups; and (4) represent discontinuities in income streams associated with retirement (pension acceptance), i.e., replacement rates.

Grad (1990) provides recent estimates of Social Security and private-pension replacement rates, computed for a variety of pre-retirement income bases; these estimates range from 30 percent to 50 percent when retirement income is expressed as a percentage of average earnings immediately before retirement. To account for this discontinuity, our model of income will include an equation for the discrete "shift" from pre- to post-acceptance of a pension. Furthermore, while there is ample evidence of a characteristically convex age-earnings (and by extension age-income) profile associated with human capital acquisition and depletion, income streams during retirement are primarily governed by other forces (Social Security benefits are indexed, and change by formula, but have in addition been irregularly adjusted via legislation; private pensions may be fixed in nominal terms, or vary essentially with real growth, depending upon their provisions).

Another outcome variable to be modeled is the transition from married to divorced (and, for those unmarried, from single/divorced/widowed to married). Although divorce rates are low in the age groups to be studied—in 1987, divorce rates for men aged 45 to 49 were 17 per 1,000, falling to 2 per 1,000 among those age 65+; the corresponding figures for women are 13.1 per 1,000, falling to 1.5 per 1,000 (NCHS 1991)—they can be expected to vary by predetermined variables such as age, race, education, duration of marriage and childbearing history. The rate of movement into marriage at these ages is somewhat higher: among divorced and widowed women aged 45 to 49, the rates are between 44 and 48 per 1,000, falling to 2 to 4 per 1,000 at ages 65+; for men in the same age/marital status groups the rates are substantially higher (NCHS 1991).

Furthermore, as noted by Johnson and Skinner (1986), there is evidence of increases in married women's labor supply preceding divorce, and further increases in their labor supply after divorce. The pre-divorce increase in women's work could arise from either anticipatory behavior, or it could be a reaction to exogenous shocks but could itself raise the probability of subsequent divorce. Under either hypothesis future divorce becomes a potential regressor in an analysis of current labor supply, but a variable representing **actual** future divorce should not be used since it may be simultaneously determined with current work effort. Johnson and Skinner formulate a two-equation model for women's current labor supply and future divorce, using cross-sectional data (from the PSID) to estimate the structural labor supply equation. They interpret their findings as evidence for the anticipatory effect of future divorce on women's current labor supply.

Following these arguments, the probability of future divorce should appear as a regressor in the equation for current annual income for married couples; however, for total couple income the anticipated sign of the coefficient on the instrument for future divorce is not obvious: divorce probabilities may rise when husbands' earnings (one component of family income) are unexpectedly low (see Becker, Landes and Michael 1977); on the other hand, wives' labor supply may rise (for reasons unrelated to the perceived risk of divorce) to compensate for negative transitory husbands' income, thereby raising family income. Johnson and Skinner's arguments also indicate that there should be an effect of **recency** of divorce in income regressions for *unmarried* women. Finally, we anticipate a correlation between the residual in the income equation for a couple in the year **preceding** divorce and that of the income equation for the former wife (and for the former husband) in the year **following** the divorce.

For the variables (a) income (b) marital status change and (c) transition to pensioner status for pre-retirement couples, we will estimate a model of the following general form. In the income equation,

$$\ln y_{it}^{c} = \alpha_1 + \gamma_1 \delta_{i,t+1}^{m} + \beta_1 X_{it} + \epsilon_{1it}, \tag{3a}$$

y_{it}^{c}, is the couple's income, X_{it} is a vector of exogenous characteristics, and $\delta_{i,t+1}^{m}$ indicates the occurrence of divorce between t and $t+1$. In specifying ϵ_{yit} we will investigate using an individual random component (ω_{yi}) and a "transitory" component u_{yit} which is potentially serially correlated. The probability of divorce is given by

$$Prob(\delta_{i,t+1}^{m}) = f(\alpha_2 + \gamma_2 \ln y_{it}^{c} + \beta_2 X_{it} + \epsilon_{2it}) \tag{3b}$$

while the probability of the transition to pensioner status (indicated by the variable $\delta_{i,t+1}^{pens}$) is given by

$$Prob(\delta_{i,t+1}^{pens}) = f(\alpha_3 + \gamma_3 \ln y_{it}^{c} + \beta_3 X_{it} + \epsilon_{3it}). \tag{3c}$$

Since spouses in two-earner couples do not necessarily retire the same year, we will have to adopt a somewhat arbitrary approach to coding the retirement transition (e.g., the first year in which pension income is received). Relaxing this restriction adds considerably to model complexity, and is a topic for subsequent work.

Recall that in a microsimulation any variable appearing anywhere in the model as a *predictor* must appear elsewhere in the model as *endogenous*, since otherwise it will be impossible to forecast trajectories of all variables. Consequently our model of income and retirement (pension acceptance) is relatively simple, using income but not pension coverage/pension wealth, assets, or health. Using a simple labor-leisure analysis of the timing of retirement, income effects should lead high-income individuals (couples) to retire sooner than those with low incomes. Moreover, higher-income individuals are more likely to have substantial public and private pension wealth; together these arguments suggest a negative relationship between income and the age of pension acceptance. In contrast, however, lower-income people will tend to have poorer health, and poor health has been shown to strongly predict early retirement (Burkhauser and Quinn 1994). Moreover, there is an offsetting substitution effect, caused by the greater opportunity cost of retirement among higher-wage workers. Whatever the effect of total income on pension acceptance (without controls for pension coverage, health, and so on), there may be a correlation between shocks to current income (given by the composite error in equation 3a) and the disturbance in equation (3c).

Following pension acceptance, the couple's income is determined by an equation similar to (3a),

$$\ln r_{it}^{c} = \alpha_4 + \beta_4 X_{it} + \epsilon_{4it}, \tag{3d}$$

with r_{it}^{c} representing the couple's retirement (post-pension acceptance) income. The random component ϵ_{rit} may [like its counterpart in equation (3a)] be partitioned into a time-invariant random component and a (possibly serially correlated) transitory component.

Identification of these equations is a serious analytic problem (as is usually the case). As a point of departure, it can be argued that variables representing the existence and ages of parents belong in the income equation (in view of recent evidence that women's work efforts and caregiving efforts respond to potential claims on their time such as the existence of older parents, see Wolf and Soldo 1994), but can be excluded from the probability-of-divorce equation. Identification of the pension-acceptance equation may be more problematic, and we may have to settle for reduced-form estimates of this "switching" equation.

The model will include equations analogous to (3a)-(3d) for unmarried women and men (although sample-size limitations may force us to impose a number of cross-equation parameter constraints). For that part of the sample observed to make a marital status and/or a pension-acceptance transition, part of the data would be modeled using the "couple" equations shown and part would be modeled using the corresponding "singles" equations (and/or part of the data would be in the pre- and part in the post-pension acceptance regimes). These subsamples of "movers" would provide information with which to identify cross-equation correlations among disturbances.

There exists a large econometric literature on the dynamics of wages and earnings (Lillard and Willis 1978; Lillard and Weiss 1979; MaCurdy 1982; Hause 1985; Abowd and Card 1989; Gottschalk and Moffitt 1993). A conclusion supported by this research is that the stochastic component of earnings dynamics can be captured in a fairly simple specification, one containing a fixed individual component and an ARMA1,1 temporally random component; i.e., the effects of temporal shocks quickly die out. Our model can represent this stochastic structure; ω_{yi} represents the time-invariant random effect, while u_{1it} and u_{4it} are autoregressive processes. We anticipate that ω_{yi} and ω_{ri} would be positively correlated: someone with pre-retirement income that is higher than expected on the basis of observable characteristics, would also be likely to have post-retirement income that is higher than expected on the basis of observables. In addition, we intend to estimate correlations between the random effect in the income equation, and the disturbances of the two probability functions (3b) and (3c). If we were to model fertility, in addition to the variables already discussed, the model would become far more complex: besides fertility it would be necessary to disaggregate couple income into separate components for each spouse's earnings (since these are jointly determined, which would further argue for separate earnings-selection and conditional earnings equations, at least for married women—i.e., generalized Tobit). It is the desire to defer analytical complexities such as these that motivates us to adopt a simpler specification, at least initially.

Data for estimating the income/pension status/marital status submodel will come from the Panel Study of Income Dynamics (PSID), the 1992 release of which will contain up to 22-year income histories (for the period 1968-89). All the endogenous and exogenous variables mentioned above are available in the PSID (the existence and ages of parents can be inferred retrospectively from kin-network questions asked of all respondents in 1988). Though the intended use of the estimated equation is to forecast year-to-year changes in income and marital status among those aged 48+, we may be able to improve the estimates by using a more broadly-defined age group.

Joint Dynamics of Functional Status, Mortality, and Income. A second set of equations will represent the joint evolution of functional status (represented by GoM scores, described above) and income. These equations will be estimated using data from the 1982, 1984, and 1989 NLTCS, which provides measures of income in addition to functional status. Variables measuring the characteristics of the spouse (age and education) are included, and the death of the spouse can be inferred from changes in the respondent's marital status (a more precise dating of the spouse's death will be obtained from the linked Medicare files). The NLTCS does include cross-sectional variables describing all living children (age, sex, marital status), the major component of the kin-network dimension of our overall model. However, since the information on children is not linked from interview to interview, it is not possible to model changes in the characteristics of kin networks jointly with changes in functional status and income. We will, however, treat the kin variables as exogenous covariates in the functional status/income equations.

This part of our model generalizes and extends past work on health and functional status conducted by Manton and colleagues. Here we begin by providing an overview of the relevant methods used in the past work for forecasting the evolution of GoM scores. Applications of the model and estimation techniques mentioned can be found in Manton et al. (1991b). We then go on to present the extensions to that past work that will be carried out in the proposed project.

There are two sets of equations to be discussed; the first are mortality functions that depend on GoM scores (and, in the proposed work, on other covariates), and the second are autoregressive functions for changes in GoM scores (and, in the proposed work, on income and other covariates). Ideally, the parameters of both types of change would be estimated using a single likelihood function. However, that likelihood does not possess a closed form solution.

Earlier we discussed several reasons why GoM is well suited to the analysis of complex high-dimensional functional status data. As we now begin to discuss the **dynamics** of functional status, a further justification for using GoM becomes apparent: if we wanted to take account of possible transitions among outcome categories using the full array of 27 functional status indicators, conventional dynamic models—such as discrete Markov chains, for example, as would be used in a multistate life-table approach—the data requirements would be overwhelming, far beyond the capacity of any existing data source. However, by mapping the numerous observed indicators into a smaller-dimensional GoM state space we can make feasible a dynamic model while representing variations in the high-dimensional space of observable indicators.

Under certain conditions a stochastic random walk over a large number of discrete states is approximated, in the limit, by a multivariate Gaussian process. Manton and Woodbury (1985) show that these processes can be estimated from variables measured at fixed times in a longitudinal study. The stochastic differential equations describing such a multivariate process are robust to departures from normality if individual changes are linear functions of "current" state variables. In the multivariate case, inferences can be made about the change of latent state variables even with measurement error.

With estimates of an individual's vector of disability scores (i.e., gs) for two (or more) times (the λs are fixed over time so that gs relate to the same $K-1$ dimensions) the model represents disability changes over age and the relationship of mortality to those changes, using a model of human aging and mortality due to Woodbury and Manton (1977).

The first equation is a quadratic mortality function, with an exponential term in age, or

$$\mu(g_{it}, age_{it}) = \left\{g_{it}^{T} Q\, g_{it}\right\} \exp(\theta age_{it}) . \tag{4}$$

In (4), mortality is a quadratic function of *g*s, and the expression provides an estimate of the instantaneous hazard of death, μ, given the *g*s. The second term in (4), $\exp(\theta \mathrm{age}_{it})$, represents the effect of age related unobservable variables on the relation of *g*s to mortality, i.e., each term in (4) is multiplied by this factor so that the mortality effects of functional impairment are scaled by age related factors. The coefficient θ in the exponential function controls the age rate of increase in mortality and is viewed as an "intrinsic" aging rate. As the information contained in the *g*s increases, however, θ decreases. If $\theta = 0$ there is no effect of unobservables on mortality and the time to death is determined only by *g*s. We will estimate the mortality function for our sample population (NLTCS list sample, for which there are 12,000+ deaths recorded in linked Medicare records through 1992), and we will also estimate a mortality function for their spouses (if married). The data contain no health/functional status information for the spouses, but we will make the spouses' mortality depend on the **respondents**' *g*s (arguing that this cross-spouse dependency reflects common unmeasured risk factors).

The second equation depicts a vector autoregressive process for changes in g_{ikt} over age and time. The equations describing changes in disability as a function of prior disability and age (*t*) are

$$g_{i(t+1)} = C_t\, g_{it} + e_{i(t+1)} , \tag{5}$$

where $C = \{c_{hk}\}$ is a K x K matrix of transition coefficients, which are functionally equivalent to regression coefficients with constraints to keep $0 \le g_{ik(t+1)} \le 1$ and $\Sigma_k g_{ik(t+1)} = 1;$ and $g_{it} = \{g_{ikt}\}$ is a K-element vector, as is $g_{i(t+1)}$. In (5), $1 - c_{kk}$ is an age specific rate of change in the kth score over t to $t+1$, and c_{hk}, $h \neq k$, is the rate at which the kth score at t contributes to the hth score at $t+1$. Since g_{ikt}s are estimated under the constraint $\Sigma_k g_{ikt} = 1$ a change in the c_{hk} in one equation (i.e., for fixed h) is numerically matched by changes in the other K-1 equations (due to the convexity constraints). The C matrix in (5) generalizes the Chapman-Kolmogorov equations of a discrete-state process to the present case, in which the elements of C represent intensities of movement, between t and $t+1$, between locations in the K-dimensional unit simplex (see Manton et al. 1992, p. 326; included as appendix D). Likewise the error vector $e_{ik(t+1)}$, whose cross moments define the diffusion or innovation process of the g_{ikt}s is bounded both in terms of means and variances, and in particular cannot have greater than Bernoulli variance.

Alternative approaches to estimating (5) are possible. One possibility is OLS, but this presents problems analogous to single-equation estimators for dummy dependent variables: the boundedness restrictions on the dependent variable may not be maintained in forecasts. Two approaches are discussed in Manton et al. (1992), one of which, the minimum-entropy (or maximum information) estimator (described on p. 327 ff of Manton et al. 1992) is the preferred estimator due to its superior theoretical properties.

In this project we will extend the model represented by equations (4) and (5) to include both exogenous and endogenous covariates. There are a number of ways to introduce covariates, with increasing degrees of complexity and conceptual richness. The simplest way to introduce an exogenous covariate is to stratify the data, and produce separate models (with separate parameters Q, θ, and C) within each stratum. Stratification is a reasonable and feasible approach if we are dealing with just a few strata (e.g., males and females) but further attempts at stratification will quickly exhaust the data. A second step is to define interactions with the g_{ikt}s for relevant exogenous variables (x_{ijt}), as in

$$g_{i(t+1)} = C(x_{ijt})\, g_{it} + e_{i(t+1)}\,, \tag{5a}$$

where C remains a KxK matrix but which is now modeled as a function of exogenous variables. The advantage of this approach is that the cross-equation constraints on the $g_{ik(t+1)}$s is maintained, and the maximum-information estimator remains applicable.

A third approach introduces covariates in marginal form (i.e., adds equations). Thus, a system of equations for the joint evolution of GoM scores and income can be written as

$$\begin{Bmatrix} (g_{i(t+1)}) \\ y_{i(t+1)} \end{Bmatrix} = \begin{Bmatrix} C & \gamma_{12} \\ \gamma_{21} & \beta \end{Bmatrix} \begin{Bmatrix} g_{it} \\ y_{it} \end{Bmatrix} + \begin{Bmatrix} e_{i(t+1)} \\ u_{i(t+1)} \end{Bmatrix}, \tag{6}$$

where γ_{12} are effects of income (at t) on GoM scores (at $t+1$), and γ_{21} is the opposite. Again, the variances of the $g_{ik(t+1)}$s would have to be adjusted for consistency with the boundary conditions. Another issue to be addressed is the distribution of the disturbance in the income equation, $u_{i(t+1)}$, which in the submodel described in the previous section was implicitly assumed normal; here, we must address the joint distribution of the GoM score errors ($e_{ik(t+1)}$) and the income-equation error. Additional exogenous variables can be added to this system of equations, either in the form of additional equations (where the variables affect GoM scores and/or income but are not affected by GoM scores or income) or as interactions (as in equation 5a).

Projection Cycles. The interval was 2 years between the 1982 and 1984 NLTCS and five years between the 1984 and 1989 NLTCS. The microsimulation requires parameters for one-year intervals. Preliminary analyses suggest that the mortality and dynamic coefficients matrices may change over the follow-up interval (Manton et al. 1991b). However, since the date of death is known for each decedent, mortality coefficients can be estimated for any length of follow-up. The problem is more difficult for the dynamic coefficients (i.e., the C matrix). We will investigate two strategies. First we will evaluate the 2nd and 5th roots of the separately estimated matrices for 1982-84 and 1984-89. In both cases the roots are only an approximation to the one-year matrix of transition parameters. This will provide estimates that can be used directly in the simulations. Second, we will modify the model to directly estimate the single year transition matrices using the Missing Information Principle (Orchard and Woodbury 1971): i.e., with *g*s observed in 1982 and 1984, the *g*s for t = 1983 are "missing," and similarly the *g*s for 1985, ... ,1988 are "missing." However, the forecasting model represented by (4) and (5) [further modified by equations (9)-(12) below] can estimate the mean and variance of the *g*s for each individual in 1983 (or the other missing years). These can be used to predict the mean and variance of the *g*s for each individual in 1984. These can be used, in turn, to form a likelihood for the observed *g*s in 1984, based on, say, the Dirichelet distribution (Narayanan 1990). The likelihood for the observed *g*s in 1989 is formed similarly, with the projection matrix applied five times, not two. With the likelihood, one can obtain maximum likelihood estimates of C_t for single year transitions.

"Tracking." Tracking refers to the temporal persistence of individual deviations from the average GoM scores. Tracking is thus a form of heterogeneity in the population, in which the expected trajectories of the *g*s are distributed in the population. Thus, the usual assumption that cases regress toward the mean may be violated, and each individual (or groups of similar individuals) have their own homeostatic point (or trajectory) toward which they tend. This type of process is consistent with what is described as a "random effects" model with correlated errors over observations on the same individual. The preceding mathematical structure represents tracking in several ways. First, the equations for the $g_{ik(t+1)}$s have no constant; any of the k vertices might be taken as a "null" or "origin" profile in the GoM state space. Second, the error terms in (5)—the $e_{ik(t+1)}$—are themselves state dependent, so that there are different rates of change (in GoM scores, through the C matrix) and in mortality, depending on the location in GoM space at t. Since the variance of the diffusion process tends to zero near a boundary (0 or 1) and tends to a maximum as g_{ik} tends to 0.5, "tracks" will exhibit differential forms of state persistence over time.

A alternative and more elegant approach, recently developed by Woodbury (and illustrated in Woodbury et al. 1992, using data from the 1982 and 1984 NLTCS), is to estimate an extended GoM model of the form

$$P_{ijl}^{t} = \mathrm{Prob}(x_{ijl}^{t} = 1.0) = \sum_{r} \gamma_{ir} [\sum_{k} \phi_{k}^{r}(t) \cdot \lambda_{kjl}]. \tag{7}$$

This model [which generalizes (2)] relates observed indicators to 2 underlying parameters. The terms γ_{ir} in this model are time-invariant effects at the person level, and can be viewed as fixed "frailty" parameters (although nonparametric; cf. Vaupel and Yashin 1985), while the λ_{kjl} are [as in (2)] time-invariant parameters associated with observable indicators. The ϕ parameters vary over time, and indicate the dynamics of group membership for the individuals in the sample. In the extended model R indicates the number of distinctive "tracks" (and $R \neq K$). This reduces to the standard GoM model when the matrix $\| \phi \|$ equals the identity matrix (and $R=K$).

Covariates can be brought into the extended model by making the γ matrix a function of age and covariates: that is, the time dependence in γ operates through a factorization to time-invariant and time-dependent components:

$$P_{ijl}^{t} = \sum_{r} \gamma_{ir} \{ (\sum_{k} \phi_{k}^{r} \cdot e^{[\beta x + \theta] \cdot \mathrm{Age}_{it}}) \cdot \lambda_{kjl} \}. \tag{8}$$

The extended model represented by (7) has been estimated (with results reported in Woodbury et al. 1992). The extended model with covariates [equation (8)] will be estimated in this project. The advantage of the proposed framework is that it can be estimated using a modification of existing GoM estimation software. Identification of the time-dependent terms representing exogenous influences [the θ parameters in (8)] depends on the specific factorization chosen for the model.

The NLTCS list sample includes people identified through a screener interview as nondisabled, while the detailed interview was administered only to those identified as disabled. Income is not known for the nondisabled respondents who answered only the screener interview, and therefore is missing for this part of the sample. Our income equation estimated using PSID data [equation (3a)] will, however, be used to impute income where it is missing in the NLTCS. We can then estimate the full model [equation (6)] for all elderly. Multiple imputation will be used to account for uncertainty associated with the use of imputed values (this is discussed more fully below).

The Dynamics of Family Composition. Kin networks will be described in our model by a **family roster** of individuals related to the "sample" members (respondents), in particular their spouse, parents, and children. Each such individual is described by an indicator of their existence, and by variables giving their sex, age, and (for children only) marital status. The initial composition of the kin network, including values of all sex/age/marital status variables, is contained in the data base representing the starting population. The microsimulation must provide for simulated changes in all these variables. Change along each of these dimensions is most readily modeled as transitions in a discrete state space, governed by continuous-time intensities or hazards (i.e., a semi-Markov model, or a more general stochastic process model).

The sophisticated model of mortality at ages 65+ (represented by equation (4) and estimated using NLTCS data) that will be used to project survival of the respondents cannot be used for respondents' *kin* (with the exception of their spouse, as described on p. 71), because (1) survival of other kin (children only, in the NLTCS) is not tracked in the data, and (2) for each child we know only age, sex, and marital status. Furthermore, most of the children of NLTCS respondents are under age 65. Thus, we will use data from other sources to represent mortality and (for children only) marital status changes.

The state space for the variables describing kin is (1) for parents, the states *alive* and *dead*, and (2) for children, the state *married*, *unmarried*, and *dead* (the NSFH data used to construct the starting population does not record the marital status of all parents; both existence and age of living parents will have to be imputed to respondents from the NLTCS). For both classes of kin, transition intensities will be modeled as depending on age, sex, and race (using the respondent's race).

Our analysis will use published sources of relevant vital rates (e.g., from NCHS) and/or series of the pertinent schedules from established sources such as the SSA actuaries' model schedules (see, for mortality, Bell et al. 1992). New techniques for modeling trends in age-specific rates (in particular, mortality schedules) developed by Lee and Carter (and summarized in Lee 1992) provide a means of exploring the implications of continuation of long-run trends in demographic parameters, and will be used in our analysis. The analysis consists of regression analysis, with log rates as dependent variables, and variables representing age and time as explanatory variables; separate models are estimated for each sex/race group. Parameters representing trends constitute one of several areas for sensitivity analysis in the context of microsimulation.

One limitation of the available data is that mortality rates specific to marital status are not available every year (see Hu and Goldman 1990), while rates aggregated over all marital statuses are available annually. One possible solution, which allows use of all the available data, is to simply assume that marital status-specific mortality differentials are, in the future, *fixed* at some value (e.g., the most recently observed value, specific to sex and age groups). This approach must be analyzed with respect to any implied assumptions concerning the relative importance of "protection" and "selection" mechanisms underlying the observed differentials (Goldman 1993).

The parameters used to fix these mortality differentials would, furthermore, become another area for subsequent sensitivity testing.

We will estimate time-series regression equations for age/sex/race-specific mortality rates (aggregated over marital status), for marital-status-specific mortality rates (distinguishing rates for those "married" and "unmarried"), and for transitions between the states "married" and "unmarried." Using a seemingly unrelated regressions approach, we will estimate correlations between residuals across equations, which if found to be nonzero would induce covariances between rates used as input parameters in the microsimulation forecasts (note, however, that Lee and Tuljapurkar 1993, investigated cross-equation correlations in their time-series model of mortality and fertility rates, and found none).

To use transition rates modeled as just described to simulate transitions made by each parent/child in their respective state space, we are forced to assume that the transitions of each individual in the kin network occur independently. This conditional-independence assumption will be used initially, but we plan to conduct analyses that will permit us to relax the assumption.

We intend to explore more complex models employing additional sources of stochasticity (e.g., unmeasured heterogeneity), considering in turn possible sources of intrafamilial dependencies. These efforts will be constrained by the availability of data. The basic form for such a model is

$$\mu_i(j,k;a_i,X_i) = \mu_o(j,k;a_i,X_i)\gamma_i\,,$$

where μ_o is the baseline hazard of an j-to-k transition at age a, X_i represents covariates for person i, and γ_i is a person-specific term interpreted (in the mortality context) as "frailty" and more generally as a measure of overall unobserved propensities to exhibit the indicated transition.

One source of intrafamily dependency involving *observable* variables that we will investigate is the influence on children's marital-status transitions of parental attributes (especially income) and the size of the child's "sibship;" e.g., in an equation for the rate of becoming married, the (log) rate could be specified as a function of parents' current income (proxying long-run economic status) and family size. PSID data permit some investigation of such dependencies, but only for parent-child pairs in which the child is young enough to have been living with the parent at baseline (in 1968). Wolf (in a separate project) is creating a data base that will support such an analysis. A further data source with which such an analysis will be conducted is the linked 1988-1992 NSFH data, which will become available during the project. This will allow us to model correlated outcomes across family members in the data used as part of the starting population for the microsimulation.

The unobserved-heterogeneity model laid out above has been widely developed and applied in recent years, in the context of fertility analysis (Heckman and Walker 1987, 1990) and mortality analysis (Manton et al. 1981, 1986; Vaupel and Yashin 1985). In principle, we would like to use models of this form to investigate intrafamily dependencies, i.e., the possibility that within families the "frailty" parameters can be regarded as drawn from a multivariate distribution. Correlated parameters of this sort can be regarded as the consequence of both shared genetic influences and shared socio-economic and physical environments. Recent empirical research on the heritability of longevity, using twins data (e.g., Hougaard et al. 1992) and has found evidence of within-pair dependence of lifetimes; however the estimated dependence is so small that knowing one twin's survival status provides only slight improvement in the ability to predict the other twin's survival. In view of this finding, it is unlikely that we will be able to estimate unmeasured-heterogeneity models that incorporate intrafamily correlations in the course of this project.

Maintaining Boundedness; Forecasting Consequences of Improved Health

The preceding section has described the development and estimation of the heart of our model: equations representing dynamics in the domains of functional status, income, and kin-network composition. However, prior to using these equations to forecast individual trajectories we must carry out a final analytic step, namely developing equations that control the moments of the simulated values in the simulated population. These "control" equations must be applied to the forecasts of functional status, represented by GoM scores, so that the moments are consistent with the assumptions of the underlying conceptual model. In particular, the equations for mortality and the autoregressive process for the *g*s requires modification to account for the fact that the *g*s are bounded.

The first equations generalized are for survival. The age specific survival function, l_{t+1}, is updated using the following equation:

$$l_{t+1} = l_t \,|\, I + V_t B_t \,|^{-\frac{1}{2}} \exp\{\frac{\mu_t(v_t) + \mu_t(v_t^*)}{2} - 2\mu_t(\frac{v_t + v_t^*}{2})\}, \tag{9}$$

where $l_{t_0} = 1$ for initial age t_0 and v_t and V_t are age dependent means and covariances of the *g*s. In (9) age dependent coefficients from the integrated quadratic hazard function (i.e., $B_t = 2Q\exp[\theta t]$), adjust survival for risk associated with the age specific average of the K *g*s and their variance-covariances matrix.

Change in the means, v_t, and variances, V_t, of the *g*s due to mortality are calculated using (10) and (11), where v_t^* is calculated for persons expected to survive the interval, and V_t^* is the variance at t for persons expected to survive the interval:

$$v_t^* = \left(v_t - V_t^* Q_t v_t\right) / \Sigma_k \left\langle v_t - V_t^* Q_t v_t \right\rangle_k \tag{10}$$

$$V_t^* = (I + V_t Q_t)^{-1} V_t. \tag{11}$$

Mortality in all three equations is affected both by the mean (i.e., v_t) and variance of the *g*s. The variance is altered by a diffusion process which is a function of time. To describe the diffusion process, assume that the diagonal elements of V_t, reflecting the variance of each g_{ikt}, have, as a maximum, Bernoulli variance (v_{kt} $(1 - v_{kt}$)). To calculate the diffusion component we assume that the correlation of the K sets of gs are constant over age where the correlation matrix R is estimated from the empirical covariance matrix after conditioning on age, and that the ratios of the variance of the gs to the Bernoulli bounds are constant over age with square roots of the ratios defining a diagonal matrix S. If W_{t+1} is a diagonal matrix with the kth element equal to the square root of the Bernoulli variance the covariance matrix for $t+1$ is

$$V_{t+1} = W_{t+1} \, S \, R \, S \, W_{t+1} \, . \tag{12}$$

With (12) diffusion is $\Sigma_{t+1} = V_{t+1} - C_t \, V_t^* \, C_t^T$ with C_t as defined above, and Σ_{t+1} is the variance of the residual in the autoregressive forecasting equation (5). The dynamics of the means of the gs among survivors are given by $v_{t+1} = C_t \, v_t^*$. For further details on these equations, and an application, see Manton et al. (1994).

Equations (9)-(12) provide a mechanism for modeling the consequences of future changes in health and mortality in specific cohorts. In particular, they describe the dynamics of health and mortality as they selectively interact over the lifetime of a cohort. They will be used in sensitivity analyses of alternative assumptions about improved health status among future cohorts of elderly.

Starting Population

The model to be simulated is defined by (a) the equations governing period-to-period change in all endogenous variables, and (b) the initial conditions. In our project, as in virtually all microsimulation undertakings, it will be necessary to merge information from two or more sources to create the initial-conditions sample on which to base the analysis, and to augment data records to compensate for incomplete data. This in turn raises issues of the calibration of sampling weights, and of matching and/or imputation, both of which are addressed below.

The data base representing the initial conditions for the model to be simulated pertain to a population at a point in time (i.e., a cross-section), although the information contained may include lagged values and other retrospective data. Creation of this data base is itself a form of microsimulation, and the resulting data will support descriptive analyses that are themselves of interest. Our plan for creating the initial-conditions data base is as follows: (1) the disabled elderly (age 65+)—both those in the community and those in institutions—will be represented by respondents to the 1989 wave of the National Long-Term Care Survey (NLTCS); and (2) the nondisabled elderly, and nonelderly (age 19-64) populations will be represented by respondents from the 1987-88 National Survey of Families and Households (NSFH).

This plan reflects the special needs imposed by the model to be simulated. In this model, the unit of analysis is the individual, linked via the kin roster to a spouse (if married) as well as to living parents and children. The desire to model kin networks dictates the use of the NSFH and NLTCS data since they, unlike most household surveys, provide rosters of the respondent's nonresident children (the NSFH, but not the NLTCS, includes data on living parents as well).

Pooling NLTCS and NSFH Records for Elderly Respondents. Potential NLTCS respondents initially receive a "screener" interview designed to identify elderly with a functional disability that has lasted, or is expected to last, for three months. Those who pass through the screening process—i.e., are determined to be functionally impaired—provide additional data on detailed interview schedules. The NLTCS screener contains specific, separate questions pertaining to several ADL and IADL items. In contrast, the NSFH includes only four items addressing the respondent's limitations with respect to ADL/IADL tasks, and, if any, the elapsed duration of the limitation (coded less than one month; one-six months; four additional categories). Thus, we cannot perfectly simulate administration of the NLTCS screener to NSFH respondents. Ideally, we would eliminate from the data base NSFH respondents deemed eligible for the NLTCS detailed interview. Additional analysis is necessary to resolve ambiguous cases, that is NSFH respondents neither clearly functionally impaired nor functionally nonimpaired (according to NLTCS criteria). However, the number of ambiguous cases is small: a tabulation of the NSFH shows that of 1,836 elderly respondents, only 58 have functional limitations with durations in either the < one or one-six month categories.

Calibrating and Adjusting Sampling Weights. A key advantage of using the 1989 wave of the NLTCS is that it is a list sample drawn from Medicare records which are known to represent over 97 percent of the target population (individuals aged 65 and older). The list from which the sample was drawn delineates the population as of a date close in time to the 1987-88 period during which the NSFH survey was administered. Analysis of the 1982-84 waves of the NLTCS data, linked to official death records, has revealed high rates of mortality among nonrespondents drawn for the list sample (see Manton et al. 1991b). This analysis permits the calculation of model-based sampling weights to be attached to the attained sample, more accurately reflecting the true mix of health conditions in the population to which the sample is controlled. Similar modeling and reweighting analyses are currently in progress using the third (1989) wave of NLTCS data, which will produce a set of weights representing the distribution of the population by disability status (hence, by proneness to mortality) as of April 1, 1988. We will use this information to calibrate the sampling weights attached to aged 65-and-older individuals in the pooled NLTCS-NSFH sample.

Augmenting Incomplete Data. As noted above, we will encounter missing data when constructing the starting population (some children's sex in the NSFH; existence and characteristics of parents in the NLTCS). Values will be imputed using imputation/matching techniques. Statistical matching of data files, particularly for use with microsimulation, is widely used (see, for example, Budd 1971; Okner 1972; Woodbury 1983; Alter 1974; Paass 1985; Barry

1988; for a recent discussion of the theoretical basis for and statistical properties of matching, see Goel and Ramalingam 1989). In a data-matching exercise, one file of records containing arrays X and Y is matched to an independent second file of records containing arrays X and Z. Synthetic "complete" records are obtained by joining pairs of records from the two files on the basis of their closeness with respect to the common variables in X. Formally, between each record i (=1, ..., n_A) in "target" file A and j (=1, ..., n_B) in "donor" file B we can calculate a distance d_{ij} which is a function of X_i and X_j; an **optimal** (or, "nearest neighbor") match is one that minimizes $\Sigma_{ij}a_{ij}d_{ij}$, where a_{ij} = 0 or 1 (and indicates an assigned match), subject to the constraints $\Sigma_i a_{ij} = 1$ and $\Sigma_j a_{ij} = 1$; this is a linear assignment problem, for which several solution algorithms exist (Goldfarb 1985; Balinski 1986; code for several algorithms is provided in Carpaneto et al. 1988).

The use of matched data files for analysis has been criticized on two grounds. First, inherent in the procedure is a necessity to maintain the assumption that the variables not observed jointly (Y and Z) are conditionally independent given X (Sims 1972, is credited with first making this observation). Second, if only a single "optimal" match is performed, and the resulting data are analyzed as if it were a file of complete observations, then the resulting statistics, including variances, understate the uncertainty associated with the matching process (Rubin 1986; 1987). The multiple-matching approach that we will adopt—in which multiple **sub**optimal matches are performed—addresses both criticisms. In the multiple-match approach, there is not a unique optimally matched file, but multiple files, each being an independent replicate of the starting population data. To ensure the existence of multiple solutions to the file-matching problem it will be necessary to calculate **interval** rather than **point** values of the distance function described in the preceding paragraph. Given the use of interval distances between potentially matched records (equivalently, of a discrete categorization of target and donor records' locations in K_A and K_B space, respectively), not-necessarily-nearest neighbors will be matched, and this will induce a distribution of the variables Z **given** X and Y. While it is impossible to create (through the matching process) an estimate of the unobserved conditional covariance of Y and Z, the variability across the multiple matched files will represent this source of uncertainty in the combined data file. By performing a series of microsimulations on each matched file separately, the variance of final outputs can be approximated using a weighted sum of "within" and "between" estimates (Rubin 1987), reflecting uncertainty about the matching process.

IMPLEMENTATION: THE MICROSIMULATION FRAMEWORK

Programming Considerations

We intend to develop software that can take advantage of future hardware developments while being flexible enough to run in a variety of current environments. This can be accomplished by splitting the program into two parts, the user interface and the microsimulation program itself. The *user interface* allows the user to set and/or override default model parameters and specifications, and to set switches/flags that control a simulation run (all of which are stored in the "parameter file"). The user interface can also monitor the progress of the microsimulation program and manipulate/download/format aggregate-data outputs produced by the microsimulation (and written to the "macrodata file"). The *microsimulation program* (the "*simulator*") will be modular, with a "supervisor" and a number of master routines and their associated subroutines. The basic work of the program is to return a vector X_{t+1} when it is passed the vector X_t.

The simulator interacts with both a micro and a macro data base. The microdata base consists initially of the starting data base, and is modified throughout the simulation run as new outcomes are simulated, year by year. The macrodata base is a file of aggregate indicators and summary measures that are produced during each simulation year; these can be scanned by the user (using the user interface program), and can be downloaded/reformatted to produce summary tables describing the simulated population. Of course, for most applications the final microdata file will also be further analyzed to produce summary output appropriate to the application in question. The main contenders for language to be used are C and FORTRAN, both of which are highly portable. Existing modules for GoM projections are written in FORTRAN, while the discrete-state continuous-time simulation software GENESIS is written in C. Another possibility is the FORTRAN-90 standard, which supports array-based processing (and incorporates the 77 standard as well).

Calibration and Validation

Before any applications of the microsimulation forecasting software can be undertaken, validity checking of the forecasts is required. However, prior to any validity testing extensive debugging and program checking must take place. The debugging phase is followed by validation analyses: internal and external validation, sensitivity analyses, and uncertainty analysis.

Validation. The debugging phase of software development is intended to establish the (mathematical) “correctness” of the software. The next important step is to establish the forecast accuracy—the “reasonableness”—of the forecasts produced by the software. This is, of course, a much more formidable task. Following the thrust of recommendations made by the NAS in its report, we will undertake extensive validity testing of the model’s output by comparing it to actual data sources where possible. We will undertake two types of validity testing, both of which entail comparisons of model forecasts to survey data. Figure 1 displays the overlap of several existing and forthcoming data sources with the time period/cohort coverage of our proposed microsimulations. For example, the 1992-93 follow up to the NSFH will include substantial overlap with our simulated data, as will the 1994 wave of the NLTCS. The first type of validity test, *parameter* validation, will be used with these two data sources, since together they serve as the starting population for the simulations. Furthermore, many of the equations in the model are estimated using data from the pre-1994 waves of the NLTCS. Discrepancies between the predicted values (for 1992, among the NSFH-based portion of the starting population; in 1994, among the NLTCS-based portion of the starting population) will indicate areas for re-estimation of model parameters. What we are calling parameter validation can also be done with other panel data sets (such as HRS and AHEAD). In these cases, it consists of applying estimated parameters to a baseline file (for example wave 1 of the HRS) and projecting forward to a time coincident with a subsequent survey data (e.g., wave 2 of the HRS), and comparing the projected with the actual data. This test can only be performed for the parts of the model that pertain to variables represented in the given panel data set (for example, we would not be able to test the parameters of the functional status equations using the HRS). However, it has the advantage that the definition of the sample population from which the external data come is constant at each point of comparison (apart from sample attrition).

The second type of validity testing consists of comparisons of model output to actual data, i.e., external validation. This does not require that the external data source be panel data, simply that there be overlap of age groups and measured variables between the simulated data and the empirical (external) data. Again, discrepancies between the two data sources (given reasonable alignment of the populations compared and variables measured, across the two data sources) indicate possible areas for re-estimation or respecification of equations in the forecasting model, and/or sensitivity testing. In all cases, the actual data to which model forecasts will be compared will be sample survey data, which while generated by actual people is nonetheless subject to sampling and nonsampling errors. Since sample survey data produce interval estimates of population values, and our microsimulation software should also be viewed as producing uncertain estimates, we will end up having to compare two uncertain (error-laden) indicators of (nominally) the same phenomenon. Because of this, we view analysis of the uncertainty associated with microsimulation output as an important question, one to which we intend to devote considerable attention.

Sensitivity Analysis. Another type of model validation stressed in the NAS report is sensitivity testing. We have noted, at several points in the preceding discussion, areas in which a need for sensitivity testing will take place. For example, survivorship and marital status changes within kin networks will depend on projected mortality rates and rates of marital status change. Alternative assumptions regarding the time path of mean rates is one candidate for sensitivity analysis. Also, the income-dynamics equation [equation (3a)] will condition on aggregate year-to-year productivity growth, and a future time path for this variable must be specified prior to any simulation run. Using a combination of published sources and analysis of past trends, we will identify a range of assumed time paths of this variable; these will also be the basis for sensitivity testing. Finally, (as noted in the NAS report) sensitivity testing should be used to assess the consequences of alternative model specifications (e.g., functional form, interactions, etc.).

Uncertainty Analysis. The NAS report on microsimulation noted the need for increased attention to uncertainty analysis, and surveyed some of the technical issues involved (Cohen 1991). We plan to address this issue in depth. There are three distinct sources of variability in the projections obtained from microsimulation that are associated with limitations of the underlying data: (1) the projections begin with a starting population that uses data from sample surveys, and therefore is subject to sampling variability; (2) one or more of the elements in the data array used to describe members of the starting population may have been unobserved for part or all of the sample, and these missing data elements may have been replaced by imputed or statistically matched variables which are imperfect (i.e., uncertain) replacements for the actual values; and (3) the parameters of the projection model are, in nearly every case, derived from statistical analysis of other data sources, and therefore are themselves subject to sampling error. Because of these sources of variability, the output of each "run" of a microsimulation should be viewed as a sample from a distribution of such outputs. Another general source of uncertainty arising with survey data is *non*sampling error, which encompasses all errors not attributable to the sampling process including nonresponse, measurement, and frame error (Lessler and Kalsbeek 1992). Each is a candidate for investigation, and for some important types (e.g., errors attributable to use of proxy respondents in the NLTCS) precise estimates can be made (Corder et al. 1993).

The propagation of sampling error in the starting population through the outcomes of a microsimulation has not been previously addressed. A number of well-established statistical methods developed for estimation of variance in complex sample surveys are applicable here, in particular those that make use of resampling or reweighting cases in the starting data sets such as the bootstrap and the jackknife (Cohen 1991), which are applicable to data obtained from complex surveys. These methods may be applied in the microsimulation context by resampling from the simulated trajectories. Typical advantages of these resampling methodologies are that they are fairly straightforward to apply even for complex statistics and for data from complex survey designs. A particular advantage in microsimulation is that the estimated sampling variances and covariances of any summaries of interest at any time can be calculated without rerunning the simulations. Thus, the creation of the simulation database, the most computationally expensive part of the microsimulation exercise, need not be replicated for variance estimation.

The second source of uncertainty, "imputation error," has been the object of considerable investigation in recent years. The multiple imputation methodology treats filled-in missing data with great generality. Multiple imputation (Rubin, 1978, 1987) involves replacing the set of missing values with $m \geq 2$ sets of plausible values. The observed data set supplemented with the m sets of imputations is called a multiply-imputed data set, and can be used to create m completed data sets. A substantial literature supports the efficacy of multiple imputation. Rubin (1977a, 1977b, 1979a, 1979b), Herzog and Rubin (1983), Heitjan and Rubin (1986), Schenker and Welsh (1985), and Rubin and Schenker (1986, 1987) develop procedures for scalar parameters. Li (1985), Rubin (1987), Raghunathan (1987), Weld (1987), Treiman, Bielby, and Cheng (1988), and Schenker, Treiman, and Weidman (1988) consider multiparameter situations.

In the first step of the multiple imputation procedure (the "modeling task"), models are fitted that predict the missing data and values of the parameters of these models are drawn from their posterior distribution, representing variability due to our uncertainty about the exact values of the parameters (these models can be as simple as cell means in a multiway table). In the second step (the "imputation task") the missing data are drawn from their predictive distribution given the parameters. By repeating the two steps several times, multiple versions of the imputed missing data are created. Inference based on the several completed data sets can be combined to reflect the added uncertainty due to the missing data, although only standard complete-data methods of analysis need be applied to each completed data set in an actual microsimulation model (Rubin, 1983; Rubin and Schenker, 1986, 1987; Clogg, et al. 1991; Treiman, Bielby and Cheng, 1988; Rubin, 1986).

The third source of uncertainty, that due to the sampling distributions of parameters used in the microsimulation, can also be modeled using multiple imputation methodology. The parallel here consists of recognizing all the statistical analysis that goes into developing the input parameters for the microsimulation as the "modeling task", while the actual running of the microsimulation is the "imputation task." As with multiple imputation for missing data, the model parameter values must be drawn from their posterior distributions given various data sets used to estimate the models, to properly represent uncertainty regarding the true values of the parameters. However, in virtually all applications of microsimulation (within the class of microsimulation efforts relevant to this proposal) analysts have obtained only *one* realization, rather than multiple realizations, of the random process represented by the "modeling task." We are aware of only one previous attempt to compute the variance of estimators derived from a large-scale microsimulation exercise, and in this case (Lohrer 1980) the covariances of only two out of the many parameters in the model were considered.

The covariance matrix of **all** the parameters of the model will contain numerous zeros. Simplifying slightly, the model will have four "blocks" of jointly-estimated parameters, say Θ_1 (quadratic mortality), Θ_2 (autoregressive function for functional status/income), Θ_3 (income/retirement/marital status), and Θ_4 (mortality/marital status intensities), each with corresponding sample covariance matrix S_m (m=1,...,4). The full covariance matrix is then block-diagonal, with zeros in place of missing covariances between blocks. We will be able to calculate the contribution to uncertainty of each of the four blocks of parameters, each time treating the other three as fixed. Only if we assume that the *true* covariances are zero can we sum the four components.

We plan to use "Latin hypercube" sampling from the posterior distribution of model parameters, which has been shown to be much more efficient than random sampling with the same number of replications (McKay et al. 1979). This entails partitioning the parameter space into equiprobable regions, and drawing a stratified sample of parameter values. An algorithm developed by Iman and Conover (1982) to deal with correlated parameters will be used in our analysis.

A microsimulation run may be regarded as an outcome of a simulated experimental trial with a number of random-effects (e.g., parameter sets, starting population replicates) and fixed-effects (e.g., functional form) factors set at selected levels. The number of conditions in a full factorial simulation experiment—one in which all possible combinations of the factors are given at least one trial—is potentially very large, possibly well beyond the range of computational feasibility. In these simulation experiments, as in a physical experiment, a fractional factorial experiment may produce adequate results at a fraction of the cost of a full factorial design. In a fractional factorial simulation experiment, a subset of the possible conditions are selected for simulation. The experiment must be carefully designed to take advantage of these typical properties of factorial experiments: (a) main effects (i.e., effects of altering a single factor while keeping others constant, averaged over levels of other factors) are typically larger than interactions between factors, and of more analytical interest; as a result, each main effect can be estimated by averaging over only a few conditions of the other factors; (b) main effects of factors may be adequately estimated from a sample as small as or smaller than that required for estimating an overall mean; and (c) one well-designed set of trials can be used to estimate main effects. Given such a design, the subsequent analysis takes the form of an analysis of variance (ANOVA). The outcome variables in this ANOVA are meaningful summaries of the simulation results, and the results include: (a) estimates of main effects of the systematic factors; (b) estimates of error variance associated with any random factors (or sets of random factor); (c) an overall standard error for each variable of interest that can be used to calculate a predictive interval for the corresponding variable; and (d) a summary of the sources of variability that most affect the overall precision of these results.

Parallel-Processing Applications

We plan to make use of a parellel-processing computer as a platform for the uncertainty analysis described above. A likely schematic for the requisite software is one in which one CPU is designated a “supervisor” (the *S* task) which initiates, monitors, and interfaces with the other

CPUs; up to 30 CPUs have a resident copy of the simulator, each using a different (and independently generated) replicate of the starting population and/or input parameter set in accordance with the fractional factorial design discussed previously (the *R1*, ... , *Rn* tasks); and, one or more CPUs would be assigned to creating a running summary of the endogenous data bases being produced, including quantifying the uncertainty in the output. The parallel-processing approach that we will use rules out the need to save a large number of large output files.

Applications of the Microsimulation

Many applications of the microsimulation program will be possible. We will undertake several applications to demonstrate the capabilities of the model and to address key policy and research questions. The strength of the modeling effort lies in its ability to allow us to observe both the interactions among a number of socio-demographic variables at any future point in time, and to look at such variables through time as well. Although the exact form of these simulations will depend upon what issues are particularly salient when the applications are conducted, we can identify several promising applications: (1) We could model policies that attempt to encourage caregiving by relatives to determine what impact they might have on need for LTC services in the future. Some policy makers have considered limited proposals such as tax benefits to try to encourage caregiving, but little is known about the extent to which there are relatives who are potential caregivers for older persons in need of care. This model would allow us to examine to what extent the supply of potential caregivers matches well to those who are in need. Only then will we be able to assess the likely success of incentive programs for relatives to provide aid. (2) At a more basic level, it would be of interest to determine the extent to which children of elderly parents are at risk of being a caregiver and, given this risk, the extent and distribution of "burden" of caregiving. The risk status depends not only on the disability status of the elderly, but also on the family situation of each child, which determines the pool of potential caregivers. Both the risk of being a caregiver, and the burden of care actually given, can be approached from both a period and a cohort (lifetime) perspective. Our microsimulation analysis will provide a much richer and complete accounting of this highly charged issue than any existing approaches. (3) A third application, suggested by recent findings on declines in disability prevalence (specific to age, sex, marital status, etc.), is to simulate the continuance of such declines and their effects on LTC use, ALE, and the economic status of the elderly.

Further applications of the microsimulation model can be envisioned, using what is sometimes called "post-processing." By this we mean taking a simulated future year (*t*), and treating the data base "at" time *t* (i.e., the cross section of X_ts and lagged values if desired) as a sample from a hypothetical population. Behavioral relationships conditioned on variables "observed" in the hypothetical population can then be superimposed, providing still more endogenous outcomes. This could be used, for example, to project family-provided community care arrangements into a future year, based on the cross-sectional model presented in Wolf and Soldo (1990). Behavioral responses to policy variables can be grafted on to the analysis in this way as well. For example, variables representing state-to-state variation in Medicaid policy (such as those recently assembled by Sloan, undated) could be imputed to observations in a forecast year; these could be made to agree with observed marginal distributions (and joint distributions involving individual-level variables) from a base year. Implied values of outcome variables (e.g., use of community-based formal care services) could then be determined using estimated behavioral relationships. While this strategy has its limitations, the alternative (in the microsimulation context) requires a full-blown model of interstate migration and possibly of the choice by states of policy instruments, and therefore is infeasible within the scope of the project.

EXTENSIONS TO THE PROJECT

The work planned is viewed as part of a larger agenda that includes model development over a longer time horizon and encompassing a broader range of endogenous outcomes. One possible extension to the work proposed here is to embed behavioral relationships modeled in the ongoing research on intergenerational transfers into the microsimulation to be developed. A similar extension can be envisioned with respect to another ongoing project in which the migration histories of parents and their adult children are jointly modeled, and the association between spatial proximity and migration is determined. Other desirable extensions to the model include: equations for private pension coverage, pension plan attributes and pension wealth accrual that faithfully tracks the recent flattening of the postwar trend in coverage; determination of health service use for both acute and chronic conditions; the dynamics of formal community-based LTC service use; and explicit treatment of the "spend down" of assets in connection with Medicaid eligibility and institutionalization; and (as discussed above) equations for fertility, which would allow us to make forecasts for a broader group of initial cohorts.

ENDNOTES

[*]Douglas Wolf is Professor of Public Administration and Director of the Maxwell Center for Demography and Economics of Aging at Syracuse University. Jan Ondrich is Associate Professor of Economics at Syracuse University. Ken Manton, Eric Stallard, Max Woodbury and Larry Corder are Research Professors associated with the Center for Demographic Studies at Duke University. This document is based on the proposal for the project that it describes. Important contributions to the development of that project were made by Donald Rubin and Alan Zaslavsky. Helpful comments were received from Michael Cohen and Debra Dwyer.

REFERENCES

Abowd, J.M. and D. Card. (1989). "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57: 411-45.

Alter, H.E. (1974). "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970," *Annals of Economic and Social Measurement*, 3/2: 373-94.

Balinski, M.L. (1986). "A Competitive Simplex Method for the Assignment Problem," *Mathematical Programming*, 34: 125-41.

Barry, J.T. (1988). "An Investigation of Statistical Matching," *Journal of Applied Statistics*, 15: 275-83.

Becker, G.S., E.M. Landes, and R.T. Michael. (1977). "An Economic Analysis of Marital Instability," *Journal of Political Economy*, 85: 1141-1187.

Bell, F.C., A.H. Wade, and S.C. Goss. (1992). "Life Tables for the United States Social Security Area 1900-2080." SSA Pub. No. 11-11536, U.S. Department of Health and Human Services, Social Security Administration, Baltimore, MD.

Budd, E.C. (1971). "The Creation of a Microdata File for Estimating the Size Distribution of Income," *Review of Income and Wealth*, 17: 317-33.

Burkhauser, R.V. and G. Duncan. (1988). "Life Events, Public Policy and the Economic Vulnerability of Children and the Elderly," In J. Palmer, T. Smeeding, and B. Torrey (eds.), *The Vulnerable*. Washington, DC: Urban Institute Press, pp. 55-86.

Burkhauser, R.V. and G. Duncan. (1991). "U.S. Public Policy and the Elderly: The Disproportionate Risk to the Well-Being of Women," *Journal of Population Economies*, 21: 217-231.

Burkhauser, R.V. and J.F. Quinn. (1994). "Retirement and Labor Force Behavior." In L. Martin and S. Preston (eds.), *The Demography of Aging*. Washington, DC: National Academy of Science Press, pp. 50-101.

Carpaneto, G., S. Martello, and P. Toth. (1988). "Algorithms and Codes for the Assignment Problem," *Annals of Operations Research*, 13: 193-223.

Citro, C.F. and E.A. Hanushek (eds.). (1991). *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling* (Volume I). Washington, D.C.: Committee on National Statistics, National Research Council.

Clogg, G., D.B. Rubin, D. Schenker, B. Schultz, and L. Weidman. (1991). "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, 86: 68-78.

Cohen, Michael (1991). "Variance Estimation of Microsimulation Models through Sample Reuse." In C.F. Citro and E.A. Hanushek (eds.), *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling* (Volume II). Washington, D.C.: Committee on National Statistics, National Research Council.

Corder, L.S., M.A. Woodbury, and K.G. Manton. (1993). "Health Loss Due to Unboserved Morbidity: A Design Based Approach to Minimize Nonsampling Error in Active Life Expectation Estimates." In J.M. Robine et al. (eds), *Calculation of Health Expentancies: Harmonization, Consensus Achieved, and Future Perspectives*. Montrouge: John-Libbey.

Coughlin, T. A., T.D. McBride, M. Perozek, and K. Liu. (1990). "Home Care for the Disabled Elderly: Predictors and Expected Costs." 3972-01. Washington, DC: The Urban Institute.

Duncan, G., and K. Smith. (1988). "The Rising Affluence of the Elderly: How Far, How Fair, and How Frail?"unpublished manuscript.

Goel, P.K. and T. Ramalingam. (1989). *The Matching Methodology: Some Statistical Properties*. New York: Springer-Verlag.

Goldfarb, D. (1985). "Efficient Simplex Algorithms for the Assignment Problem," *Mathematical Programming*, 33: 187-203.

Goldman, Noreen. (1993). "Marriage Selection and Mortality Patterns: Inferences and Fallacies," *Demography*, 30: 189-208.

Gottschalk, P. and R. Moffitt. (1993). "Trends in the Covariance Structure of Earnings in the U.S.: 1967-87." Department of Economics Working Paper No. 93-9. Prividence, RI: Brown University.

Grad, S. (1990). "Earnings Replacement Rates of New Retired Workers," *Social Security Bulletin*, 53 (October): 2-19.

Guralnik, J.M., M. Yanaagishita, and E.L. Schneider. (1988). "Projecting the Older Population of The United States: Lessons from the Past and Prospects for the Future," *Milbank Quarterly,* 66: 283-308.

Hause, J.C. (1985). "The Fine Structure of Earnings and the On-the-Job Training Hypothesis," *Econometrica*, 52: 1013-30.

Heckman, J.J. and J.R. Walker. (1987). "Using Goodness of Fit and other Criteria to Choose among Competing Duration Models: A Case Study of Hutterite Data." In Nancy Tuma (ed.), *Sociological Methodology 1987*. San Francisco: Jossey-Bass.

Heckman, J.J. and J.R. Walker. (1990). "The Relationship Between Wages and Income and the Timing and Spacing of Births: Evidence From Swedish Longitudinal Data," *Econometrica*, 58: 1411-1441.

Heitjan, D.F. and D.B. Rubin. (1986). "Inference From Coarse Data Using Multiple Imputation." In T.Boardman (ed.), *Proceedings of the 18th Symposium of the Interface of Computer Science and Statistics*. Washington, DC: American Statistical Association, pp. 138-143.

Herzog, T. and D.B. Rubin. (1983). "Using Multiple Imputations to Handle Nonresponse in Sample Surveys." In W.G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data Sample Surveys (Vol. 2): Theory and Bibliographies*. New York: Academic Press, pp. 209-245.

Holtz-Eakin, D. and T.M. Smeeding. (1992). "Income, Wealth and Intergenerational Economic Relations." Prepared for NAS Committee on Population Workshop on the Demography of Aging, December 10-11.

Hougaard, P., B. Harvald, and N.V. Holm. (1992). "Measuring the Similarities between the Lifetimes of Adult Danish Twins Born between 1881-1930," *Journal of the American Statistical Association,* 87: 17-24.

Hu, Y. and N. Goldman. (1990). "Mortality Differentials by Marital Status: An International Comparison," *Demography,* 27(2): 233-250.

Iman, R.L. and W.J. Conover. (1982). "A Distribution-Free Approach to Inducing Rank Correlation among Input Variables," *Communications in Statistics-Simulation and Computation*, 11(3): 311-334.

Johnson, W.R. and J. Skinner. (1986). "Labor Supply and Marital Separation," *The American Economic Review*, June: 455-469.

Katz, S., L.G. Branch, M.H. Branson, J.A. Papsidero, J.C. Beck, and D.S. Greer. (1983). "Active Life Expectancy," *New England Journal of Medicine*, 309: 1218-1223.

Kobrin, F.E. (1976). "The Fall in Household Size and the Rise of the Primary Individual in the United States," *Demography,* 13: 127-138.

Le Bras, H. (1978). "Parents, Grandparents, Bisiaeux" (English translation). In K.W. Wachter, E.A. Hammell and P. Laslett (eds.), *Statistical Studies of Historical Social Structure*. San Diego, CA: Academic Press.

Lee, R. (1992). "The Lee-Carter Approach to Forecasting Life Expectancy." Presented at the NAS/IOM workshop on forecasting survival, health, and disability, Washington, D.C., March 9-10 1992.

Lerman, S. and C. Manksi. (1981). "On the Use of Simulated Frequencies to Approximate Choice Probabilities." In C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.

Lesser, J.T. and W.D. Kalsbeek. (1992). *Nonsampling Error in Survey*s. New York: Wiley and Sons Publishers.

Levy, F. (1988). "Incomes, Families, and Living Standards." In R.E. Litan et al. (eds.), *American Living Standards: Threats and Challenges*. Washington, DC: Brookings Institute.

Li, K.H. (1985). "Hypothesis Testing in Multiple Imputation-with Emphasis on Mixed-up Frequencies in Contingency Tables." Ph.D. Thesis, Department of Statistics, University of Chicago.

Lillard, L.A. and Y. Weiss. (1979). "Components of Variation in Panel Earnings Data: American Scientists, 1960-1970," *Econometrica*, 47: 437-54.

Lillard, L.A. and R. Willis. (1978). "Dynamics Aspects of Earnings Mobility," *Econometrica,* 46: 985-1012.

Lohrer, M. (1980). "Estimation of the Variance of Prediced Labor Supply Responses." Working Paper Series SRI-SD21. SRI International, Socioeconomic Research Center.

Macunovich, D.J., R.A. Easterlin, E.M. Crimmins, and C.M. Schaeffer. (1995). "Echoes of the Baby Boom and Bust: Recent and Prospective Changes in Intergenerational Living Arrangements of Elderly Widows in the United States," *Demography,* 32: 17-28.

MaCurdy, T. (1982). "The Use of Time Series Processes to Model the Error Structure of Earnings in a Longitudinal Data Analysis," *Journal of Econometrics,* 18: 83-114.

Manton, K.G. (1988). "A Longitudinal Study of Functional Change and Mortality in the United States," *Journal of Gerontology,* 43: 153-161.

Manton, K.G., L. Corder, and E. Stallard. (1993a). "Changes in the Use of Personal Assistance and Special Equipment from 1982 to 1989: Results from the 1982 and 1989 NLTCS, *The Gerontologist*, 33: 168-176.

Manton, K.G., L. Coder, and E. Stallard. (1993b). "Estimates of Change in Chronic Disability and Institutional Incidence and Prevalence Rates in the U.S. Elderly Population From 1982, 1984, and 1989 National Long Term Care Survey," *Journal of Gerontology: Social Sciences,* 48: S153-S166.

Manton, K.G. and E. Stallard. (1991). "Cross-Sectional Estimates of Active Life Expentancy for the U.S. Elderly and Oldest-Old Populations," *Journal of Gerontology,* 46: S170-S182.

Manton, K.G. and M.A. Woodbury. (1985). "A Continuous-Time Multivariate Gaussian Stochastic Process Model of Change in Discrete and Continuous State Variables." In N. Tuma (ed.), *Sociological Methodology 1985*. San Francisco: Jossey-Bass.

Manton, K.G., E. Stallard, and J.W. Vaupel. (1981). "Methods for Comparing the Mortality Experience of Hetergeneous Populations," *Demography,* 18: 389-410.

Manton, K.G., E. Stallard, and J.W. Vaupel. (1986). "Alternative Models for the Heterogeneity of Mortality Risks among the Aged," *Journal of the American Statistical Association*, 81: 635-644.

Manton, K.G., E. Stallard, and J.W. Vertrees. (1991a). "Estimates of Costs of Medical Conditions in Elderly Populations." Working Paper Series No. M372, Center for Demographic Studies. Durham, NC: Duke University.

Manton, K.G., E. Stallard, and M.A. Woodbury. (1991b). A Multivariate Event History Model Based Upon Fuzzy States: Estimation from Longitudinal Surveys with Informative Nonresponse, *Journal of Official Statistics*, 7: 261-293.

Manton, K.G., M.A. Woodbury, and E. Stallard. (1991c). "Statistical and Measurement Issues in Assessing the Welfare Status of Aged Individuals and Populations," *Journal of Econometrics*, 50: 151-181.

Manton, K.G., E. Stallard, M.A. Woodbury, and J.E. Dowd. (1994). "Time Varying Covariates in Models of Human Mortality and Aging: Multidimensional Generalizations of the Gompertz," *Journal of Gerontology: Biological Sciences*, 49: B169-B190.

Manton, K.G., M.A. Woodbury, L.S. Corder, and E. Stallard. (1992). "The Use of Grade of Membership to Estimate Regression Relationships," *Sociological Methodology*, (Marsden P. Ed.), forthcoming.

McFadden, D. (1989). "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica*, 57: 995-1026.

McKay, M.D., R.J. Beckman, and **??** Conover. (1979). "A Comparison of Three Methods for Selecting Value of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, 21: 239-245.

Michael, R., V. Fuchs, and S. Scott. (1980). "Changes in the Propensity to Live Alone: 1950-1976," *Demography*, 17: 39-56.

Narayanan, A. (1990). "Computer Generation of Dirichlet Random Vectors," *Journal of Statistical Computation and Simulation*, 36: 19-30.

National Center for Health Statistics. (1987). *Vital Statistics of the United States, 1987: Volume III, Marriage and Divorce*. Washington, DC: U.S. Government Printing Office.

National Center for Health Statistics. (1993). "Advance Report of Final Natality Statistics." Monthly vital statistics report Vol 41:9. Maryland: Public Health Service.

Okner, B.A. (1972). "Constructing a New Data Base From Existing Microdata Sets: The 1986 Merge File," *Annals of Economic and Social Measurement*, 1/3: 325-42.

Orchard, R. and M.A. Woodbury. (1971). "A Missing Information Principle: Theory and Application." In L.M. LeCam, J. Neyman, and E.L. Scott (eds.), *Proceedings Sixth*

Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA: University of California Press, pp. 697-715.

Orcutt, G. (1957). "A New Type of Socio-Economic System," *Review of Economics and Statistics*, 39: 773-797.

Paass, G. (1985). "Statistical Record Linkage Methodology," *45th session of the International Statistical Institute*.

Pakes, A. and D. Pollard. (1989). "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57: 1027-57.

Quinn, J.F. (1987). "Economic Status of the Elderly: Beware of the Mean," *Review of Income and Wealth*, March: 63-82.

Raghunathan, T.E. (1987). "Large Sample Significance Levels From Multiply-Imputed Data," Ph.D. Thesis, Harvard University.

Rivlin, A.M. and J.M. Wiener. (1988). *Caring for the Disabled Elderly: Who Will Pay?* Washington, DC: Brookings Institute.

Rogers, A., R. Rogers, and A. Belanger. (1989). "Active Life among the Elderly in the United States: Multistate Life-Table Estimates and Population Projections," *The Milbank Quarterly*, 67: 370-411.

Rubin, D.B. (1977a). "Formalizing Subjective Notions About the Effects of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association,* 72: 538-543.

Rubin, D.B. (1977b). "The Design of a General and Flexible System for Handling Non-Response in Sample Surveys." Manuscript prepared for the U.S. Social Security Administration.

Rubin, D.B. (1978). "Multiple Imputations in Sample Survey-A Phenomenological Bayesian Approach to Nonresponse." In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20-34. (Also with addendum, discussion, and rejoinder in *Imputation and Editing of Faulty or Missing Survey Data*. Washington, D.C.: U.S. Department of Commerce, pp. 1-23.)

Rubin, D.B. (1979a). "Illustrating the Use of Multiple Imputations to handle Nonresponse in Sample Surveys." In *Proceedings of the ISI-IASS Manila*. Voorburg, The Netherlands: International Statistical Institute.

Rubin, D.B. (1979b). "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74: 318-328.

Rubin, D.B. (1986). "Statistical Matching Using File Concentration with Adjusted Weights and Multiple Imputations," *Journal of Business and Economic Statistics,* 4: 87-94.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley and Sons Publisher.

Rubin, D.B. and N. Schenker. (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," *Journal of the American Statistical Association*, 81: 366-374.

Rubin, D.B. and N. Schenker. (1987). "Interval Estimation from Multiply-Imputed Data: A Case Study Using Census Agriculture Industry Codes," *Journal of Official Statistics*, 3: 375-387.

Schenker, N. and A.H. Welsh. (1985). "Multiple Imputation for Interval Estimation from Surveys with Ignorable Nonresponse." Ph.D. Dissertation, Department of Statistics, University of Chicago.

Sims, C.A. (1972). "Comment" [on Okner], *Annals of Economic and Social Measurement*, 1/3: 343-5.

Singer, B. (1989). "Grade of Membership Representations: Concepts and Problems." In T.W. Anderson, Krishna. B. Athreya, and Donald L. Iglehart (eds), *Probability, Statistics, and Mathematics: Papers in Honor of Samuel Karlin*. New York: Academic Press, Inc., pp. 317-334.

Sloan, F.A. (undated). "1991 State Medicaid Policies Regarding the Elderly in Need of Long-Term Care." Vanderbilt University: Health Policy Center, mimeo.

Soldo, B.J. and K.M. Manton. (1985a). "Demographic Challenges for Socioeconomic Planning," *Socio-Economic Planning Sciences*, 19: 227-48.

Soldo, B.J. and K.M. Manton. (1985b). "Health Status and Service Needs of the Oldest Old: New Perspectives and Future Trends," *Milbank Quarterly*, 63: 286-319.

Tolley, H.D. and K.G. Manton. (1991). "Large Sample Properties of a Fuzzy Partition," *Journal of Statistical Mathematics*.

Treiman, D.J., W. Bielby, and M. Cheng. (1988). "Evaluating a Multiple-Imputation Method for Recalibrating 1970 Census Detailed Industry Codes to the 1980 Standard," *Sociological Methodology* 18.

U.S. Bureau of the Census. (1992). Current Population Reports, Special Studies, P23-178RV, *Sixty-five Plus in America*. Washington, DC: U.S. Government Printing Office.

Vaupel, J.W. and A. Yashin. (1985). "The Deviant Dynamics of Death in Heterogeneous Populations." In N. Tuma (ed.), *Sociological Methodology 1985*. **City:** Jossey-Bass Publishers.

Weld, L.H. (1987). "Significance Levels from Public-use Data with Multiply-Imputed Industry Codes." Ph.D. Dissertation, Department of Statistics, Harvard University.

Wiener, J.M., R.J. Hanley, R. Clark, and J.F. Van Nostrand. (1990). "Measuring the Activities of Daily Living: Comparisons across National Survey," *Journal of Gerontology*, 45: S229-S237.

Wilkins, R. and O. Adams. (1983). *Healthfulness of Life*. Montreal, Canada: Institute for Research on Public Policy.

Wolf, D.A. (1986). "Simulation Methods for Analyzing Continuous-Time Event-History Models," *Sociological Methodology 1986*. San Francisco: Jossey-Bass.

Wolf, D.A. (1988). "Kinship and Family Support in Aging Societies." In UN Population Division, *Economic and Social Aspects of Population Aging*. New York: The United Nations.

Wolf, D.A. (Forthcoming). "Changes in the Households of Older Women: A Cross-National Study." Presented at the annual meetings of the Population Association of America, Baltimore, Maryland. Forthcoming in *The Gerontologist*..

Wolf, D.A. (1994). "Coresidence with An Elderly Parent: Lifetime Patterns and Sensitivity to Demographic Change." In UN Population Division, *Aging and the Family*. New York: The Untied Nations.

Wolf, D.A. and B.J. Soldo. (1990). "Family Structure and Caregiving Portfolios." Presented at the annual meetings of the Gerontological Association of America, Boston, November 9, 1990.

Wolf, D.A. and B.J. Soldo. (1991). "The Dynamics of Disability, Family Composition, and Caregiving." Presented at the annual meetings of the Gerontological Association of America, San Francisco, November 17, 1991.

Woodbury, M.A. (1983). "Statistical Record Matching For Files." In W.G. Madow and I. Olkin (eds.), *Incomplete Data in Sample Surveys (Vol 3): Proceedings of the Symposium*. New York: Academic Press.

Woodbury, M.A., J. Clive, and A. Garson Jr. (1978). "Mathetical Typology: A Grade of Membership Technique for Obtaining Disease Definition," *Computers and Biomedical Research*, 11: 277-298.

Woodbury, M.A. and K.G. Manton. (1977). A Random Walk Model of Human Mortality and Aging," *Theoretical Population Biology*, 11: 37-48.

Woodbury, M.A., K.G. Manton, and L.S. Corder. (1992). "Representing Changes Over Time in the Grade of Membership Model." Duke University Center for Demographic Studies.

World Health Organization. (1984). "The Uses of Epidemiology in the Study of the Elderly." Report of a WHO Scientific Group on the Epidemiology of Aging. Technical Report Series 706, WHO, Geneva.

Center for Policy Research
Tim Smeeding, Director
Richard V. Burkhauser, Associate Director for Aging Studies

Center for Demography and Economics of Aging
Douglas A. Wolf, Director

The Maxwell Center for Demography and Economics of Aging was established in September 1994 with core funding from the National Institute on Aging (Grant No. P20-AG12837). The goals of the Center are to coordinate and support research activites in aging, publish findings and policy analyses, and convene workshops, seminars, and conferences intended to train data users and disseminate research findings and their implications to the research and policy communities.

The Center produces two paper series, the Aging Studies Paper Series and Papers in Microsimulation.

Many Center publications are available via anonymous ftp. To access these resources attach to our World-Wide Web home page at http://gerosun.syr.edu.

For more information contact: Martha Bonney (mbonney@maxwell.syr.edu).

Center for Policy Research
426 Eggers Hall
Syracuse University
Syracuse, New York 13244

(315) 443 3114

FAX: (315) 443 1081

PAPERS ON MICROSIMULATION SERIES

Paper No.	Title	Authors	Date
1	A Model for Simulating Life Histories of the Elderly: Model Design and Implementation Plans	Wolf, Ondrich, Manton, Stallard, Woodbury, and Coder	August 1995
2	Family Structure and Institutionalization: Results from Merged Data	McNally and Wolf	August 1995