A MICROSIMULATION MODEL FOR HEALTH STATUS PROJECTIONS

George C. Myers
Alfred M. Pitts
Roger Hillson
Eric Stallard

Duke University

INTRODUCTION

The development of a microsimulation model that captures the sickness-death process has been a central focus of several research projects undertaken in the past few years by the Center for Demographic Studies at Duke University. Implicit in this development has been the intent that such a model provides a suitable framework for producing national population projections that contain not only age, sex and race specificity, but also estimates of the health status of these population sub-groups. Moreover, the model affords the researcher an experimental tool for assessing the changes that may be experienced in the incidence of specific diseases and the probabilities of dying (or surviving) from the diseases. In this paper, the main features of the model are described and two applications are discussed.

The projection of the health status of future national populations is clearly of great importance in anticipating the demands that will arise in medical manpower, facilities and fiscal support systems. Moreover, it is likely that existing differentials in health status by age, sex, race, socioeconomic and other social characteristics will persist to varying degrees in the future. Thus, these health status projections must be disaggregated for important segments of the population if they are to be responsive to the growing concerns with health policy formulations and health service program planning.

Demographic specialists concerned with national population projections have largely ignored considerations of health status change and disease processes in examining the differentiation of mortality risk in human populations. Although the compositional variables that are typically used to explain changes in mortality might be regarded as major, implicit correlates of morbidity, these projection models tend to view changes in aggregate survivorship as arising *pari passu* with changes in sociodemographic composition alone. The processes of disease onsets, virulence, and recovery, through which the effects of compositional change are transmitted, have not been a traditional projection concern.

MODEL FOR PROJECTING HEALTH STATUS

The health status model that has been developed is quite straightforward in its conceptualization. The general approach is similar to that of the POPSIM simulation model developed by Horvitz, *et. al.*(1) although there are a number of important differences that cannot be discussed here. The model constructs for each member of a random sample of the United States population, a health status history by stochastically exposing the individual to a set of health status change probabilities. Although these probabilities are allowed to vary cross-sectionally by age, race, and current health status, in effect, we are projecting the future health status of the population to the year 2000 on the assumption that the probabilities observed in 1970 remain unchanged. After repeating these simulations for every member of the sample sufficiently often to assure that the relative age distributions of deaths and population are free of random experimental error, the results for the sample are extrapolated to the national population.

Health statuses are divided into three acute and eight chronic conditions; these conditions are indicated in the tables that follow. In addition, individuals also can die from external causes (e.g., accidents). However, the model does not capture the temporary or permanent disability that might result from a non-fatal external incident, though nothing in the structure of the model precludes such a refinement.

Paper presented at the Annual Meetings of the American Statistical Association held August 15-18, 1977, in Chicago, Illinois. Support for this research was from a contract from the Adult Development and Aging Branch of the National Institute of Child Health and Human Development, No. 1-HD-2-267, and a contract from the National Academy of Sciences' Committee on Health Care Resources in the Veterans Administration.

The model projects a life history for each member of the simulation sample by establishing the time of disease onsets and deaths, on the basis of a number of simplifying assumptions.

First, the model assumes that the onset or presence of any one health status condition is not correlated with the onset or presence of any other condition or group of conditions (except via an indirect path through mortality). Therefore, health status changes occur as independent events.

Second, the onset of any acute condition has a fixed initial duration of three months, during which time the afflicted individual experiences a risk of dying from that condition. Persons who experience an acute onset are allowed to "recontract" that condition prior to the termination of the three-month onset period. The effect of recontracting the disease is to extend the recovery date of the condition by another three months from the month of recontract; during this second onset interval, the condition-specific mortality risk remains at the same level at which it was during the initial onset period.

Third, although acute illnesses constitute transient health statuses, the onset of a chronic condition results in a permanent independent increase in the risk of dying, although the amount of the increase is allowed to vary as the person ages forward from the time of onset. In other words, persons who experience a chronic onset never recover, in the sense that they never experience a remission of the rise in mortality risk that results from the onset of a chronic condition at any time during their lifetimes.

Finally, the experience of an "external incident" has no effect on the individual except to bring about an instantaneous, momentary rise in one's risk of dying. That is to say, that the model explicitly concerns itself only with those external events that are immediately lethal.

Mortality, except from an external cause, is handled by the model as an age and health status contingent process. Individuals die according to a set of independent probabilities of death from each of the health conditions that they are experiencing at a given moment. Persons will experience a continuous rise in their death risk as they accumulate more and more conditions.

Perhaps the most problematic aspect of the above formulation is that it ignores considerations of disease latency and recovery. We would argue, however, that the model's lack of any explicit representation of latency or recovery is less relevant to the task at hand--namely, that of assuring consistency between the model's health status projections and its projections of overall mortality--if one uses a broad definition of "recovery" or latency that includes any elements of the disease process that do not directly influence mortality risk. If the remission of a condition does not carry with it a reduction in the risk of dying, the individual cannot be considered as having "recovered" from the standpoint of the model. The mere resumption of normal activity--if, for example, this is what one means by "recovery"--has no relevance in the present context.

Thus the model embodies an "ever experienced" notion of morbidity, unlike the "currently manifest" notion implicit in most point-prevalence measures of health status. In a sense, it accordingly captures the conceptualization of disease latency that characterizes the standard multiple decrement, cause-elimination life table, in which the survivorship column (l_x) of the life table is segregated into subpopulations of individuals who are ultimately and inevitably destined to die from a specific cause--"marked for life", so to speak, by the onset of a specific condition.

THE MICROSIMULATION FRAMEWORK OF THE MODEL

In its present application, this formulation is operationalized by drawing upon its implications for the timing of changes in health status and death. Simply stated, each individual's life span is segmented into a series of age intervals over each of which the individual is assumed to be at constant risk of experiencing a given event. A stochastic procedure then is applied to determine whether the individual survives through each consecutive interval without experiencing the event. When the interval is finally reached in which the event is projected to occur, the model assigns that event to a precise time point within the interval.

Derivation of the appropriate timing functions is relatively straightforward. Define,

- t' as the time at which an event is projected to occur
- $t(i)$ as the number of years in the i-th age interval

$p(i)$ as the probability that an event will not occur in the i-th age interval

r as a number that is randomly selected from a rectangular distribution of numbers between 0 and 1

If one assumes that the hazard of an event remains constant over all age categories, it can be shown that,

$$(a) \qquad p(i)^{t'} = r$$

Appropriately transposing t' in (a) yields the expression,

$$(b) \qquad t' = \frac{\ln r}{\ln p(i)}$$

Expression (b) specifies the time at which the event is expected (projected) to occur, if the hazard of its occurrence remains constant over and across all age intervals.

To derive an estimated failure time from hazards that vary across age intervals, consider first the case in which

$$t' > t(1) \text{ and } t' < [t(1) + t(2)]$$

that is, the case in which an event is projected to occur within the second age interval. It can be demonstrated that expression (a) and the randomness of r together imply that,

$$(c) \qquad r = P(1)^{t(1)} \cdot p_{(2)}^{[t'-t(1)]}$$

Appropriately transposing (c) yields the expression,

$$(d) \qquad t' = t(1) + \frac{\ln r}{\ln p(2)} - \frac{\ln \left[p(1)^{t(1)}\right]}{\ln p(2)}$$

which is the precise time within the second age interval at which the event is projected to occur. This result can be further expanded to yield the general timing expression,

$$(e) \quad t' = \frac{\ln r}{\ln p(j)} + \sum_{i=1}^{j} \left[t(i-1) - \frac{\ln p(i-1)^{t(i-1)}}{\ln p(i)} \right]$$

where,

$$t' < \sum_{i=1}^{j} t(i) \text{ and } t' > \sum_{i=1}^{j-1} t(i)$$

and j denotes the age interval within which the event is projected to occur.

Expression (e) is thus used in the model to determine the timing of disease onsets and deaths. During risk intervals having constant occurrence hazards, it in effect assumes a Poisson process for the event.

Because the model assumes the onset of any given condition to occur independently of the presence or onset of all and any other conditions, one can use (e) to separately project the expected onset times of each of the eleven conditions. Later, when age at death is determined, onsets that are projected to occur after death are "erased". With regard to chronic onsets, recall that the model postulates non-recovery. Hence, only one random number needs to be generated to determine for the entire lifespan the projected period during which the individual will have a given condition not already being experienced at the start of the projection. The projection of acute onsets varies in detail, though not in principle, from that of chronic onsets.

Once the health history is available, it is then possible to determine the precise age at which the individual will die, using a single cast against the general timing function (e). The relative event probability, $p(i)$, is the joint probability of surviving all of the conditions extant in the i-th risk-homogeneous interval. Since disease onsets can occur at precise time points within age intervals over which the hazard of disease onset is constant, the mortality risk homogeneous intervals will be bounded not only by the age boundaries across which changes in the condition-specific probabilities of dying occur, but also by the time points when the individual experiences the onset of additional conditions.

APPLICATIONS

Two main applications of the model have been made to date, each intended for specific projection purposes that made the final output somewhat different, but each involving the general strategy as previously outlined. The first entailed a national projection of the elderly population (65 years of age and older) to the year 2000 by age, sex, race and marital status categories and twelve health condition states, as noted previously. In addition, of course, deaths could arise from accidents in the model as well as deaths from any of the eleven disease conditions. The elderly population segment is therefore decremented by deaths arising from the sickness to death process, and it is incremented by persons turning 65 years of age.

The second application involved projection to 2000 of a national population; in this case the male veterans entitled

to benefits from the Veterans Administration. These projections were specific by age and race. These projections took into account that the veterans population is continuously being incremented by discharges from the military by means of estimates provided by the Department of Defense.

Parameter Estimation

The most serious obstacle to the success of any model is the degree to which adequate data can be derived to apply and test it on actual populations. The numerous decisions regarding specification of the parameters made in these two applications relate fundamentally to data considerations, but a full discussion of these matters is clearly not appropriate here.

In brief, disease incidence and prevalence rates for the projections were estimated using data from the 1970 Health Interview Survey, using procedures similar to those employed by the National Center for Health Statistics in preparing its national morbidity estimates. Estimates of disease prevalence were obtained by pooling the point-prevalence data of the HIS over the entire data year and then averaging. A gross adjustment for prevalence underreporting was attempted by adding the total number of deaths from a given cause to the estimated prevalence of that condition; in effect, the prevalence estimates assume that all of the deaths from a given cause occurred among individuals that were not covered by the Health Interview Survey. Data on deaths by underlying cause were tabulated from NCHS 1969 complete file of United States death certificates, adjusted to reflect 1970 levels of total death rates.

For the elderly projection a sample of 49,000 individuals, age 34 years and over in 1970, comprised the "start population". The statistical theory upon which the model depends requires that the projected sample be genuinely random in nature. As the 49,000 HIS cases that comprised our projected sample were, in fact, differentially weighted, an adjustment of the file was required before the actual simulation was carried out. Consequently, each case in the sample was duplicated by a factor equal to its case weight divided by the lowest case weight found in the sample. The result of this adjustment procedure was to expand the original sample to approximately 110,000 projected cases.

For the projection of the veteran population, the sample that was actually used for the microsimulation consisted of the 16,000 United States male veterans surveyed in the 1970 Health Interview Survey. The result of the sample replication procedure was to expand the veteran sample to approximately 31,000 projected cases.

Experimental Error

The stochastic nature of the event-timing functions implies that the projection contains an element of random variance, "experimental error". A simulation will lead to a "correct" projection only if the simulation of each sampled life history is carried out "sufficiently often", as on the familiar coin toss experiment. One way to assure that the projection is relatively free from experimental error is to repeat the simulation of all sampled cases repeatedly until the relative frequencies of each life history characteristic, specific by whatever demographic categories are of interest, cease to change. In our applications of the model, there were good reasons to suspect that such stability had been attained after only a single simulation was run, thanks to our large sample size. Indeed, stability could have been achieved with a much smaller sample than the one that was used. Stability was tested in the following manner.

After the simulation was carried out, the resultant sample of projected life histories was randomly divided in half. The hypothesis was then tested that the relative distributions of selected characteristics in the half samples could be reflective of two different sampling universes. As the primary interest in devising the model was to prepare joint projections of population size and health status in which survivorship patterns were consistent with patterns of disease prevalence, we selected for the test the distribution of projected deaths jointly tabulated on age of the decedent in 1970, age at death, and conditions present at death, with age specified in terms of 5-year age intervals. For each condition, the age-specific death rates in one-half of the file were regressed against those in the other half. The results for the veterans are shown in Table 1.

Projection Results

Illustrative results for the elderly white male population are presented in Tables 2 and 3. These tables reveal that the results for this exercise are generally in accord with our expectations -- that is the model behaves correctly. In the actual results obtained in this test, the total numbers of the elderly exceed those estimated in other national projections. At the same time, the preva-

lence of chronic disease (in both numbers and rates) is sharply increased.

These results indicate that the structure of the model is basically sound and does capture the interaction of disease prevalence and age structure. A major deficiency would seem to lie in the assumption regarding non-recovery since a disease is acquired that creates overestimates of prevalence as the projections proceeds. What is clearly called for are refinements in the estimated levels of prevalence in the start population and the incidence rates of disease conditions that drive the model. These considerations have received additional treatment in further elaborations of the model.

The veterans application offered a possibility of assessing the population results against alternative procedures. Table 4 provides the results of this exercise. The alternatives are Method I which involved projecting the 1970 veteran population forward on a cohort-component basis, with future discharges from the military taking the place of births and using survivorship ratios derived from the life table for all United States males including non-veterans.

A second approach -- Method II -- was to prepare a cohort-component projection similar to that of Method I, using veteran-specific survivorship ratios. Although it is true that the mortality data that would enable one to directly estimate a veteran-specific life table are not readily available, it is at least conceivable that such estimates might be obtained on an indirect basis, as follows.

For any cohort of veterans alive in a given year, the number of veterans alive after a certain period of time has passed is equal to the number alive initially plus the number of military discharges entering the cohort in the interval less the number of deaths occurring among both the initial cohort and those who were discharged from the military into the cohort during the interval. Suppose that the veteran population is closed to migration. Because such surveys as the Current Population Survey provide data on the size of veteran cohorts at successive intervals in time, and also is available, it was possible to estimate the rate of decrement over given time intervals.

Table 4 indicates that all three projections show the same general trend--a veteran population of 27.2 million in 1970 that increases to somewhat more than 28.5 million in the late 1970's, but then begins an uninterrupted decline through the year 2000. The main reason for the decline is that the large World War II veteran cohort will be entering old age, and the projected number of new entrants into the veteran population is simply not sufficient to offset the increased number of veteran deaths that are expected to occur as a result.

The populations projected under Methods I and II both peak in 1980, while the peak for the microsimulation is reached in 1975. Under Method II, with its veteran-specific survivorship, however, the population reaches both a higher peak size--28.8 million--and declines to a lower level in the year 2000--25.4 million, than occurs in the Method I projection with its total male survivorship schedule (somewhat less than 28.7 million, and 25.5 million, respectively). Method III, on the other hand, generates lower projected numbers at all times than are projected by the other two methods. Method I generates the highest proportion of projected elderly--27.9 per cent of the veteran population, as compared to the 26.4 per cent projected under Method II and the 23.2 per cent generated by the microsimulation, for the year 2000.

The survivorship dynamics hold the key for explaining these differences. Methods I and II assume that constant levels of survivorship prevail throughout the period 1970 through 2000. Method I subjects veterans to the schedule of survivorship rates that was experienced by all United States males, including non-veterans, in 1974. Method II, on the other hand, uses a schedule of survivorship experience in which veterans experience lower levels of mortality during the younger ages, but in which the favorable differentials fades with age. Indeed, examination of the Method II life expectancies reveal a slight mortality crossover in the older age categories; mortality rates derived from the 1970-73 CPS for ages past 55 years (not shown here) are actually slightly higher than those observed for all males in 1974. As a result, Method II population grows somewhat faster than the Method I population, as long as it is a relatively younger population; when the projected age distribution under Method II becomes a relatively older one than that of Method I, the Method II population declines much more rapidly because of its lower survivorship levels at the older ages.

In the microsimulation projection, however, survivorship levels vary over time as the relative prevalence of disease varies. An examination of the Method III life expectancies, derived from the age-specific death rates generated by the projection, reveal that the

implied survivorship function for veterans, on the assumption that they experience fixed rates of disease onsets and condition-specific mortality characteristic of the total United States male population, endows them with substantially lower mortality at the younger ages (initially) than that which characterizes either Method I or Method II, but that the rates rise more rapidly with age and lead to much higher levels of mortality at the older ages than is characteristic of either of the other two survivorship regimes. The result is a considerably smaller population projected for the year 2000, with a smaller proportion of aged individuals.

Thus, the Method II projection presents the population size implications of a regime of veteran mortality in which veterans are somewhat favored over non-veterans initially, with gradual convergence to the mortality experience for all males (assuming that age is a relatively good proxy for time since discharge), and with a slight mortality cross-over at the older ages. Method III shows the implications of a regime with initially much lower veteran mortality, rapid convergence and a much deeper cross-over.

(1) Horvitz, D. G., F. G. Giesbrecht, B. V. Shah, and P. A. Lachenbruch "POPSIM, a Demographic Microsimulation Model". Monograph 12. Chapel Hill, North Carolina: Carolina Population Center, University of North Carolina at Chapel Hill. 1971.

Table 1

Condition	Regression Slope	Correlation Coefficient
Acute infectious disease.....	1.077	0.992
Acute respiratory disease.....	0.811	0.986
Miscellaneous acute diseases.....	1.017	0.998
Chronic respiratory disease.....	0.983	1.000
Malignant neoplasms.....	0.998	0.998
Endocrine and metabolic disorders.	0.990	0.998
Cardiovascular disease.....	1.006	1.000
Cerebrovascular disease.....	0.922	0.999
Arterioschlerosis.....	1.043	0.999
Chronic digestive, liver disease..	0.970	0.995
Miscellaneous chronic disease.....	1.002	1.000
External events.....	0.943	0.978

Table 2

Projected Population of the Elderly, 1980-2000

White Males Age 65 Years and Over

Population By Age:	1980	1990	2000
All Ages, 65+ Years	10931462	14325538	15303999
65-69 Years of Age...	3808479	4388216	3861230
70-74 Years.....	2802199	3623962	3650442
75-79 Years.....	2028967	2816234	3263160
80-84 Years.....	1230640	1835174	2320493
85-89 Years.....	770708	1112436	1444209
90-94 Years.....	241681	417990	605811
95 Years and Over....	48788	131526	158654
Percentage Age Distribution:	**1980**	**1990**	**2000**
All Ages, 65+ Years	100.0	100.0	100.0
65-69 Years of Age...	34.8	30.6	25.2
70-74 Years.....	25.6	25.3	23.9
75-79 Years.....	18.6	19.7	21.3
80-84 Years.....	11.3	12.8	15.2
85-89 Years.....	7.1	7.8	9.4
90-94 Years.....	2.2	2.9	4.0
95 Years and Over....	0.4	0.9	1.0
75 Years and Over....	39.5	44.1	50.9
85 Years and Over....	9.7	11.6	14.4
Selected Characteristics:	**1980**	**1990**	**2000**
Mean Age (Years)	74.25	75.06	76.17
Percent of all Persons Age 65	38.98	39.24	38.86
Est. Annual Growth Rate (Percent)	3.70	1.76	0.18
Total Percentage Change Since 1970	47.61	93.45	106.66
Annual Deaths Per 1000 Persons	41.28	46.50	51.51

Table 3

Projected Total Prevalence
Of Selected Conditions Among The Elderly, 1980-2000

White Males Age 65 Years and Over

Persons With Selected Conditions:	1980	1990	2000
Total	10931462	14325538	15303999
No ill health	1997641	1383519	881611
Acute infectious disease	250296	338111	317779
Acute respiratory disease	682174	865988	868661
All other acute conditions	732784	913415	914767
Chronic respiratory disease	2654727	4541182	5827826
Malignant neoplasms	918501	1669400	2022090
Chronic endocrine and metabolic diseases	809535	1463881	1993063
Chronic cardiovascular disease	3125858	5595072	7010214
Chronic cerebrovascular disease	892016	1564239	1924073
Arterioschlerosis	349738	687425	836493
Chronic digestive and liver disease	764422	1513946	1985268
All other chronic conditions	4914559	7818069	9692446
Crude Total Prevalence Rates:	**1980**	**1990**	**2000**
No ill health	182.74	96.58	57.61
Acute infectious disease	22.90	23.60	20.76
Acute respiratory disease	62.40	60.45	56.76
All other acute conditions	67.03	63.76	59.77
Chronic respiratory disease	242.85	317.00	380.80
Malignant neoplasms	84.02	116.53	132.13
Chronic endocrine and metabolic diseases	74.06	102.19	130.23
Chronic cardiovascular disease	285.95	390.57	458.06
Chronic cerebrovascular disease	81.60	109.19	125.72
Arterioschlerosis	31.99	47.99	54.66
Chronic digestive and liver disease	69.93	105.68	129.72
All other chronic conditions	449.58	545.74	633.33
Mean Number of Conditions Per Person:			
All persons, age 65+ years	1.47	1.88	2.18
Persons, age 65+ years, w. 1+ conditions	1.80	2.08	2.32

*Crude total prevalence rates are expressed in terms of prevalent conditions per 1000 persons, age 65 years and over, of specified age, race and sex.

Table 4

Alternative Projections
Of The U.S. Male Veteran Population:
Selected Characteristics, 1970-2000

Population Size (1000's):

Year	Method I	Method II	Method III
1970	27,203	27,203	27,203
1975	28,614	28,647	28,390
1980	28,670	28,801	27,931
1985	28,398	28,595	27,061
1990	27,766	27,866	25,757
1995	26,772	26,711	24,224
2000	25,485	25,353	22,559

Method I: Cohort-component projection, assumes observed 1974 survivorship for total U.S. male population.

Method II: Cohort-component projection, assumes CPS estimate of veteran-specific survivorship for 1970-1973.

Method III: Sickness-death microsimulation, assumes HIS estimates of disease onset and virulence rates for total U.S. male population, 1970.