Analytical Studies Branch

Research Paper Series

Statistics Canada Statistique Canada

Canada

POHEM -- A FRAMEWORK FOR UNDERSTANDING AND MODELING THE HEALTH OF HUMAN POPULATIONS

by

Michael C. Wolfson

No. 44

Social and Economic Studies Division
Analytical Studies Branch
Statistics Canada
1992

The analysis presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada.

Aussi disponible en français

POHEM -- A Framework for Understanding and Modeling the Health of Human Populations

Michael C. Wolfson[1]
Analytical Studies, Statistics Canada and
Canadian Institute for Advanced Research

a paper prepared for the
Annual Research Conference
of the U.S. Census Bureau
Washington D.C., March 24, 1992

Abstract

This paper describes a new approach to the development of statistical information with respect to health. We start with the diagnosis of two fundamental problems in the current set of Canadian health-related statistics. The first is the serious imbalance in coverage, with far more data collected on the inputs to the health care system, and much less on the health status of the Canadian population. The second problem is the lack of coherence. Unlike other major statistical systems such as the System of National Accounts and population demography, health data do not "add up". They appear in compendia primarily as a matter of juxtaposition. These problems are being addressed by the development of a new conceptual framework for a System of Health Statistics. The Population Health Model (POHEM) is a central component of this ongoing work at Statistics Canada. This paper gives a general overview of the raison d'etre of POHEM, its construction, and recent illustrative results, particularly insofar as they allow problems of non-sampling error to be addressed.

Key words: health statistics, conceptual framework, microsimulation, non-sampling error.

---

1 The views expressed are my own, and not necessarily those of Statistics Canada or the CIAR. The development of POHEM has been a team effort, and I am deeply indebted to my colleagues for their many contributions, and to the CIAR for providing a unique intellectual milieu. I of course remain responsible for any errors or infelicities.

A. Introduction

This paper is about the development of a new approach to health information. In part, this development was inspired by the dinner speech given five years ago at the Annual Research Conference of the U.S. Census Bureau by the former Chief Statistician of Canada, Dr. Martin B. Wilk. In his remarks, he raised a fundamental question about the statistical community's preoccupation with mathematical formalism and sampling error. He drew attention to all the myriad other sources of error in statistical work. One general suggestion was offered to ameliorate this professional bias toward the more readily quantifiable sampling sources of statistical error. This was to address the problem of non-sampling error by the construction of statistical data *systems*.

Statistical systems reach in the direction of science more than mathematics for their inspiration and structure. They draw on the empirically grounded style of substantive scientific thinking, and the interplay between theorizing and observation. This scientific -- as opposed to a more abstract or axiomatic -- style of thought can provide a basis for coherent conceptual structures which in turn can influence what and how data are collected. The resulting image is a properly designed set of data collection processes where the flow of data is integrated to form a coherent structure of information. Then, through built in redundancy and "data confrontation", problems of non-sampling error would be continually visible. In turn, given this visibility, these errors could be monitored, and changes in data collection and processing could be made to reduce them. A major example of such data confrontation occurs with the income and expenditure sides of the National Accounts, and the balancing process used in constructing the input/output tables.

Shortly after the Wilk (1987) speech, and not unrelatedly, Statistics Canada began a research project in precisely this direction. It started out as a request for a "think piece" on a satellite account for health statistics. The motivation for this work was a wide ranging sense of unease with Canada's health information. The initial think piece quickly became an exercise in developing the conceptual basis for a System of Health Statistics (Wolfson, 1991). The project has since evolved into the construction of a prototype, the core of which is the Population Health Model, POHEM. These efforts within Statistics Canada have also become integrated with a larger review of health information in Canada (Wilk, 1991). As a result, momentum has been building for a basic rethinking of the health information system. This includes the creation of

joint institutional mechanisms involving the major constituencies in the health field, so that new efforts will be coordinated and based on consensus. The ideas and conceptual framework of POHEM and related work are at the core of these broader developments.

This paper gives a general overview of the raison d'etre of POHEM, its construction, and recent illustrative results. We start with an argument as to why a coherent conceptual framework is needed in the health area, and what some of the desiderata are for such a framework. The paper next considers what "theory of health" should underlie the statistical framework. We then turn to a description of microsimulation methods generally, and POHEM specifically as the core of a coherent system of health statistics. The paper concludes with examples of data integration and confrontation enabled and produced by POHEM, particularly as they illustrate the application of systems of statistics to the question of non-sampling error.

B. The Need for a Conceptual Framework

Health is a major area of public interest and public policy concern. It must certainly rival the economy in importance, yet in comparison the statistical base is confused, fragmentary, and incoherent. This is not to say that we know any more about the determinants of healthy economic growth than we do about the determinants of a healthy populace;[2] but at least we have a coherent statistical framework for economic information in the System of National Accounts.

A convenient indication of the incoherence of health statistics can be gained by perusing the statistical compendium, Health United States (NCHS, various years). Both Health USA and a National Accounts publication contain page after page of numerical tables; and both have table headings that have an apparent relationship to the book titles. However, the SNA numbers obey a series of arithmetic identities. Furthermore, economic series like unemployment rates and interest rates, while not connected arithmetically to the SNA series, are related in various macroeconometric models developed and housed in nearby organizations. The same kind of mathematical structure certainly does not exist for the data in Health USA. The only thing that

---

2 Indeed our scientific knowledge in areas like biology and biochemistry is relatively strong, certainly as compared to economics. Still, these "white lab coat" areas of science constitute only a part of the knowledge required to understand population health -- the levels of health of large populations of individuals, not specific individuals presenting with a known malady. It is this much broader concern with health at the level of populations that is central to this paper.

binds the numbers together is a general pertinence to health, and the binding of the book itself. (Health statistics, unlike the SNA, also use a wide variety of numeraires -- from person-years to bed-days to dollars, but this need not be problematic, as will be discussed below.)

A less abstract and equally fundamental problem with health statistics is their imbalance. Much more effort is spent measuring the inputs to the health care system than in measuring how healthy the population is. There can be no clearer evidence for this proposition than the fact that many people know how much health care costs as a percentage of GDP, but no one knows if the "average healthiness" of the population is increasing or decreasing from one year to the next. Of course, almost as many people who know the cost numbers also know life expectancies. But life expectancy is a form of average age at death; it need not be correlated with health status while living.

The current incoherence and imbalance in our countries' collections of health statistics is more than sufficient reason to develop a conceptual framework, but how to proceed in this effort? One step is to articulate some of the other desiderata for such a framework. In addition to coherence and balance, let us emphasize three main features. The first is the centrality of individuals. Health is basically about people, so thinking about statistical systems in the health area should start with measures of the health of the members of a population. A key corollary point is that health is a dynamic process. Expression of genetic predispositions, diseases, and risk factor exposures, for example, evolve through time and the underlying bio-medical phenomena often exhibit very long latencies. Thus, individual population members should be considered in a life-cycle context.

A second area concerns units of measure. Every effort should be made to measure health and health-related activities and phenomena in their own most natural units. Certainly, the principal numeraire should not be dollars; rather, it should be something like "health status-adjusted" or "full health equivalent" life-years. Such measures of health status are fundamental to remedying the current imbalance in the health area where costs figure so prominently. Moreover, common use of such measures is essential if, for example, clinical trials results are to be comparable to broadly measured population baseline survey results.

At the same time as there is a core common numeraire like health status-adjusted life years for measuring individuals' health status, it is also important to allow some variety in units of measurement. For example, part of the statistical system can use the visit or service encounter as the basic unit of observation for interactions with the health care system. Of course, dollars should also be used, but only where it is most appropriate, as in descriptions of the resources used in the health care sector. Indeed, these amounts should be grouped according to function and other classifications used in the National Accounts in order to provide a linkage between the two systems of statistics.

A third desideratum of a conceptual framework for health statistics is that it be seen as the foundation for a decision support system. One reason is simply that unless this is the case, experience shows that long term stable funding for data collection, integration, and analysis will not be forthcoming (Pommier, 1981). More importantly, systematic feedback to health care providers and to the health system more generally is absolutely essential for continuing improvements in health care, and in population health more generally.

There are important corollary implications of this decision support role. Feedback must be based on high quality and appropriate information -- i.e. information that can support decisions at various levels ranging from individual surgeons and care givers to ministers of health. Often in the health area, there is a pressing need to make choices between very strongly contending claims on resources. Thus, an ability to support cost-benefit analysis -- using the term in the broadest sense -- is vital. At the same time, the statistical system must support distributional analysis, to indicate who might gain or lose from any policy initiative.

Finally, decision support entails the capacity to pose and answer "what if" questions. The construction of hypothetical scenarios is central, and this requires simulation of one sort or another. While producing numbers is widely accepted as part of the core mission of national statistical agencies, the construction and use of simulation tools is not. For example, statistical agencies generally do not get into the business of macroeconometric modeling, though the data producers typically have very close contacts with the modelers. On the other hand, long run demographic projections are usually developed by national statistical agencies. The health domain, in this respect, is virgin territory. It may be that the longer term goal is to have external organizations like ministries of health or state and provincial governments become the principal users of "what if" analytical tools in the health area. In the meantime, however, it is entirely reasonable for a

national statistical agency to take the initiative, to anticipate this most fundamental use of their data, and to develop new data systems in conjunction with the decision support analytical tools, including simulation models.

In addition to these general desiderata, there are several more practical premises for any statistical system to be associated with a new conceptual framework. The first is the need to build on explicit microdata foundations. This means, for example, that data on groups of individuals are derived by aggregating data on the individuals in each group (or a sample thereof). This approach may appear self-evident, but it is not the case with the National Accounts and demographic projections. In each of these existing statistical systems, the smallest level of disaggregation is a group like an industrial sector or population in a given age interval. The natural units of observation, firms or individuals respectively, are not accessible.

The semi-aggregate character of these economic and demographic statistical systems is understandable, given that they pre-date the revolution in computing. However, from the perspective of contemporary computing power and the analytical benefits of explicit microdata foundations, it is an unfortunate limitation that should be avoided in the development of new statistical systems. The System of Health Statistics should provide the capacity to disaggregate down to the level of natural units of observation such as individuals, health care providers, and health care service encounters.

An obvious corollary and another practical premise is that new systems of statistics should *not* be seen principally as a series of print publications. Rather, the fullest incarnation of the statistical system should be in electronic form -- a database combined with tailored retrieval and analytical software (subject always to constraints imposed by considerations of privacy and confidentiality). The image should be one of offering the user an interface whereby s/he can ask questions and get answers. Of course, many users will still want printed results, but these should be derivative.

Finally, it is essential, when devising a new statistical system, to avoid unnecessary constraints. One of these is the existing range of data feeder systems; they are themselves part of the problem. Thus, the multitude of incompatible hospital and administrative databases, and limitations of current household surveys, have generally been ignored, at least in our conceptual work. If we are to plan for new kinds of health statistics, then it is appropriate to accept as constraints that

the proposed data collection processes be feasible and not too costly. However, the planning would be crippled from the outset if we had to assume that the current babel of concepts, definitions and categories was immutable.

C. A "Theory of Health"

Ordinarily, the development of a conceptual framework and system of statistics presupposes some sort of theory. In the case of the System of National Accounts, the underlying theory was Keynesian economics. For demographic projections, at one level the theory is quite simple -- population next year equals population this year plus births and immigrants minus deaths and emigrants. At a deeper level, considerable work is devoted to trying to understand the determinants of these basic population flow rates.

Do we have a corresponding theory of health? The fairest answer is no, there is no widely agreed *overall* theory – a theory that operates at the level of populations of individuals, as well as organs, cells and biochemicals. On the other hand, many partial theories exist; and there is general agreement about the main ingredients in these various theories of the determinants or causal pathways involved in health processes. Where there is contention is in the significance, materiality, and sometimes the direction of influence, and in the ways the various partial theories should be combined. To paraphrase an old song, there is general agreement that the risk factor is connected to the disease, and the disease is connected to the symptom, and so on. But there is controversy about the details of various causal pathways.

For example, there is general agreement that poor diet is associated with higher incidence of coronary heart disease. However, there is substantial debate as to whether the strategic point for intervention is lowering the levels of certain serum cholesterol fractions, and if so whether drugs are the most effective method. To take another example, there is overwhelming evidence that wealthier people are healthier and live longer; but there is continuing debate as to which way the major forces of causality run -- from poor health to low wealth or vice versa. (Wolfson et al., 1992, provide evidence for the latter.)

Given such contending partial theories, it is possible to set out the basic *structure* for an overall theory. There are two basic parts to this structure, a description of the variables of interest and a description of how they evolve over time. In descriptions of dynamical systems like the motions

of astronomical bodies, these are referred to respectively as the state space and the laws of motion. Figure 1a provides an image that can serve as the basis for describing a comprehensive state space for population health. This is a monochrome reduction of the animated graphics software image in the Health Information Template (Wolfson, 1992b) which was developed as part of the National Task Force on Health Information (Wilk, 1991).[3]

[Figure 1a about here]

The basic image of the Template shown in Figure 1a divides the health field into three broad domains. At the center of the image (deliberately so) is a population of individuals and their characteristics. The visual nuance of a set of files is also deliberate -- it indicates that the data in the state space for the population must be at the individual level.

Surrounding but not completely enveloping the population in Figure 1a (visually and metaphorically) is the "External Milieu". It in turn has been structured into four kinds of environments. The popular media tend to identify the "environment" only with the first of these, physico-chemical environments. However, there is strong evidence that socio-cultural and economic environments are at least as important to human health as risks derived from physico-chemical exposures. Fourth, some aspects of the health system are best considered as an environment (e.g. the existing stock of hospital buildings and medical equipment).

Finally, as the product of our conscious intended actions, the Template defines a third domain of "health-affecting interventions", a phrase deliberately chosen to be broader than "health care". These interventions are of two general kinds. The first is "individual" health-affecting interventions which act on us as individuals one-on-one, and take the form of encounters with providers

---

3 The Health Information Template is an experiment in animated computer graphics pedagogy. It was developed in 1991 and used as part of the consultative process of the National Task Force on Health Information, in part to show the breadth of information potentially pertinent to health, and to introduce a classification structure or "roadmap" for the vast domain of health information. It built on the work underlying the System of Health Statistics (SHS, Wolfson, 1991) and the work underway on POHEM. The current form of the Template was heavily influenced by the project team who helped create it, and the cross-country consultations with various health constituencies. In turn, these ideas have broadened our thinking in respect of the SHS and POHEM.

of various services. The second is "collective" health-affecting interventions. These take the forms of government programs and regulations that act on us collectively, though indirectly, through the external milieu.[4]

The Template software from which the image in Figure 1a is taken also allows the various domains and sub-domains to be "exploded" to show finer levels of underlying classification structure. This is illustrated for the domain of Individual Characteristics in Figure 1b, using a conventional hierarchical classification structure, and building on the WHO's ICIDH (1980).

[Figure 1b about here]

An alternative structure for the variables in each individual's state space emphasizes the time dimension. This is shown in Figure 2 with an adjusted version of the basic Template image. The central portion of the image now shows a biographical or life cycle structure for the information on individuals, as well as some shading to illustrate an actual hypothetical biography of events and states.[5] This state space for individuals is not unlike a sketch of a longitudinal microdata set. Each row corresponds to a group of variables or fields in the record layout. The main groups of variables are as follows:

[Figure 2 about here]

- SES (socio-economic status) -- years of schooling and educational attainment, labor force participation, earnings, other income, wealth, tenure, career stage, marital status, fertility, etc;

- risks -- various risk factors including genetic predispositions, physical factors (e.g. cholesterol, obesity, hypertension), lifestyle factors (e.g. smoking, substance abuse), and risks

---

4 To extend the analogy with dynamical systems to the ideas of control theory, health affecting interventions constitute the "control variables" for influencing the trajectories of individuals directly, and indirectly via the trajectories of the environments of the external milieu.

5 The Template software displays an animated progression of events culminating in this image in order to illustrate the dynamics of an individual's life cycle. Note that in Figures 1a and b, the time dimension was missing; the images represent data at a point in time. While time is explicit for the individuals in Figure 2, it is still implicit for the external milieu and for health-affecting interventions.

deriving from environmental exposures (including physico-chemical sources of noxious agents and socio-cultural environmental factors affecting the ability to cope with functional limitations or disabilities);

- diseases -- clinically defined symptoms and diseases, e.g. coronary heart disease, cancer, musculoskeletal diseases, dementias;

- functional status -- activities of daily living (ADL), instrumental activities of daily living (IADL), multidimensional functional status indicators such as the International Classification of Impairments, Disabilities and Handicaps (World Health Organization, 1980), multi-attribute health status scales (e.g., Ontario, 1990), etc.;

- costs -- health service utilization (e.g. physician visits, home health services, hospital visits, surgery, drugs, nursing home use, etc. measured in a variety of natural units, e.g. procedures by type, bed-days), a summary dollar amount for health services consumed, plus other economic costs like foregone earnings and the value of informal social support; and

- health -- a summary health status value score (i.e. a number between zero and one).

The dynamics or laws of motion for these variables in an individual's biography are not explicit in this image. All that is shown is the top levels of a classification structure. However, descriptions of the dynamics of various health-related processes are an essential part of the "theory" and the health information that should be encompassed by this framework.

Before indicating how these dynamic relationships can be incorporated, a very brief digression on theories and models is necessary -- particularly to clarify terms. Theories as well as dynamic processes are often represented by models. Unfortunately, the term "model" is so widely used that it is quite ambiguous. For example, "model" is used to describe an abstract conceptual framework (like the Template image in Figure 1a), a deterministic system of differential equations (e.g. to describe the mechanical forces acting on a bicycle in motion), a specific mathematical relationship represented by a single equation to be estimated or tested statistically (e.g an econometric or hazard regression), an algorithm describing a particular process or "toy world" (e.g. the tesselation automaton of A.K.Dewdney's ecology game WA-TOR, 1984), and a complete numerical simulation system containing many specific process algorithms (e.g. a complete macroeconometric model). We shall generally use the term model in the latter sense.

The term theory is often used interchangeably with model. For our purposes, a theory is a structured set of hypotheses about how something works. A key criterion of the adequacy of a theory is its predictive ability. There is always an interplay between observation and theorizing. We can break into this cycle arbitrarily by starting with a classification structure, like that shown by the Template for the health field. This classification embodies the results of prior theorizing and observation. It is used to structure subsequent observation, including the generation of data on various health processes.

Data, when analyzed, often give rise to empirical regularities -- the data have structure, and empirical analysis reveals it. Typically, empirical analysis is parametric; an a priori functional form is assumed, and the best fitting coefficients are estimated. Age/parity-specific fertility rates are a very simple example; a logistic risk function for coronary heart disease based on the Framingham data is a richer example. More rarely, less parametric methods are used as in the Dowd et. al., (1992) analysis of the Framingham data using highly multivariate, fuzzy set, grade of membership (GoM) techniques. Often, as in the examples just noted, these empirical analyses generate precise quantitative descriptions of the dynamics for a given health process. These kinds of dynamic processes -- exposures to risk factors, occurrences of events (heart attacks and births), and the formulae connecting the two -- are clearly central to any sensible set of health statistics.

These dynamic processes are implicit in Figure 2 as follows. At any moment in calendar time, living individuals have completed only portions of their life courses. As a result, the cubbyholes in the biography or microdata record layout are only filled in up to that point in time (where time is moving to the right along the horizontal axis). For example, Figure 2 shows the completed life course of an individual, where the gray shade levels indicate (illustratively) the levels of scalar values for a set of individual attributes. The dynamics representing various theories or knowledge relating to health then consist of those rules, mappings or formulae that allow the next column of an individual's biographical array (assuming discrete time) to be "filled in" (i.e. estimated, calculated or imputed) as a function of all the information to the left.

These dynamical algorithms apply to a heterogeneous population, so they will likely be complex and multivariate, and may have substantial stochastic components. The *theories* comprehended by the conceptual framework are embodied in these explicit descriptions of the dynamics of the various individual attributes. For example, the same fertility theory that underlies demographic

projections is naturally incorporated by having the event of giving birth represented by one of the rows in each individual's biographical record layout, and taking as the law of motion or dynamics for this type of event a stochastic process described by a set of age/parity-specific fertility rates.

There are contending theories of the determinants of fertility. Other variables like marital status, ethnicity and labor market behavior may be significant determinants. The framework in Figure 2 is agnostic in this regard. As long as these kinds of independent or determining variables are part of the state space, they can be incorporated as inputs to an algorithm describing the dynamic process. The same point applies to health processes. There is general agreement that hypertension and elevated cholesterol increase the risk of coronary heart disease; there is continuing debate in the epidemiological literature about the magnitudes of these effects, and the range of other variables that are also significant determinants of coronary events. The framework in Figure 2 in principle can absorb or include any such theory or observed empirical regularity, and can in fact encompass alternative or contending versions. The only proviso is that they are well-defined, i.e. they can be expressed algorithmically.

We can, in fact, think of a library of dynamic process algorithms, one or more for each state variable. A given instance of a theory then consists of the choice of one of these algorithms for each process. The academic literature is replete with these kinds of "process algorithm" data, i.e. proposed and estimated descriptions of dynamic processes. However, these results are generally fragmentary, with one or at most a few process descriptions per journal article in a plethora of clinical and cohort studies. In each case, one or more statistical tests of significance are performed piecemeal. But we know relatively little of the potential biases and range of non-sampling errors in these results. To achieve Wilk's (1987) objective of rigorous assessment of these kinds of error, we need an overall framework plus a major synthetic effort. This effort would bring together the pieces of fragmentary theory and evidence into a coherent whole wherein the internal consistency of the pieces could be tested.

In effect, the "theory of health" we are proposing is a joining of an overall conceptual framework with a wide variety of elements. Each of these elements is a process description -- a nugget of empirical regularity which is itself the subject of scientific exploration, albeit in a partial way. This "theory" is not unlike a macroeconometric or cosmological model. The overall structure of an economy is represented in a macroeconometric model by a basic set of income and expendi-

ture identities, and formulae describing production possibilities. Then, a large number of time series econometric relationships are estimated. The resulting stochastic equations constitute the process algorithms or laws of motion for the model. They are generally estimated independently of one another, and brought together coherently within and by the framework. A realistic model of the evolution of a star cluster starts with an initial description of the positions and velocities of each member of a population of stars. Then the trajectory of each star is projected according to dynamic formulae encapsulating the basic laws of gravity and other relevant forces.

In both of these examples, the models are quite complex but still embody major simplifying assumptions -- in the direction of limiting complexity by ignoring less important interactions. For example, the gravitational forces acting on a star depend on all the other stars in the cluster, but most strongly only on the nearest stars. The test of whether it is reasonable to ignore the weaker forces of the more distant stars comes in assessing the agreement between the theory -- expressed as a numerical simulation of the evolution of the star cluster -- and observation.

Similar problems must be confronted in the health field. There is increasing evidence that the socio-cultural milieu and economic circumstances have a profound effect on health (Marmot, 1986; Wolfson et al., 1992). Increasingly, public policy is turning from the question of how many people have high cholesterol to what is it about our communities that predisposes members to certain dietary patterns (e.g., Ontario Premiers Council, 1991). Again, the framework in Figure 2 is agnostic. By design, it can incorporate significant results of this form, namely dependencies of individuals' dynamics on the attributes of their communities. The state space can be extended to include the relevant variables, and the dynamic processes can be expressed as functions of these community level variables. The only proviso is that they be measurable -- i.e. they can either be derived from individual level variables (e.g. the neighborhood's poverty rate) or are separately collected (e.g. ambient air pollution levels).

D. Role of Microsimulation

The paper so far has set out the basic problematique with regard to statistical information in the health field, and has sketched a broad theoretical structure. We turn now to consider the general role of microsimulation modeling (MSM) methods, including the way MSM can both represent and "solve" the theoretical model of health just described.

It might appear unconventional for an MSM to be at the center of a statistical system. That it is indeed natural in this case can be argued in stages. The first stage is the notion that abstract or formal mathematical models should be part of a statistical system. This is clearly necessary to clarify concepts and definitions, and to inform empirical observation. The systems of difference equations describing (some would say bastardized versions of) Keynesian macroeconomic theory and the matrix algebra of Leontief input/output analysis clearly play this role for the system of National Accounts.

The second stage is that such models must also be concretely realized in the form of numerical simulation models. In one sense, this is not news. Formal models of sufficient complexity are rarely analytically tractable. Also, people are often interested in the specific situation that flows form the current state of affairs -- i.e. the quantitative results for given initial conditions, not a general proposition for some abstract class of starting points. As already noted, macroeconometric simulation models have an almost symbiotic role with the National Accounts, and the use of input/output table-based economic models is a service offered by Statistics Canada. Similarly, much of the set of demographic statistics is motivated by the demand for population projections, another form of numerical simulation. Indeed, while it is not generally appreciated, life expectancy itself is a measure that is the result of a simulation -- the average age at death for a simulated steady-state population age structure. Thus, national statistical agencies are already closely involved with numerical simulation.

However, in another sense, simulation in the health area is needed for a novel reason. In economics or demography, to continue with these examples, all the data or statistics for a recent period are essentially directly observable. Life expectancy, on the other hand, can never be so observed. It is a synthetic indicator that requires observation plus the simulated answer to the "what if" question just noted. In turn, a family of "statistics" that is a generalization of the notion of life expectancy will also require numerical simulation to be "observed". Since such indicators are fundamental to creating good summary population health status statistics, numerical simulations to answer such standardized "what if" questions are an essential part of a System of Health Statistics, even in the absence of a need to generate projections.

Of course, most economic and demographic statistics are not collected just to feed into some simulation model; they are of broad interest in their own right. The same should be true in the area of health statistics. However, simulation models are of central *strategic* importance because

they give coherence. Without some sort of integrating analytical framework like a simulation model, data series risk being a hodgepodge, as is the case currently in the health area. Turning this point around, it is easy to consult with various data using constituencies and then compile a wish list of needed data, as was observed in the course of the Task Force consultations (Wilk, 1991). But if the wish list is at least partly driven or framed by a coherent structure like a simulation model, the proverbial whole has the possibility of being greater than the sum of its parts. The myriad constituent data series will be of interest in and of themselves. Moreover, when combined in the model, they will be useful for simulations and decision support; and with proper designed-in redundancy they will also generate data confrontations which can then serve as a check on non-sampling error, as will be illustrated below.

While it is not completely new for numerical simulation models to play a strategic role in statistical systems, the use of *micro*simulation is much more recent. This is the third stage in establishing our claim as to the centrality of MSM in a System of Health Statistics. The feasibility of MSM is intimately bound up with the revolution in computing power of recent decades. The need for microsimulation follows essentially from the heterogeneity of individuals and their behavior. The conventional partially aggregated or cell-based approaches of macroeconomics and much of demography are simply inadequate to capture the richness and texture of the phenomena of interest in the health area. (Indeed, we would argue that the same is true of economics and demography.) This in turn has been reflected in the structure introduced above in Figure 2. A microanalytic approach is not only computationally feasible and needed to represent accurately population heterogeneity. It is also needed to provide the common foundation for a range of familiar statistics, and a further range of highly desirable statistics, i.e. the generalizations of life expectancy described below. The data on individuals envisioned by the Template comprise an ideal *micro*data set.

Unfortunately, these data are not readily observed directly. The implied longitudinal household survey is impractical -- not least for reasons of respondent burden, privacy and confidentiality concerns, and the century we would have to wait until it was complete. Moreover, by the time the century of longitudinal follow-up was complete, the information would likely be useless because so much had changed. We need to be able to observe recent trends and regularities in behavior, and then make predictions about their implications.

In this context, MSM is a methodology for synthesizing the requisite biographies using more practical, albeit fragmentary, data sources. MSM can be thought of as a form of super imputation where a variety of partial pieces of data and partial descriptions of dynamic processes are annealed into a set of realistic though synthetic individual life cycle histories. We sketch the details of this synthetic process below. It must be supported by a large effort not only of meta-analysis, as increasingly undertaken in the epidemiologic literature, but more accurately a process of meta-synthesis. In meta-analysis, a number of cohort studies considering the same phenomenon, say the relationship between serum cholesterol and coronary heart disease events, are examined and efforts are made to pool the results so as to increase the effective sample size. MSM goes beyond this by taking quantitative results from a variety of domains (e.g. labour force participation transitions, risk factor dynamics, disease incidence hazard functions) and seeking to draw out their joint implications.

The result is not just one "baseline" instance of the longitudinal microdata set implicit in Figure 2. It is also a simulation model capable of constructing hypothetical alternative versions of this dataset where one or more factors have been changed. A much more familiar example of this kind of hypothetical simulation experiment is cause-deleted life expectancy -- how long could we expect to live if there was no mortality at all from a particular kind of cancer, for example.[6] MSM is able to construct generalizations of this kind of indicator, and to base the computations on explicit models of the complex interactions among risk factors and comorbidities.

In addition, by using MSM methods, the underlying longitudinal population data are always available for more detailed analysis. Thus, other kinds of related statistics can be readily computed. Moreover, these synthetic individual biographies are always available for assessing the plausibility of the results of the simulation. This is not possible with conventional (partially aggregated or cell-based) multi-state life table methods. As shown below, when the life paths implicit in life table analyses are made explicit, they can be highly implausible.

---

6 Note that cause-deleted life expectancies are quite simplistic. They are based implicitly on a model that treats each cause of death as strictly independent. One example where this is clearly false is for heavy cigarette smokers. If they are hypothetically prevented from ever dying of lung cancer, as in a lung cancer-deleted life expectancy calculation, they would still be at elevated risk of dying from coronary heart disease, chronic obstructive pulmonary disease, etc. However, this underlying elevated risk is completely ignored in the usual cause-deleted life table estimates.

The hypothetical alternative versions of the basic longitudinal microdata set of Figure 2 generated by MSM can provide the basis for decision support. For example, cause-deleted life expectancies have often been used to form a league table of the most "important" diseases, with heart disease and cancers in first and second place, since they account for the largest expectations of years of life lost in developed countries. As an extension of this idea, MSM allows the creation of a much broader league table that includes as the ranking measure not only mortality (based on increments in life expectancy due to the hypothesized elimination of a given disease) but also morbidity and disability while alive. These measures would be based on increments in life expectancy adjusted for how sick people are year by year -- health status-adjusted life expectancy or population health expectancy (PHE, for short).

Furthermore, MSM-based analyses can encompass not just diseases but also risk factors and health-affecting interventions. In this broader "contest", it is likely that arthritis and dementia would appear much closer to the top of the human health problem league table. Also, risk factors like smoking, radon, and cholesterol could be ranked by comparing the impact of deleting each risk factor in turn and estimating the resulting changes in PHE as well as in life expectancy (e.g. see Gentleman et al., 1991). Correspondingly, interventions like coronary artery bypass graft surgery could be compared to smoking cessation programs and then ranked according to a consistent and popularly accessible health outcome measure -- PHE.

Indeed, regular publication of PHE by the national statistical agency would provide a major benefit in remedying one of the key problems in the health statistics area noted at the beginning of this paper. This is the imbalance between information on inputs and outputs. PHE is a simple index that would balance the current preoccupations with health care spending as a percentage of GDP. However, estimation of PHE requires microsimulation[7]. Thus, if PHE is to be a core concept in a system of health statistics, then MSM must also lie at the core of this new system.

MSM can play a role in a statistical system beyond decision support and the construction of summary indices. It is also significant for basic research. National statistical agencies are the closest analogue for social scientists to the particle accelerators and radio telescopes of physicists and astronomers. They all provide huge volumes of basic data. For the astronomers, where to point

---

7 Multi-state life table methods can also be used, but they are highly restrictive compared to the general range of phenomena the model should ideally encompass. Also, as already noted, life tables are a form of simulation, albeit not *micro*simulation.

the radiotelescope, and what frequencies to monitor, are not random choices. Rather, they flow from very detailed deductive processes which integrate previously collected data with theories, typically by use of models which are often solved by numerical simulation. Thus, data collection strategies in areas like radio astronomy are intimately bound up with analysis of previously collected data and theorizing based on simulation modeling.

Similar processes are involved in basic social science research and socio-economic statistics, but data collection and data analysis seem to be less tightly coupled. In our view, having an MSM at the core of a statistical system will induce a tighter coupling and hence more fruitful basic research. This effect can be illustrated as follows. An MSM is built to estimate PHE. In a number of areas, data are very limited or even non-existent. In these areas, "guesstimates" are made, and then a sensitivity analysis is undertaken to see which of the various guesstimates is most important to estimates of PHE. The result should be increased priority to the collection of data relevant to the most important guesstimate.

Even if such sensitivity analysis to various guesstimates does not feed back immediately to data collection priorities, the process has value to basic social science research. Much of this research is conjectures about possible causal pathways and their magnitudes. For example, how important is unobserved heterogeneity with respect to some kind of innate "frailty" (Vaupel and Yashin, 1985); what would be the impact of relaxing the assumption of independence of competing risks? These basic questions cannot be addressed by new data collection, either because we simply do not know how (innate frailty) or because it is logically impossible (dependent competing risks; Elandt-Johnson, 1980). The alternative is to use a model. Just as numerical computer simulation has become an essential part of the "laboratory apparatus" of meteorologists and aerodynamicists, microsimulation models can become a standard part of the laboratory apparatus for social science research.

E. How To Microsimulate

We have now sketched a theoretical framework for population health, and indicated the key role that numerical microsimulation models can play in both holding all the requisite information, and in "solving" or drawing out theimplications of the observed empirical regularities. In this section we turn to a more detailed description of one instance of an MSM, POHEM. By design, POHEMs structure is very similar to that shown in Figure 2. The rows in the individual's biog-

raphy represent, in effect, the record layout for the hypothesized longitudinal microdata set, and as discussed earlier, the "state space" for the model. The columns represent the individual at various ages. While it was explicit in Figure 1a, it is only implicit in Figure 2 that there is a series of "slices" or planes going back into the page representing a population of individuals.

Microsimulation synthesizes this data set in a series of three nested loops. At the innermost level, the simulation process creates one column vector in an individual's biography by synthesizing each element in the vector working from top to bottom. This process starts with the "birth" of a healthy alive individual at age zero. This is followed by the repeated application of the appropriate dynamic algorithms to fill in column vectors for successive ages until the individual dies. This is the second loop in the simulation. Finally, the third and outermost loop synthesizes a large sample of individuals.[8]

We can use an example of recent and typical life table analysis to illustrate the microsimulation process by sketching just how an MSM would be constructed to reproduce the life table results. It is not as efficient computationally to use MSM instead of life table methods where the latter are feasible (provided the only desired results are summary statistics like life expectancy, since life tables do not have explicit microdata foundations). Our purpose rather is to take a simple example (from the viewpoint of MSM) as an illustration. Subsequently, it will be seen that POHEM as an instance of MSM is considerably richer and more detailed. The specific example that will be used is the estimation of disability-free life expectancy (DFLE) as produced by increment-decrement multi-state life tables (e.g., Rogers et al., 1989; Crimmins et al., 1989). In MSM terms, this life table analysis corresponds to a very much simplified version of the biographical record layout in Figure 2. We need at most three rows -- one for alive or dead, one for healthy or disabled, and one for the health status-adjusted or full health equivalent value of the life-year. In other words, individuals' state space in each year of life consists of a three-tuple or three element column vector.

---

8 The order of the two outermost loops can be reversed. In other words, the simulation could proceed individual by individual until a full (pre-specified) sample of column vectors is completely synthesized. Then as the outermost loop, the simulation could proceed year by year until the last individual dies (assuming the simulation applies to a synthetic birth cohort as is typical of life table analyses).

These three-tuples are simulated starting at birth and moving from left to right in the sense of Figure 2 using "laws of motion" or dynamic algorithms defined by simple functions as follows: Being in the alive/dead and healthy/disabled states at age *a* is assumed to depend stochastically only on age and prior disability. If the individual is alive and healthy at age *a-1*, the possible transitions to age *a* are: no change, become disabled, or death. Similarly, if the individual is alive and disabled at age *a-1*, the possible transitions are: no change, become healthy, or death. The transitions are based on observed probabilities. In other words, this is a first order markov process.

In an actual simulation, a monte carlo process is applied where random numbers are drawn, and depending on the draws and the individual's state at age *a-1*, the individual may be simulated to die, become disabled, or get better. To represent this, assume that an individual's biography has been synthesized up to age *a-1*. Empirical observations provide the basis for two sets of transition probabilities. These are looked up from the model's input data -- $m(i,a)$ and $d(i,a)$, the probability of dying, and the probability of changing disability level respectively, at age *a* given disability status *i* at age *a-1*. The simulation process proceeds by drawing a random number from a uniform distribution over the range 0 to 1. If the number is in the $[0, m(i,a)]$ interval, the individual is simulated to die; if in the $[m(i,a), m(i,a) + d(i,a)]$ interval, the individual survives but changes disability status; if the number is in the $[m(i,a) + d(i,a), 1]$ interval the individual survives and remains in the same disability state.

This process of drawing random numbers and testing them against the exogenously estimated transition probabilities is repeated over ages to complete the individual's synthetic biography. Finally, many such biographies are generated to synthesize a large sample[9]. We thus have completely synthesized a population of complete life cycle biographies with two of the three elements in the state space -- alive/dead and healthy/disabled. Lastly, the third health status value is computed from the contemporaneous alive/dead and healthy/disabled values according to a (simplistic) function that assigns a value of one if alive and not disabled, zero otherwise.

---

9 The sample size is chosen to bring the monte carlo sampling error down to the level where the resulting estimates have the desired level of precision. Monte carlo error can be determined by sample reuse methods.

In this example of using MSM to "solve" for the life table resulting from these simple dynamic algorithms, the conventional summary statistics can be computed as follows: summing across ages in the alive/dead row (assuming alive has a value of one and dead a value of zero), and then averaging over all the individuals in the sample results in an estimate of life expectancy. The same summing and averaging applied to the third health status row gives DFLE. As shown in Wolfson and Manton (1992), a very wide range of models of disability, risk factor, and disease processes can be expressed in terms of the state space of Figure 2 combined with monte carlo microsimulation. MSM clearly nests conventional life tables as a special case.

While life table models typically embody strong simplifying assumptions, it is more realistic to consider many processes as interdependent and simultaneous. For example, getting married, buying a house, finishing school, and entering the labour market are decisions or socio-economic transitions occurring in early adulthood that are often jointly determined. Such interaction can be included in MSM. One way is to assume that the "laws of motion" for processes such as these are at least block recursive. In other words, each row variable has dynamics whose formula has on its right-hand side only variables to the left in Figure 2 (i.e. any lagged variable or any attribute in the individual's biography up to age $a-1$) plus variables in the same column but above. Then if the simulation starts at the top of the column for the i-th individual at age a, and works down the column synthesizing one variable at a time, all the required information is always available.

However, this block recursive structure may not be appropriate for some processes such as co-morbidity. For example, if the simulation is built on an annual time step, within each year an individual who is a heavy smoker will be at elevated risk of dying from both lung cancer and a heart attack. Such joint risks can be simulated by computing all the competing mortality hazards to which the individual is exposed that year, and then explicitly simulating their combined impacts as well as the progression of each incident disease. The major problem here is not any constraint imposed by the MSM methodology. Rather, it is the difficult empirical questions raised in gathering the appropriate data and estimating any interactions among hazard functions.

Let us turn now to a sketch of the current version of POHEM. To begin, POHEM creates not just individuals, but male-female pairs. This is done in anticipation of a marriage or a common law union. As well, children and remarriage partners are explicitly included in this extended

family structure or "case". The full lifecycle of each case is simulated, not just one individual at a time. A case is completed with the death of the last adult (and the last child leaving home) before another is commenced.

Unlike the very simple DFLE life table model just described in order to illustrate the MSM method, POHEM includes a large number of sometimes complex processes or dynamic algorithms. For some processes, algorithms for several variants are included in the software. The processes explicitly available in a POHEM simulation, grouped in the same way as the rows in Figure 2, are as follows:

Socio-Economic Status

Educational Attainment -- endowed at birth by drawing from univariate distributions and husband-wife correlations based on Canadian census data.

First Union -- either legal marriage or common law union (CLU); probability at each age represented by a multivariate hazard function of age, sex, education, fertility (for females), labour force history, CLU history, and pre-ordained marriageability (for unobserved heterogeneity, Rowe and Wolfson, 1990).

First Spousal Age Difference -- based on age at marriage, and observed joint distribution of brides' and grooms' ages.

Fertility -- probability based on age, parity, and marital status.

Union Dissolution -- either divorce or separation; probability at each age represented by a multivariate hazard function of age, duration of marriage, presence of children, labour force experience, age at marriage (Rowe and Wolfson, 1990).

Child Custody -- dependent on marital status.

Child Leaving Home -- probability based on age, sex, and birth order.

Remarriage -- probability based on age, sex, divorce versus widow(er).

Second Spousal Age Difference -- based on marrying person's age at marriage, sex, and prior marital status drawn from the observed joint distribution of brides' and grooms' ages.

Labour Force Participation -- probability of entry or exit at each age represented by a set of multivariate hazard functions of age, sex, marital status, presence of children by age group, educational attainment, and duration in state (Picot, 1989).

Labour Market Earnings -- dollar level each year based on an autoregressive stochastic process with parameters based on age, sex, and strength of labour force attachment (Kennedy, 1986).

Risks

Radon -- endowed at birth by drawing from a lognormal fit to the observed distribution of levels within residential dwellings.

Blood Pressure, Obesity, Smoking and Cholesterol -- quadrivariate joint density at age *a* derived as first order Markov function of quadrivariate joint density at age *a-1*, age, and sex based on analysis of the 1978 Canada Health Survey (Gentleman and Robertson, 1991).

Diseases

CHD -- exactly as in Weinstein et al., (1987) except that Canadian risk factor distributions and treatment protocols are substituted (Wolfson and Birkett, 1989).

Lung Cancer -- incidence conditional on cumulative radon and tobacco exposure up to age *a-10*; type, and stage assigned based on cancer registry distributions by age and sex; progression and case fatality conditional on type and stage based on meta-analysis of clinical literature (Gentleman et al., 1991).

Breast Cancer -- incidence based on age, parity, age at first birth; progression and case fatality based on fourth-order Markov transitions among disease-free, localized recurrence, and metastatic states.

Dementia -- incidence based on age and sex; progression based on duration since onset (Forbes and Barham, 1989).

Arthritis/Rheumatism -- under development; incidence and progression based on combination of the 1986 post-censal Health and Activity Limitations Survey and expert consensus (Tugwell et al., 1992, Chambers et al., 1991)

Mortality from Other Causes -- based on age, sex, and marital status in turn derived from vital statistics and census data.

Functional Status -- under development; 1990 Ontario Health Survey (Ontario, 1990) categories are planned (gross motor, dexterity, hearing, seeing, speech, cognitive, emotion, pain); an alternative module is available currently that has three states (mild, moderate, and severe disability) based on first order Markov transitions derived from the 1986 post-censal Health and Activity Limitations Survey (this module ignores risk factors and diseases).

Costs -- CHD treatments based on McMaster-Chedoke hospital; Lung Cancer treatments based on expert medical oncology advice (Will and Gentleman, 1992).

Health -- Torrance-style multi-attribute value scale is planned (Torrance, 1987). Currently using arbitrary weights for disease states similar in style to Wilkins and Adams (1983).

These processes or "laws of motion" are applied year by year and individual by individual in POHEM. The simulation is exactly analogous to (but considerably more complex than) the process sketched for disability-free life expectancy (DFLE) using increment-decrement life tables in the context of Figure 2 at the beginning of this section. In this way, complete synthetic biographies are built up for a representative sample of the population (more precisely, a steady-state birth cohort in the sense of a period life table). In effect, a complete longitudinal microdata set has been imputed or annealed from diverse and partial descriptions of the dynamics of health and health-related processes. Stacking all the resulting individual biographical "rectangles" (state space along the vertical axis, age along the horizontal, individuals going back along the third dimension) results, with some poetic license, in a "data cube" like that shown in Figure 3a.

[Figure 3a about here]

F. Coherent Sets of Statistics

This data cube provides a microdata foundation and hence a coherent basis for a variety of derived statistics and graphs. These include families of related statistics. As one illustration, the health status information along the bottom rows can be readily transformed into a conventional survival curve by the following algorithm: extract the bottom plane of the data cube; convert all non-zero entries to one's; sum across the columns to derive a vector of life lengths (LLs); sort these life lengths in ascending order; graph the resulting distribution of LLs starting in the upper-right of the survival curve and proceeding toward the lower-left.

Survival curves are, of course, very conventional, and of decreasing relevance with the increase in chronic disease. It is of vital importance to have indicators of how healthy people are while they are alive, as well as the expected distribution of their life lengths. The POHEM data cube readily supports construction of a generalization of the survival curve that both captures this notion and provides the basis for a set of related statistics. This time, we do not throw away the health status information in the bottom plane of the data cube -- the values *between* zero and one in the years when individuals are alive but in less than full health.

Our software routine again extracts the bottom plane of the data cube. But this time, we use a series of gray shade levels to represent varying degrees of severity of illness, and a notion like contour plotting. The conventional survival curve is constructed as before. But we now begin shading the area under the curve. We start in the lower-left with a pure white area representing those in full health. Then for each of a series of threshold levels x of less than full health (e.g. .9, .8, .7, ...), we find the intervals (there may be more than one) in each individual's biography where s/he was alive and had a health status index value above this threshold x; cumulate the durations of these intervals for each individual -- call these individuals' life lengths spent in health states valued better than x or $LL(x)$ (so that $LL(0) = LL$); sort the sample of individuals by these $LL(x)$ durations; plot the contour (or equivalently the survival curve in health with a value score better than x); and shade the space between this and the previous contour with a slightly darker gray.

Figures 3b and 3c illustrate this process using images from the Template. The top portions of these figures show the bottom plane of the data cube turned up on its edge. The bottom portion of Figure 3b assumes that there is one clearly defined disability state and shows the survival curve for DFLE, while the bottom portion of Figure 3c assumes a number of levels of disability. Finally, Figure 4a shows actual results from a POHEM run using the 1986 Canadian post-censal Health and Activity Limitations Survey data on disability prevalences and first order markov transitions.

[Figures 3b, 3c, and 4a about here]

If we compute the area under the conventional survival curve, it is simply life expectancy. If we recompute this area with weights corresponding to the average health status index values in each range represented by the gray shade levels, we have our preferred summary index of health status, Population Health Expectancy (PHE). Of course, PHE could be computed directly from the data cube by summing over the entire bottom plane and then dividing by the sample size.

In addition, PHE can be readily disaggregated, given the full data cube. For example, it could be computed for various subsets of the population -- e.g. by gender, and whether the individual ever had a given disease like coronary heart disease or arthritis. It could also be broken down by age interval, to get for example the portion of the discrepancy between PHE and life expectancy (LE) attributable to women suffering from arthritis after age 65. This is analogous to a disaggregation of the "all items" consumer price index into components for food, transportation, etc.

The data cube produced by an MSM process like POHEM also supports distributional analyses. The gray shaded survival curve in Figure 4a ignores the individual structure of spells or sojourn times in various levels of health. But the POHEM data cube provides an explicit microdata foundation which can be analyzed from a variety of perspectives. For example, let health status-adjusted life length (HSALL) equal the sum for any individual of the bottom row of the data cube, i.e. life length weighted by the value of the summary health status index that year. Figure 4b gives a scatter plot of the proportion of each individual's life length spent in the equivalent of full health against life length, i.e. a plot of HSALL/LL against LL, for a subset of the POHEM simulated population underlying Figure 4a. It also shows the univariate distributions of LL and HSALL/LL adjacent to the horizontal and vertical axes respectively -- all disaggregated by gender.

Figure 4c shows a different kind of distributional data. Here we have the frequency distribution of spell durations in a disability-free state. Two such distributions are compared. One is from the same simulation underlying Figures 4a and 4b, which in turn is similar to the increment-decrement life table model described earlier. The other is from a POHEM run where the first order markov transitions were ignored (i.e. disability at each age *a* is assumed to be independent of disability at age *a-1*), and only the cross-sectional prevalences are used. This represents the conventional Sullivan method (1971) for computing disability-free life expectancy (DFLE, e.g., see Robine, 1986). This graph clearly shows the dramatic impact of alternative assumptions regarding the "laws of motion" even in the simplest models -- data that are virtually impossible to obtain from multi-state life table methods. The Sullivan method durations, based as they are on an assumption of complete independence in disability state from one year to the next, are clearly unrealistic.

In addition to displaying statistical results from one "baseline" simulation of the data cube, hypothetical alternative scenarios could also be simulated. For example, suppose a set of cause-deleted simulations were run for coronary heart disease (CHD) and arthritis. When this is done using life tables and life expectancy, CHD comes out at the top of the league table as the most serious disease (especially since arthritis is generally not fatal). In the POHEM context, with explicit diseases and comorbidity, such simulations (currently underway) might yield quite different impressions. For example, it may turn out that even for individuals with CHD, arthritis contributed a greater proportion of the "morbidity gap" (the difference between LL and HSALL) than CHD.

G. Illustrations of Data Integration and Confrontation

A key issue raised at the outset of this paper was non-sampling error. It was claimed that the coherence of statistical *systems* could be used to generate "data confrontations" and hence a method for revealing and assessing the extent of non-sampling error. In this section we return to this claim and give two illustrations of the use of POHEM for this purpose. In other words, a coherent system of health statistics built around a microsimulation model will be used to elucidate inconsistencies among diverse data sets, and hence give some quantitative indication of the importance of non-sampling error.

The first example focuses on the POHEM lung cancer module. Lung cancer can be simulated by POHEM with three variants. The simplest uses standard age/sex-specific mortality rates by cause, with lung cancer as one cause of death. These data come from vital statistics and the census. They take no account of incidence and morbidity, so lung cancer is effectively modeled as having an infinitesimally short morbid phase always followed by death; all incident cases of lung cancer are fatal.

The next variant explicitly distinguishes incidence and case fatality. Incidence is based on cancer registry data by age and sex. Incident cases are then disaggregated by cell type and stage based on a special chart review study. Finally, disease progression and case fatality are explicitly modeled based on a detailed literature review and expert clinical judgement (Will and Gentleman, 1992).

The most detailed variant builds on the second. It is the same except that incidence rates are adjusted for risk factor exposures. Data from residential dwellings are used to assign radon exposures, and data from the 1978 Canada Health survey are used to generate age profiles of cigarette smoking. Then a risk function from the epidemiological literature (Whittemore and McMillan, 1983) is used to scale each individual's risk of contracting lung cancer as a function of his/her personal risk factor history. Given an incident case, lung cancer progression and case fatality are modeled in the same way as the second variant.

Table 1 shows the results of POHEM simulations with each of these three variants. For each simulation, life expectancy should in principle be identical, as should average age at death from lung cancer, and this is in fact virtually so. However, there does seem to be a problem with incidence and case fatality. The product of these two rates, which is deaths from lung cancer, should also be identical across all three scenarios. But the simulations suggest an inconsistency with cancer incidence registry-based rates yielding lung cancer death rates that are about 0.7 to 1.1 percentage points lower than those coming from death certificates. We are currently trying to clarify the source of this discrepancy. It does not appear to be the case fatality rates, since these have been checked against the special chart review sample of cancer registry cases used to determine cell type and stage, which have also been followed up for mortality. Nor is it any obvious differences in the ICD codes used for lung cancer from the different data sources. In any case, this is a clear example of data confrontation from otherwise disparate data sources that has been rendered feasible by the coherent structure of POHEM.

[Table 1 about here]

The second example is one of *potential* data confrontation. Unlike lung cancer where cancer registry data can be used, there are no broad-based population data on CHD incidence. However, we can use the POHEM CHD model to estimate what the age/sex specific incidence rates would have to be, given CHD mortality rates (from vital statistics), risk factor distributions (from the 1978 Canada Health Survey, Gentleman and Robertson, 1991), and a model of relative risks and CHD progression and case fatality taken from Weinstein et al. (1987).[10] In effect, we have used POHEM to solve for the basic incidence rates which, when combined with the other disease onset and progression data, will exactly reproduce the observed pattern of CHD mortality rates. Figure 5 gives the resulting incidence curves for males and females, as well as the observed CHD mortality rates.[11]

[Figure 5 about here]

The incidence curves are notable in two respects. First, as hypotheses they are eminently testable. A survey or some sort of data collection process is readily imagined that can measure CHD incidence directly, so that we would have a data confrontation, and hence the possibility of clues regarding consistency (or lack thereof) and the extent of non-sampling error. Second, even in the absence of further data, the blip in the incidence curve for males suggests some sort of anomaly. Our current view is that this blip may be due to an inconsistency in the Weinstein et al. model in the CHD case fatality rates in the highest open-ended age range in relation to those at earlier ages. In any case, the key point is that by assembling and integrating diverse statistics in a systematic and coherent framework, we are now able to detect possible errors that have nothing to do with sampling.

---

10 We are grateful to Dr. Milton Weinstein for providing us with the complete details of his model.

11 A related form of data confrontation on which we are working is to ask what rates of hospitalization by age, sex, and type of CHD (e.g. acute myocardial infarction, uncomplicated angina pectoris) would exactly reproduce current hospital morbidity statistics -- for which a complete register does exist.

H. Conclusions

This paper has motivated, described and illustrated the use of the POHEM microsimulation model in the context of a number of broad statistical and health science issues. The principal claims and conclusions can be summarized in the following points:

- Current health information systems are seriously deficient in at least two major respects. First, there is a serious imbalance with relatively more data available on the inputs to health care, and relatively little on the health of the population. Second, they are not really systems; health data tend to be a hodgepodge with no coherence.

- Coherence in statistics, while a rather abstract notion, is fundamental. Achieving coherence is possible by using a conceptual framework and developing a *system* of statistics.

- A conceptual framework and system of statistics necessarily embodies some theory. An individual life cycle framework is vital in this case to a "theory of health". The Health Information Template provides a good general framework for a theory of health.

- A major benefit of coherence is a rigorous means of addressing data quality, particularly the relatively more difficult questions of non-sampling error.

- Much desired information is not directly observable. A key example is life expectancy. It is the result of a numerical simulation used to construct the answer to a hypothetical question, even though the vast majority of people who use this kind of statistic do not realize it.

- Numerical simulation models can play a key role in health statistics. First, by producing indicators that are generalizations of life expectancy like population health expectancy (PHE), they can remedy the imbalance just noted. Second, they can provide the realization of a conceptual framework, and thus coherence and the possibility of data confrontations, and in turn a means for addressing non-sampling error. POHEM is such a model.

- Simulation models like POHEM can be readily used to construct answers to hypothetical "what if" questions. For example, generalizations of the notion of cause-deleted life expectancy are easily constructed.

- A coherent conceptual framework like that underlying POHEM can be used to set data collection priorities. Areas where data are weak or non-existent can be the subject of guesstimates or inferences. If these areas turn out to be quantitatively significant for important concerns, data collection priorities should be correspondingly high.

- A model like POHEM can provide the focus for a statistical system in a national statistical agency. Such organizations would benefit from augmenting their traditional roles of conducting and disseminating piecemeal a diversity of surveys by developing coherent statistical systems. The National Accounts today provide the only major example.

- Simulation models like POHEM are tools for basic research. For example, they can be used to formulate hypotheses and obtain (indirect) bounding quantitative estimates regarding the importance of various phenomena.

- Simulation models like POHEM, with their capacity for "what if" analyses, are tools for decision support. They can play an important role in policy analysis and provide a focus for policy analysis groups within government agencies.

POHEM is clearly a work in progress. Still, it has allowed us to reinforce and support our initial diagnosis -- that current health statistics suffer from incoherence and imbalance. Moreover, POHEM is a "proof by construction" for many of the propositions just listed. Based on experience to date, POHEM and its underlying methodology are very promising.

References

Chambers, L.W., D.L. Reynolds, E.M. Bradley, K.J. Bennett, C.H. Goldsmith, G.W. Torrance and P. Tugwell (1991), "Physical Disability Among Canadians Reporting Musculoskeletal Disease", mimeo, Statistics Canada, March, Ottawa.

Crimmins, E.M., Y. Saito, and D. Ingegneri (1989), "Changes in Life Expectancy and Disability-Free Life Expectancy in the United States", *Population and Development Review* 15:235-267.

Dewdney, A.K. (1984), *Scientific American*

Dowd, J.E., K.G. Manton, E. Stallard and M.A. Woodbury (1992), "Multivariate Risk Factor Dynamics and the Burden of CHD and other Noncommunicable Diseases", forthcoming, International Heart Health Conference, Victoria, May.

Elandt R.C., N.L. Johnson, (1980), *Survival Models and Data Analysis*, John Wiley & Sons Inc., North Carolina.

Forbes, W.F. And J. Barham (1989), "An Investigation of the Prevalence and Incidence of Dementia in Populations Aged 65 Years and Over", Research Report 2, Program in Gerontology, University of Waterloo, Waterloo, March.

Gentleman, J.F. and D. Robertson (1991), "Smoothing Procedures for Simulated Longitudinal Microdata", *Proceedings of the 1990 Statistics Canada Symposium on Analysis of Data in Time*, Ottawa.

Gentleman, J.F., M.C. Wolfson, and B.P. Will (1991), "A Microsimulation Submodel for Lung Cancer", *Proceedings of the Statistical Computing Section of the American Statistical Association, 1991 Joint Statistical Meetings*, Atlanta Georgia.

National Centre for Health Statistics (various years), *Health United States*, U.S. Department of Health and Human Services, Public Health Service, Maryland, March, DHHS Pub. No. (PHS) 90-1232.

Kennedy, B. (1986) "The LIPPS Earning Module", mimeo, Social and Economic Studies, Statistics Canada.

Marmot, M.G. (1986,) "Social Inequalities in Mortality: The Social Environment" in R.G. Wilkinson (Ed.), *Class and Health: Research and Longitudinal Data*, London, Tavistock Press, 21-33.

Ontario Ministry of Health (1990), Ontario Health Survey Questionnaire, Toronto.

Ontario Premiers Council on Health Strategy (1991), *Nurturing Health: A Framework on the Determinants of Health*, Healthy Public Policy Committee, March, Toronto.

Picot, G. (1989), "Modelling the Lifetime Employment Patterns of Canadians", in *The Family in Crisis: A Population Crises?*; J. Legare, T.R. Balakrishnan and R.P. Beaujot Eds; Proceedings of a Colloquium organized by the Federation of Canadian Demographers and Sponsored by the Royal Society of Canada, University of Ottawa, November.

Pommier, P. (1981), "Social Expenditure: Socialization Expenditure? The French Experience with Satellite Accounts", *Review of Income and Wealth*, December.

Robine, J.M. (1986), "Disability-Free Life Expectancy, General Indicators of the Health of Populations", Conseil des affaires sociales et de la famille, Scientific Report, Quebec.

Rogers, A., R.G. Rogers, and L.G. Branch (1989), "A Multistate Analysis of Active Live Expectancy," *Public Health Reports* 104:222-225.

Rowe, G. and M.C. Wolfson (1990), "Biased Divorce: Validation of Marital Status Life Tables and Microsimulation Models", paper presented to the United Nations Economic Commission for Europe, Seminar on Demographic and Economic Consequences and Implications of Changing Population Age Structure, Ottawa, Canada. September 24-28.

Sullivan, D.F. (1971), "A Single Index of Mortality and Morbidity", *HSMHA Health Reports* 86:347-354.

Torrance, G.W. (1987), "Utility Approach to Measuring Health-Related Quality of Life", *Journal of Chronic Diseases*, Vol. 40, No. 6, pp. 5930-600.

Tugwell et al. (1992) arthritis progression ***XXX***

Vaupel, J.W. and A.I. Yashin (1985), "Heterogeneity's Ruses: Some Surprising Effects of Selection of Population Dynamics", *American Statistician*, 39:176-185.

Weinstein, M.C., P.G. Coxson, L.W. Williams, T.M. Pass, W.B. Stason, and L. Goldman. (1987), "Forecasting Coronary Heart Disease Incidence, Mortality and Cost: The Coronary Heart Disease Policy Model", *American Journal of Public Health* 77:1417-1426, and personal communication.

Whittmore, A.S., A. McMillan (1983), "Lung Cancer Mortality Among U.S. Uranium Miners: A Reappraisal", National Cancer Institute, 71:489-499.

Wilk, M.B. (1987), "The Concept of Error in Statistical and Scientific Work", *Proceedings of the Census Bureau Annual Research Conference*, Washington, D.C.

Wilk, M.B. (1991) *Report of the National Task Force on Health Information*, R.H. Coats Building 24A, Ottawa, K1A 0T6.

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

Will and Gentleman (1992) (lung cancer module documentation) ***XXX***

Wolfson, M.C., G. Rowe, J. Gentleman and M. Tomiak (1992) "Career Earnings and Death: A Longitudinal Analysis of Older Canadian Men", *Journal of Gerontology*, forthcoming.

Wolfson, M.C. and K.G. Manton (1992), "A Review of Models Incorporating Notions of Population Health Expectancy", presented to the fifth international meeting of REVES, Ottawa.

Wolfson, M.C. (1992a), "POHEM - A New Approach to the Estimation of Health Status Adjusted Life Expectancy", paper presented to the first workshop of the International Network on Health Expectancy, Quebec City, September 11-12, 1989 and *Cahiers quebecois de demographie*, forthcoming.

Wolfson, M.C. (1992b) "A Template for Health Information", *World Health Organization Statistical Quarterly*, forthcoming, March. (software diskette also available from the author)

Wolfson, M.C. (1991), "A System of Health Statistics -- Toward a New Conceptual Framework", *Review of Income and Wealth*. March.

Wolfson, M.C. and N. Birkett (1989), "POHEM and an exploration of CHD", presented to the first annual Canadian Epidemiology Conference, August, Ottawa.

World Health Organization (1980), *International Classification of Impairments, Disabilities, and Handicaps: A Manual of Classification Relating to the Consequences of Disease*, Geneva.

Figure 1a

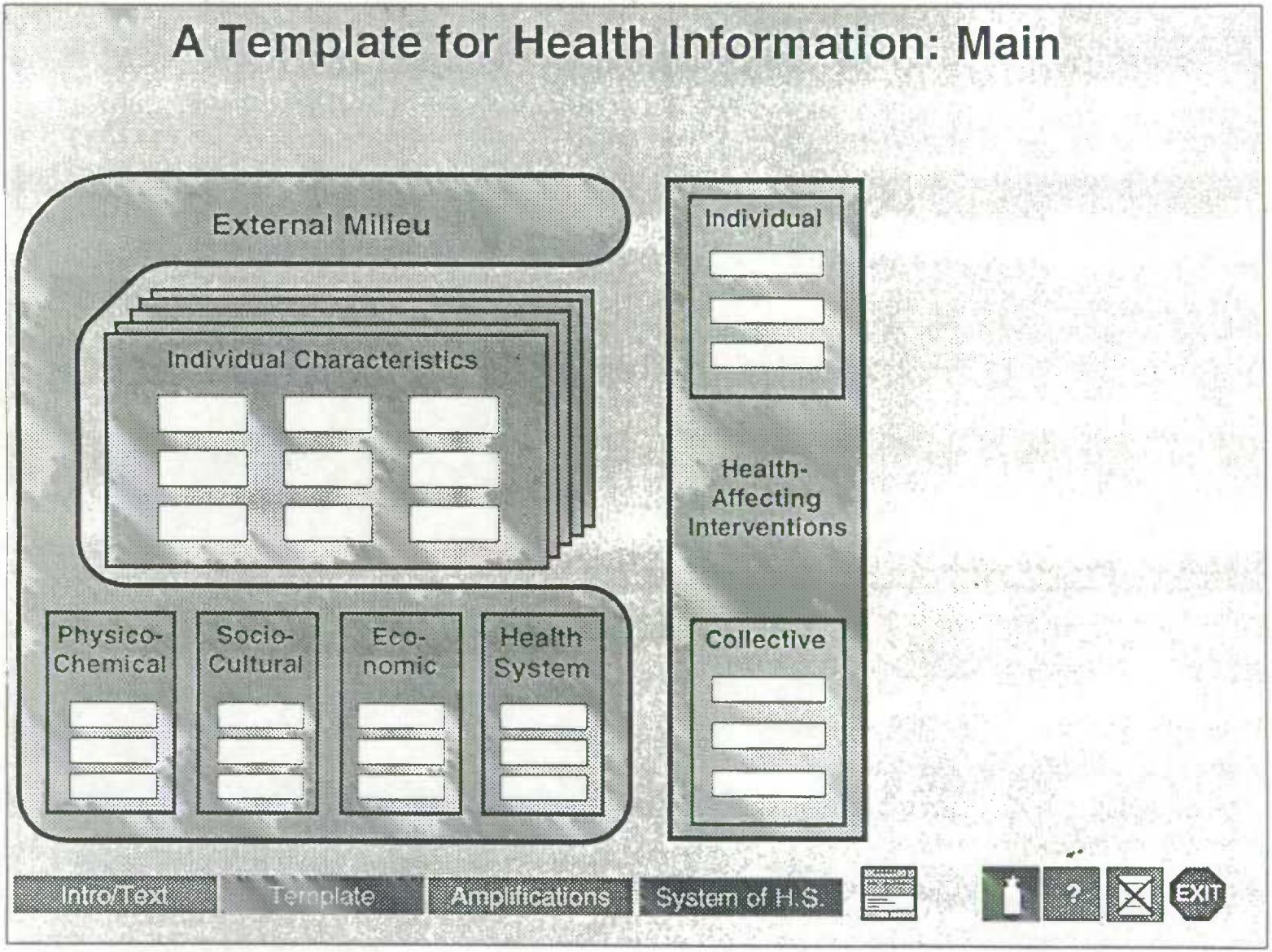
A Template for Health Information: Main
External Milieu
Individual Characteristics
Physico-Chemical
Socio-Cultural
Eco-nomic
Health System
Individual
Health-Affecting Interventions
Collective
Intro/Text
Template
Amplifications
System of H.S.
?
EXIT

Figure 1b

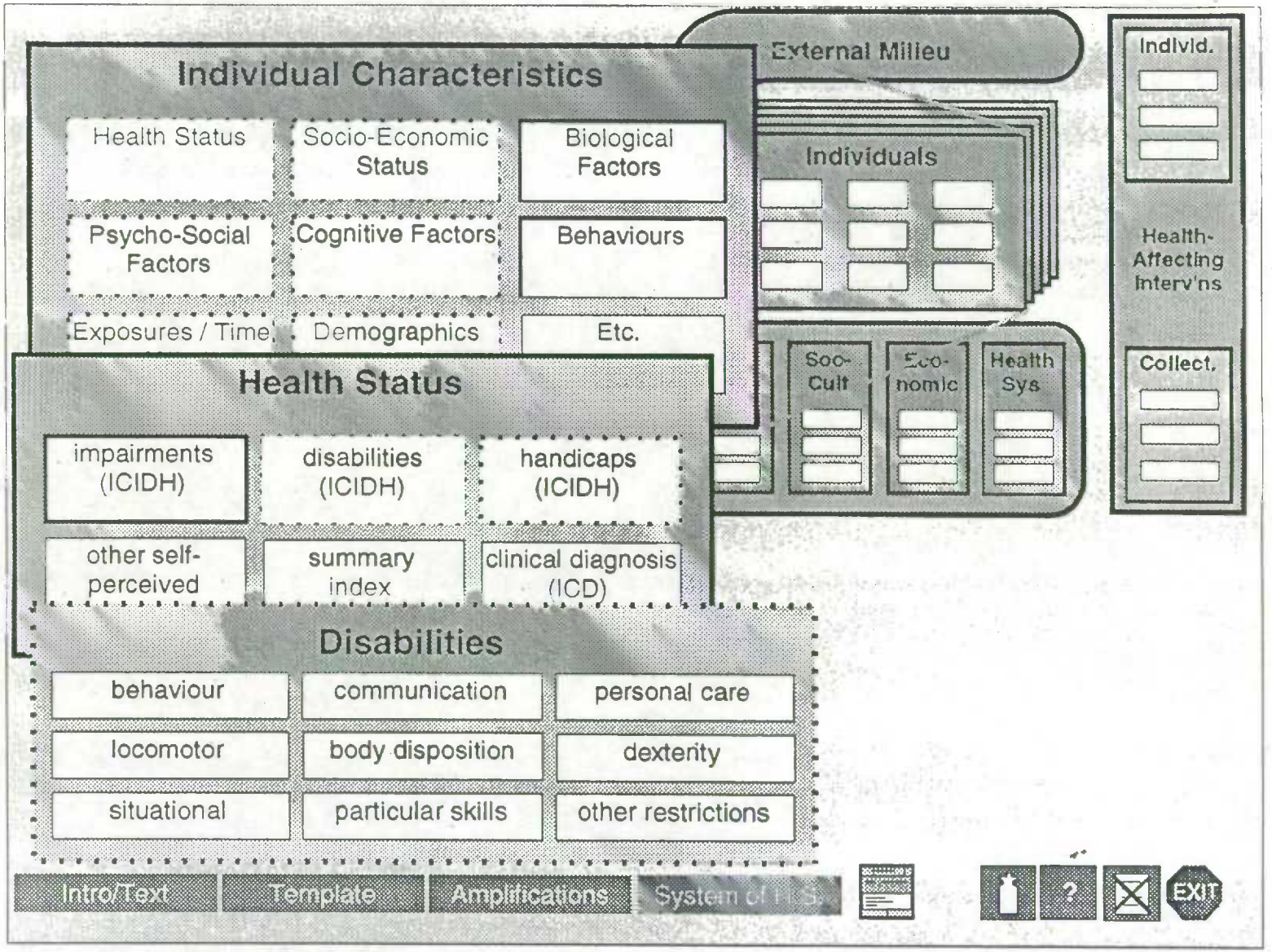
Individual Characteristics
Health Status
Socio-Economic Status
Biological Factors
Psycho-Social Factors
Cognitive Factors
Behaviours
Exposures / Time
Demographics
Etc.
External Milieu
Individuals
Soc-Cult
Eco-nomic
Health Sys
Indivld.
Health-Affecting Interv'ns
Collect.
Health Status
impairments (ICIDH)
disabilities (ICIDH)
handicaps (ICIDH)
other self-perceived
summary index
clinical diagnosis (ICD)
Disabilities
behaviour
communication
personal care
locomotor
body disposition
dexterity
situational
particular skills
other restrictions
Intro/Text
Template
Amplifications
System of H.S.
?
EXIT

Figure 2

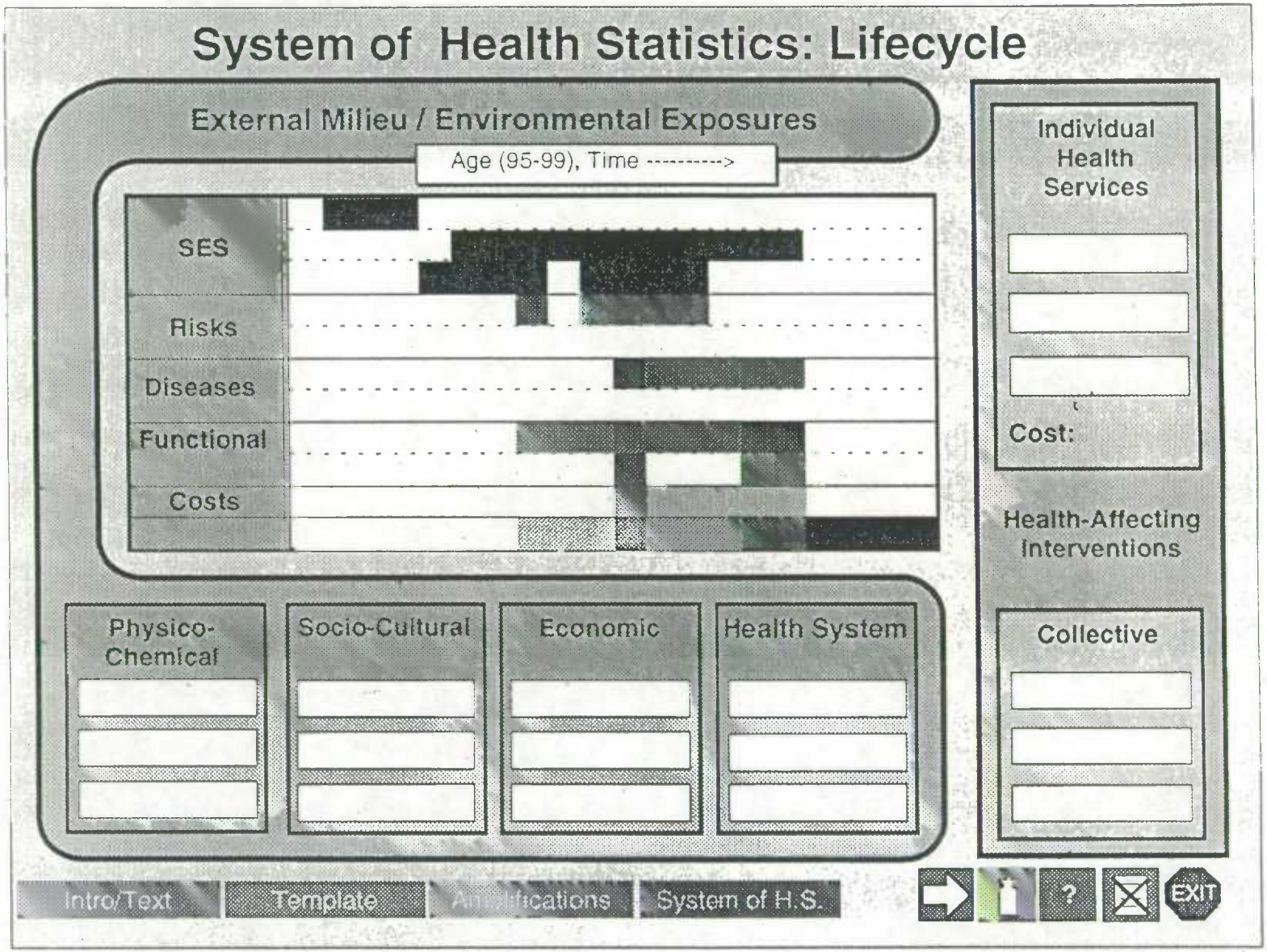
System of Health Statistics: Lifecycle
External Milieu / Environmental Exposures
Age (95-99), Time ---------->
SES
Risks
Diseases
Functional
Costs
Physico-Chemical
Socio-Cultural
Economic
Health System
Individual Health Services
Cost:
Health-Affecting Interventions
Collective
Intro/Text
Template
Amplifications
System of H.S.
?
EXIT

Figure 3a

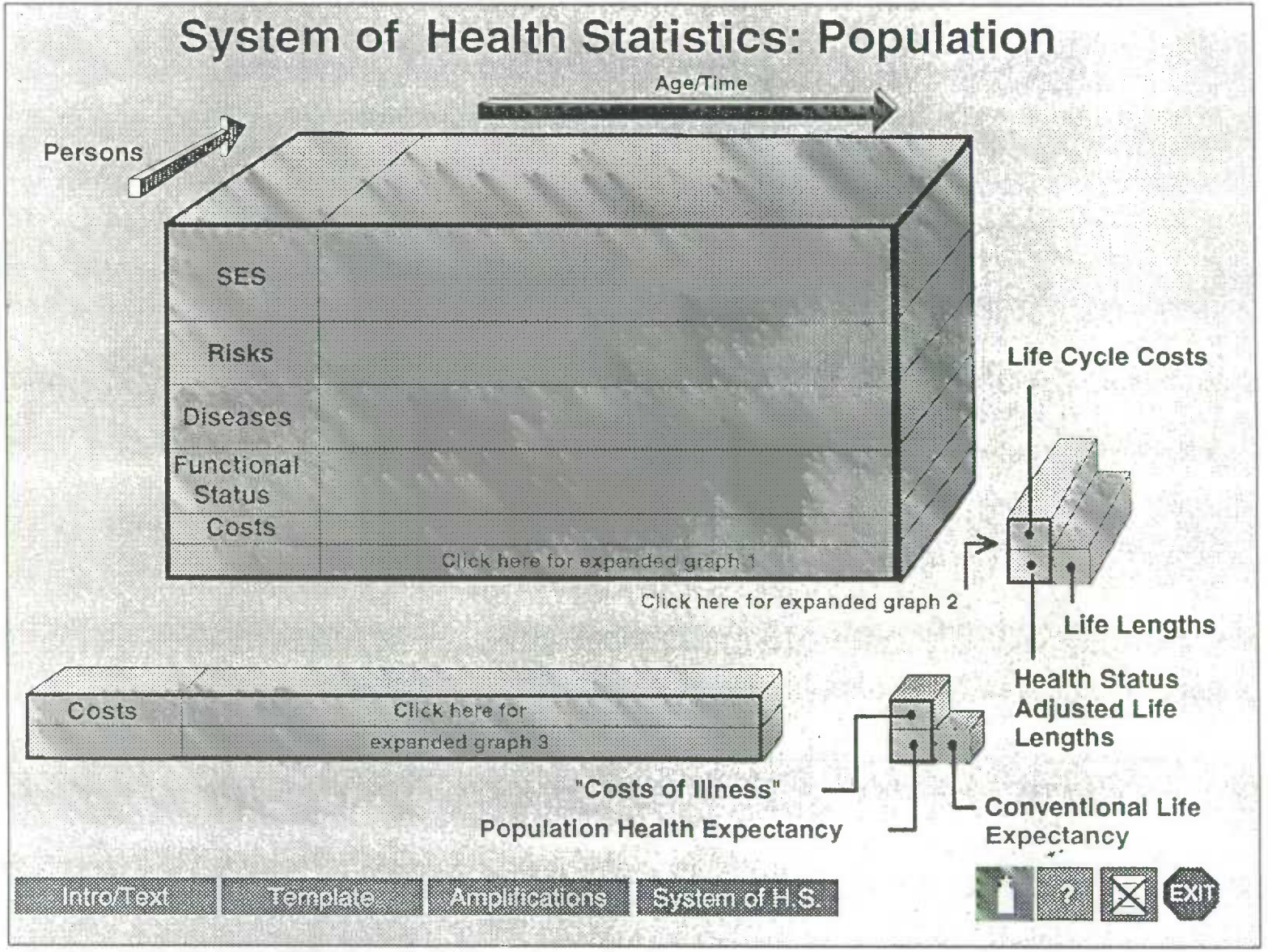
System of Health Statistics: Population
Age/Time
Persons
SES
Risks
Diseases
Functional Status
Costs
Click here for expanded graph 1
Click here for expanded graph 2
Life Cycle Costs
Life Lengths
Health Status Adjusted Life Lengths
Costs
Click here for expanded graph 3
"Costs of Illness"
Population Health Expectancy
Conventional Life Expectancy
Intro/Text
Template
Amplifications
System of H.S.
?
EXIT

Figure 3b

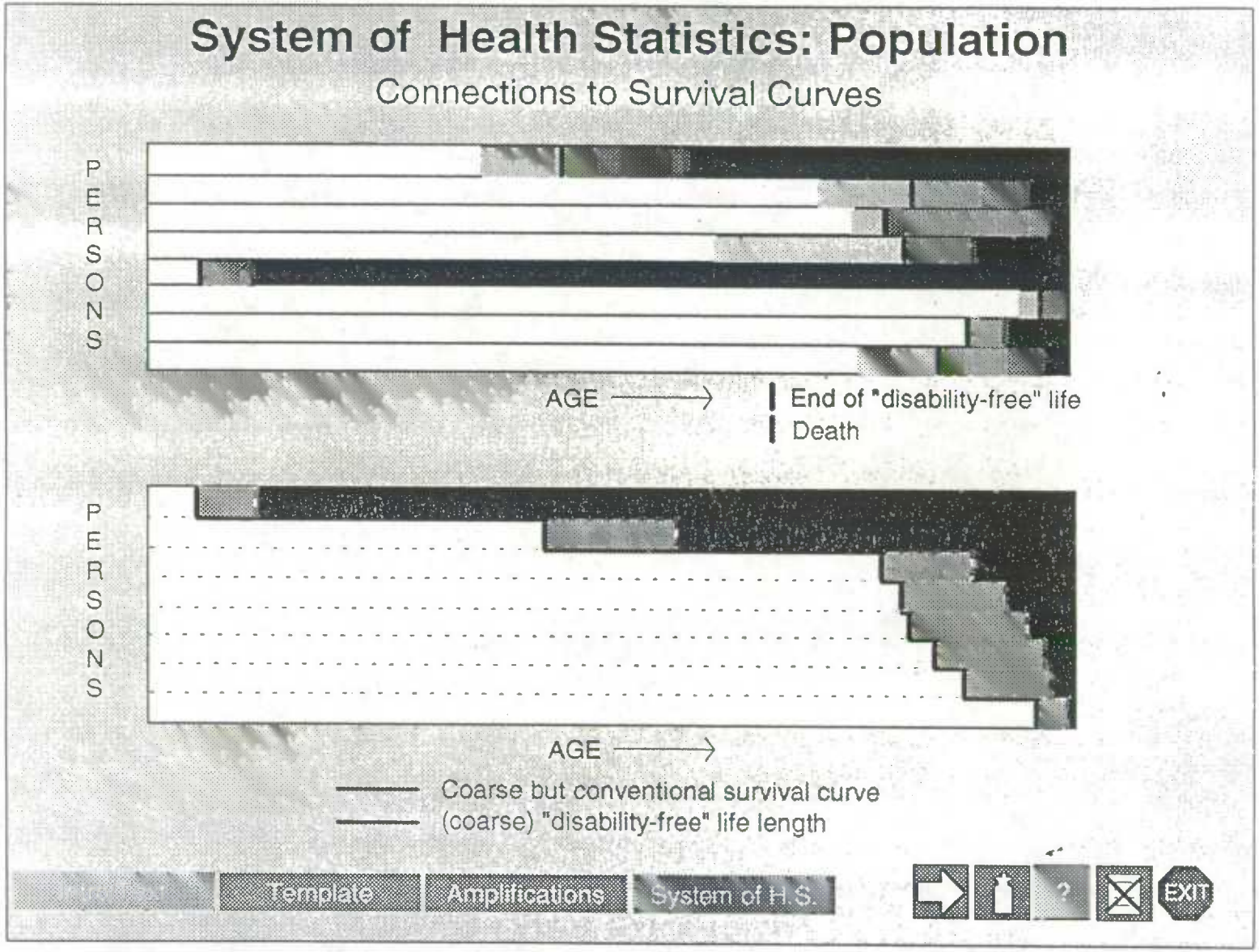
System of Health Statistics: Population
Connections to Survival Curves
PERSONS
AGE
End of "disability-free" life
Death
PERSONS
AGE
Coarse but conventional survival curve
(coarse) "disability-free" life length
Template
Amplifications
System of H.S.
EXIT

Figure 3c

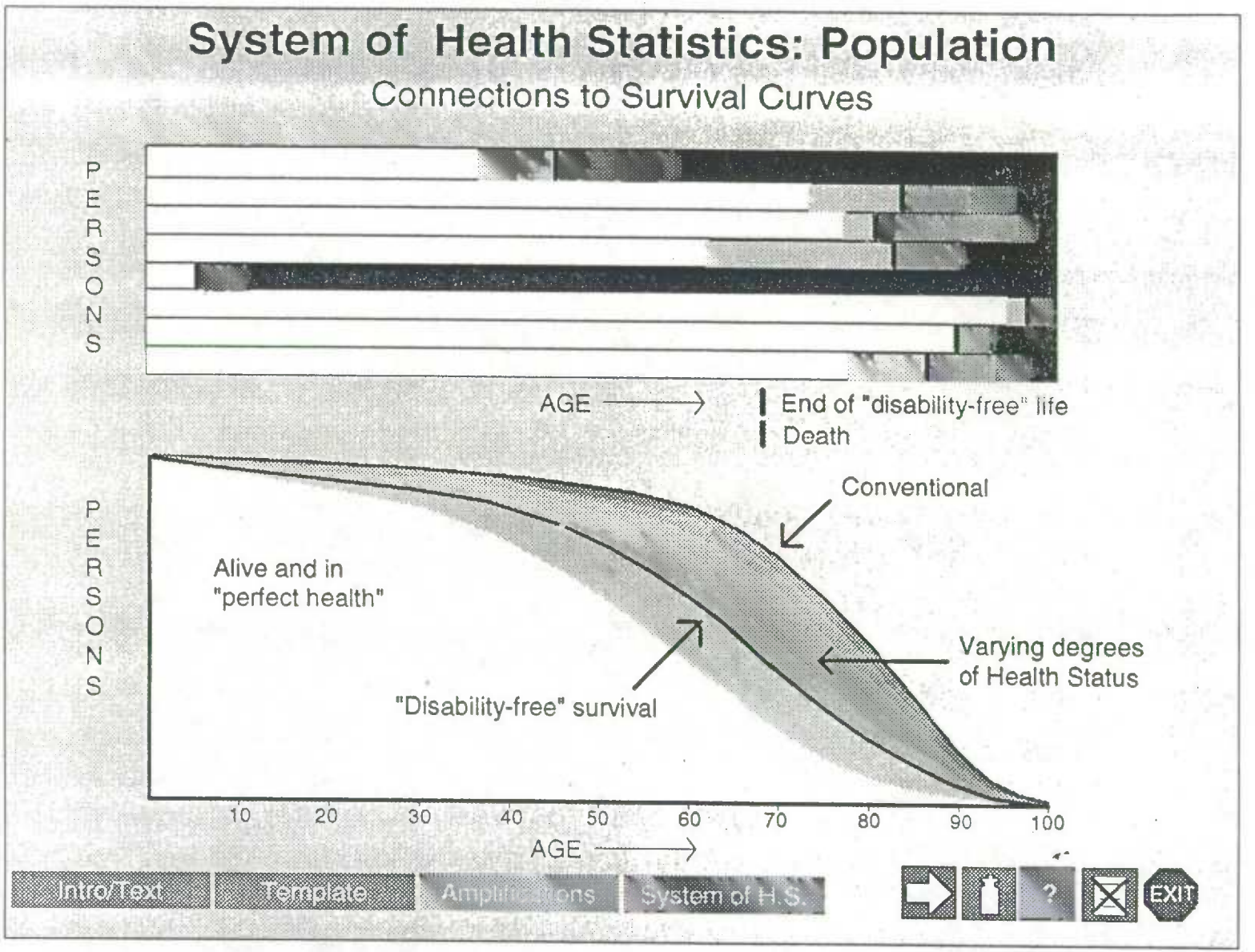
System of Health Statistics: Population
Connections to Survival Curves
PERSONS
AGE
End of "disability-free" life
Death
PERSONS
Conventional
Alive and in
"perfect health"
Varying degrees
of Health Status
"Disability-free" survival
10
20
30
40
50
60
70
80
90
100
AGE
Intro/Text
Template
Amplifications
System of H.S.
?
EXIT

Figure 4a

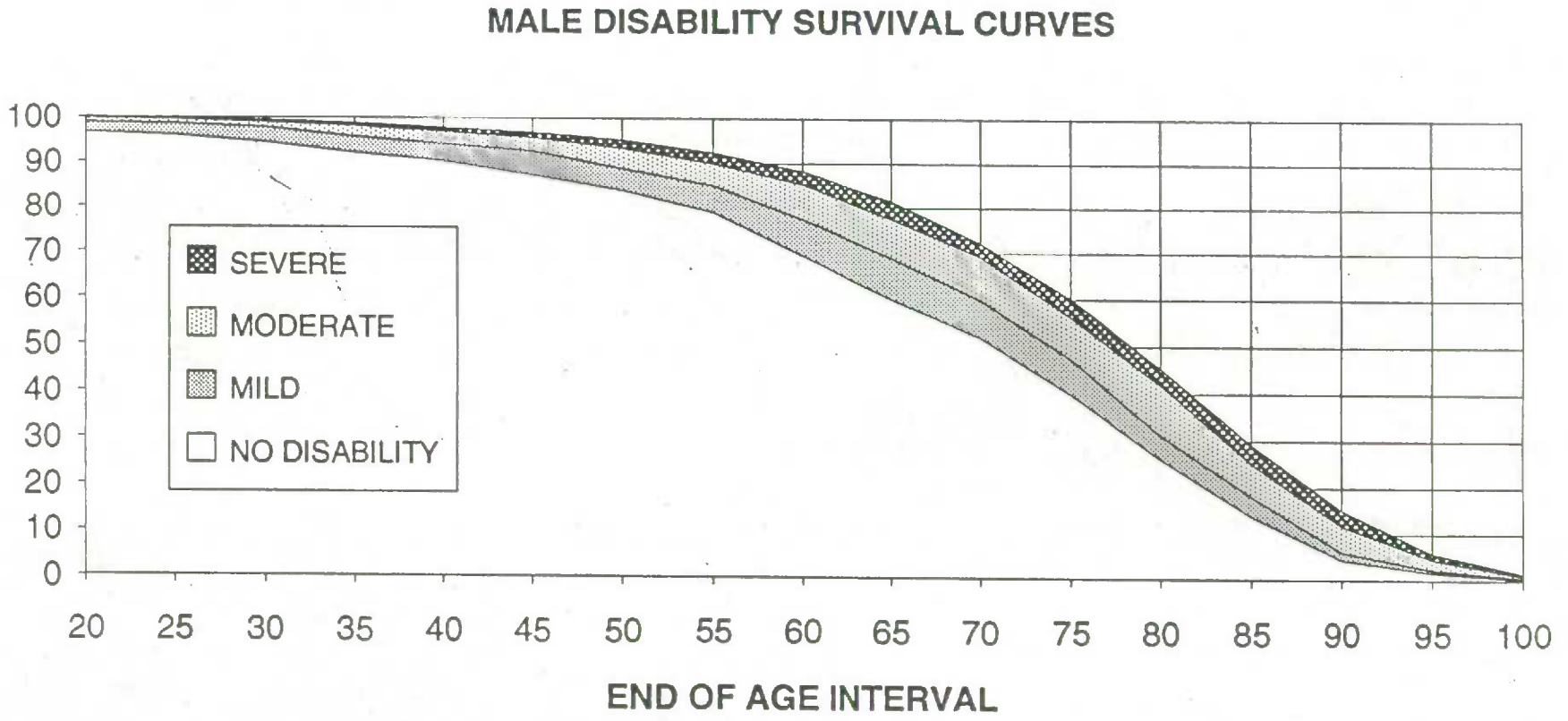
MALE DISABILITY SURVIVAL CURVES
100
90
80
70
60
50
40
30
20
10
0
SEVERE
MODERATE
MILD
NO DISABILITY
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
END OF AGE INTERVAL

Figure 18

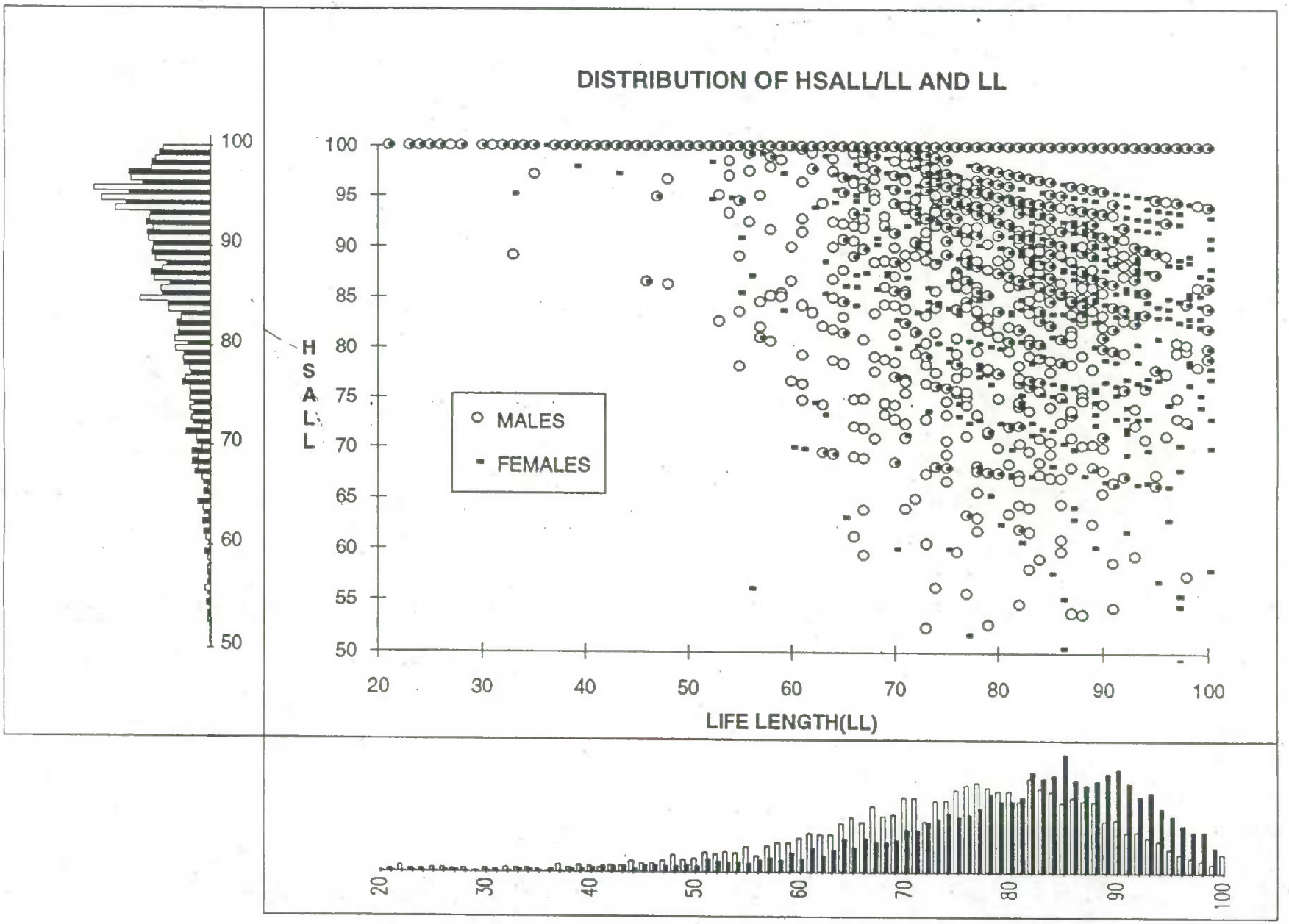
DISTRIBUTION OF HSALL/LL AND LL
HSALL
LIFE LENGTH(LL)
MALES
FEMALES
100
95
90
85
80
75
70
65
60
55
50
20
30
40
50
60
70
80
90
100

Figure 42

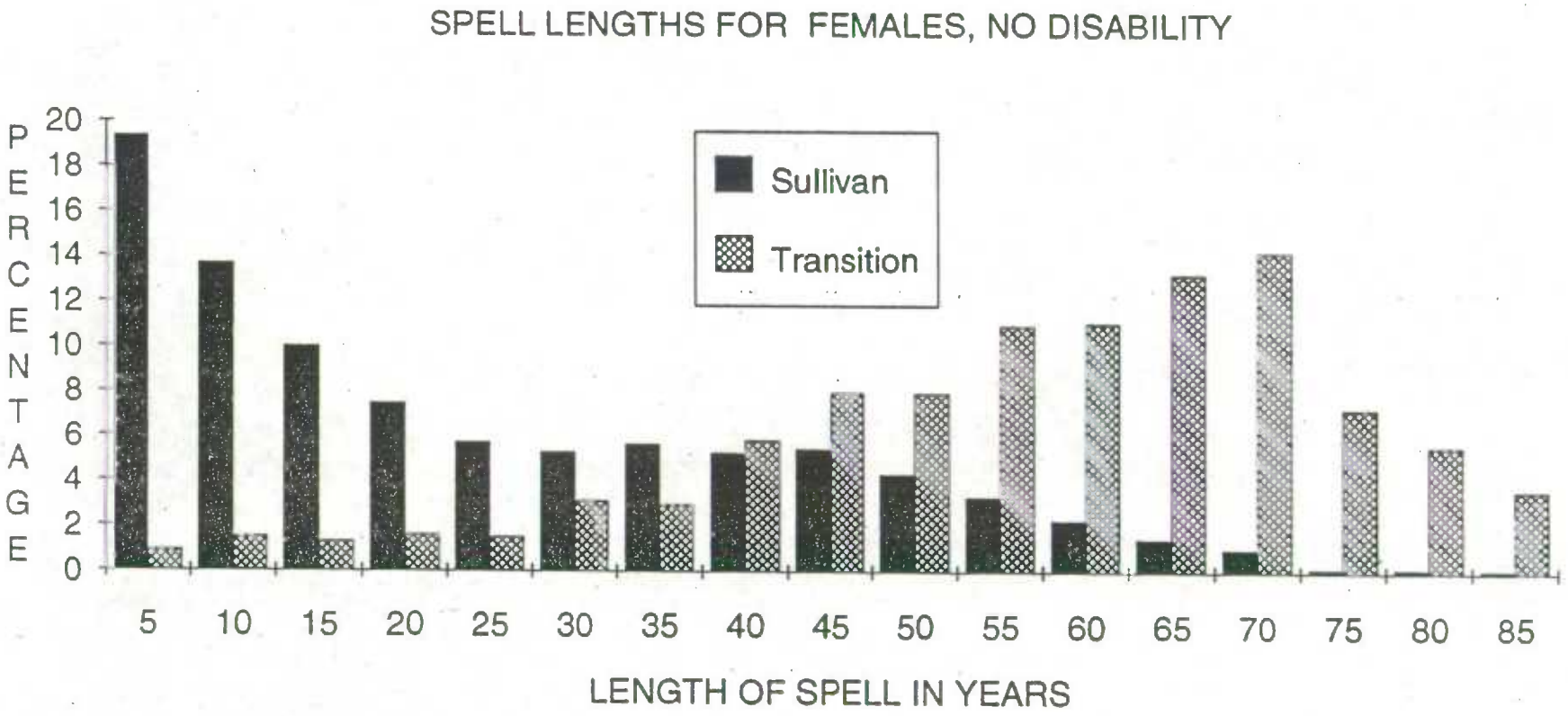
SPELL LENGTHS FOR FEMALES, NO DISABILITY
PERCENTAGE
20
18
16
14
12
10
8
6
4
2
0
Sullivan
Transition
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
LENGTH OF SPELL IN YEARS

Figure 5

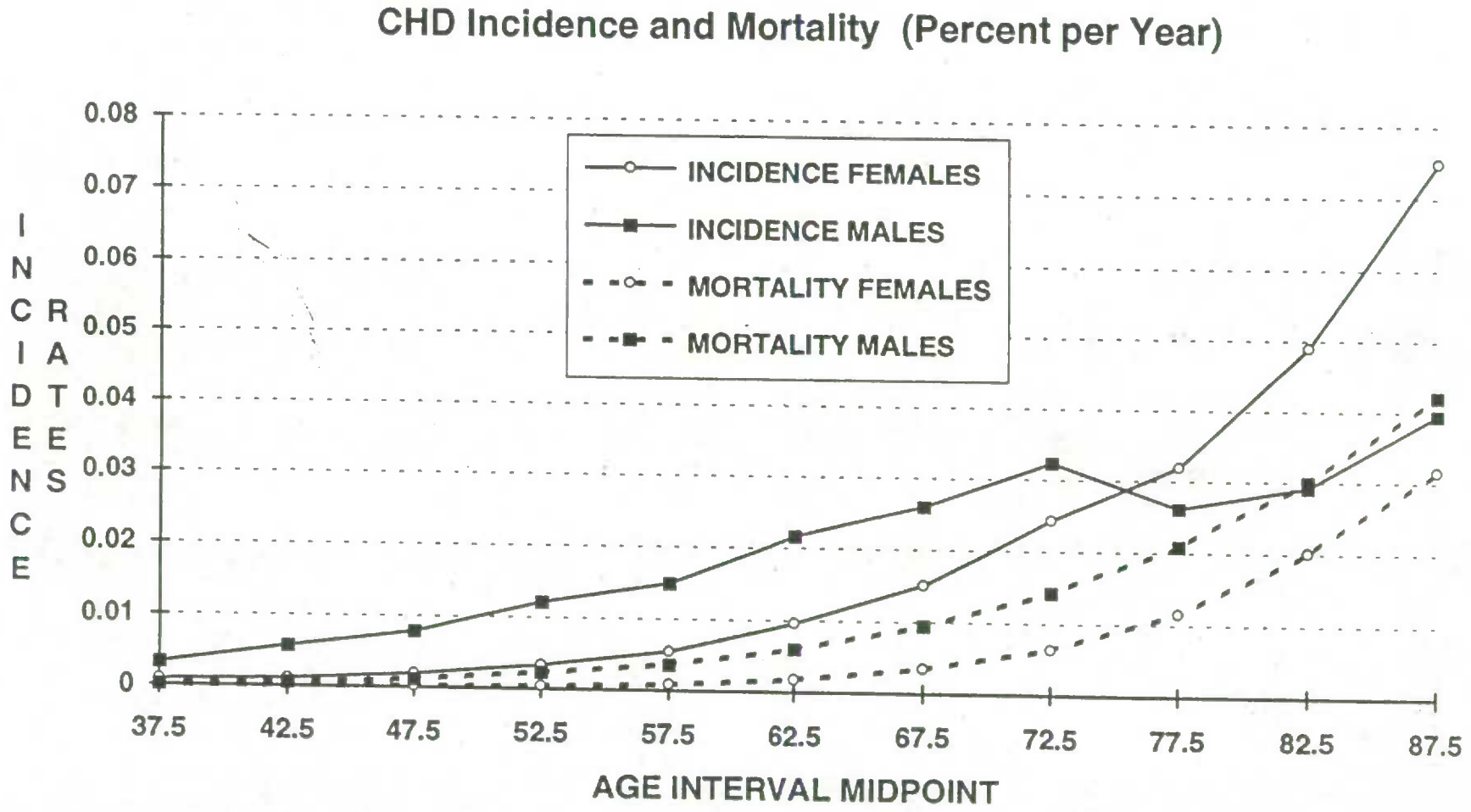
CHD Incidence and Mortality (Percent per Year)
INCIDENCE RATES
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
INCIDENCE FEMALES
INCIDENCE MALES
MORTALITY FEMALES
MORTALITY MALES
37.5
42.5
47.5
52.5
57.5
62.5
67.5
72.5
77.5
82.5
87.5
AGE INTERVAL MIDPOINT

Table 1. Lung Cancer Incidence and Progression Under Three Scenarios

	Modeling Scenario		
	Mortality Only	Incidence and Survival	Relative Risk and Incidence and Survival
FEMALES			
Lifetime Incidence	3.46 (.013)	3.30 (.021)	3.28 (.026)
Lung Cancer Cases			
Average Age at Diagnosis	68.41 (.049)	67.90 (.062)	67.79 (.096)
Average Age at Death from Lung Cancer	68.41 (.049)	68.11 (.085)	68.14 (.092)
Case Fatality (%)	100.00	83.22 (.258)	84.93 (.220)
Overall Life Expectancy	79.58 (.010)	79.58 (.012)	79.56 (.024)
MALES			
Lifetime Incidence (%)	8.62 (.025)	9.41 (.026)	9.27 (.058)
Lung Cancer Cases			
Average Age at Diagnosis	69.06 (.020)	68.76 (.056)	68.32 (.035)
Average Age at Death from Lung Cancer	69.06 (.020)	69.24 (.056)	68.70 (.029)
Case Fatality (%)	100.00	81.06 (.212)	81.29 (.129)
Overall Life Expectancy	73.23 (.009)	73.33 (.026)	73.36 (.020)

Source: POHEM simulations, sample size = 100,000 cases for first scenario, 50,000 cases otherwise.

Standard errors in parentheses based on independent sub-samples of 5,000 cases.

ANALYTICAL STUDIES BRANCH
RESEARCH PAPER SERIES

No.

1. *Behavioural Response in the Context of Socio-Economic Microanalytic Simulation,* ***Lars Osberg***

2. *Unemployment and Training,* ***Garnett Picot***

3. *Homemaker Pensions and Lifetime Redistribution,* ***Michael Wolfson***

4. *Modelling the Lifetime Employment Patterns of Canadians,* ***Garnett Picot***

5. *Job Loss and Labour Market Adjustment in the Canadian Economy,* ***Garnett Picot and Ted Wannell***

6. *A System of Health Statistics: Toward a New Conceptual Framework for Integrating Health Data,* ***Michael C. Wolfson***

7. *A Prototype Micro-Macro Link for the Canadian Household Sector,* ***Hans J. Adler and Michael C. Wolfson***

8. *Notes on Corporate Concentration and Canada's Income Tax,* ***Michael C. Wolfson***

9. *The Expanding Middle: Some Canadian Evidence on the Deskilling Debate,* ***John Myles***

10. *The Rise of the Conglomerate Economy,* ***Jorge Niosi***

11. *Energy Analysis of canadian External Trade: 1971 and 1976,* ***K.E. Hamilton***

12. *Net and Gross Rates of Land Concentration,* ***Ray D. Bollman and Philip Ehrensaft***

13. *Cause-Deleted Life Tables for Canada (1972 to 1981): An Approach Towards Analyzing Epidemiologic Transition,* ***Dhruva Nagnur and Michael Nagrodski***

14. *The Distribution of the Frequency of Occurence of Nucleotide Subsequences, Based on Their Overlap Capability,* ***Jane F. Gentleman and Ronald C. Mullin***

15. *Immigration and the Ethnolinguistic Character of Canada and Quebec,* ***Réjean Lachapelle***

16. *Integration of Canadian Farm and Off-Farm Markets and the Off-Farm Work of Women, Men and Children,* ***Ray D. Bollman and Pamela Smith***

17. *Wages and Jobs in the 1980s: Changing Youth Wages and the Declining Middle,* ***J. Myles, G. Picot and T. Wannell***

18. *A Profile of Farmers with Computers,* ***Ray D. Bollman***

19. *Mortality Risk Distributions: A Life Table Analysis,* ***Geoff Rowe***

20. *Industrial Classification in the Canadian Census of Manufactures: Automated Verification Using Product Data,* ***John S. Crysdale***

21. *Consumption, Income and Retirement,* ***A.L. Robb and J.B. Burbridge***

22. *Job Turnover in Canada's Manufacturing Sector,* ***John R. Baldwin and Paul K. Gorecki***

23. *Series on The Dynamics of the Competitive Process,* ***John R. Baldwin and Paul K. Gorecki***

    A. *Firm Entry and Exit Within the Canadian Manufacturing Sector.*
    B. *Intra-Industry Mobility in the Canadian Manufacturing Sector.*
    C. *Measuring Entry and Exit in Canadian Manufacturing: Methodology.*
    D. *The Contribution of the Competitive Process to Productivity Growth: The Role of Firm and Plant Turnover.*
    E. *Mergers and the Competitive Process.*
    F. *(in preparation)*
    G. *Concentration Statistics as Predictors of the Intensity of Competition.*
    H. *The Relationship Between Mobility and Concentration for the Canadian Manufacturing Sector.*

24. *Mainframe SAS Enhancements in Support of Exploratory Data Analysis,* ***Richard Johnson and Jane F. Gentleman***

25. *Dimensions of Labour Market Change in Canada: Intersectoral Shifts, Job and Worker Turnover,* ***John R. Baldwin and Paul K. Gorecki***

26. *The Persistent Gap: Exploring the Earnings Differential Between Recent Male and Female Postsecondary Graduates,* ***Ted Wannell***

27. *Estimating Agricultural Soil Erosion Losses From Census of Agriculture Crop Coverage Data,* ***Douglas F. Trant***

28. *Good Jobs/Bad Jobs and the Declining Middle: 1967-1986,* ***Garnett Picot, John Myles, Ted Wannell***

29. *Longitudinal Career Data for Selected Cohorts of Men and Women in the Public Service, 1978-1987,* ***Garnett Picot and Ted Wannell***

30. *Earnings and Death - Effects Over a Quarter Century,* ***Michael Wolfson, Geoff Rowe, Jane F. Gentleman adn Monica Tomiak***

31. *Firm Response to Price Uncertainty: Tripartite Stabilization and the Western Canadian Cattle Industry,* ***Theodore M. Horbulyk***

32. *Smoothing Procedures for Simulated Longitudinal Microdata,* ***Jane F. Gentleman, Dale Robertson and Monica Tomiak***

33. *Patterns of Canadian Foreign Direct Investment Abroad,* ***Paul K. Gorecki***

34. *POHEM - A New Approach to the Estimation of Health Status Adjusted Life Expectancy,* ***Michael C. Wolfson***

35. *Canadian Jobs and Firm Size: Do Smaller Firms Pay Less?,* ***René Morissette***

36. *Distinguishing Characteristics of Foreign High Technology Acquisitions in Canada's Manufacturing Sector,* ***John R. Baldwin and Paul K. Gorecki***

37. *Industry Efficiency and Plant Turnover in the Canadian Manufacturing Sector,* ***John R. Baldwin***

38. *When the Baby Boom Grows Old: Impacts on Canada's Public Sector,* ***Brian B. Murphy and Michael C. Wolfson***

39. *Trends in the distribution of Employment by Employer Size: Recent Canadian Evidence,* ***Ted Wannell***

40. *Small Communities in Atlantic Canada: Their Industrial Structure and Labour Market conditions in the Early 1980s,* ***Garnett Picot and John Heath***

41. *The Distribution of Federal/Provincial Taxes and Transfers in rural Canada,* ***Brian B. Murphy***

42. *Foreign Multinational Enterprises and Merger Activity in Canada,* ***John Baldwin and Richard Caves***

43. *Repeat Users of the Unemployment Insurance Program,* ***Miles Corak***

44. *POHEM -- A Framework for Understanding and Modelling the Health of Human Population,* ***Michael C. Wolfson***

45. *A Review of Models of Population Health Expectancy: A Micro-Simulation Perspective,* ***Michael C. Wolfson and Kenneth G. Manton***

46. *Career Earnings and Death: A Longitudinal Analysis of Older Canadian Men,* **Michael C. Wolfson, Geoff Rowe, Jane Gentleman and Monica Tomiak**

47. *Longitudinal Patterns in the Duration of Unemployment Insurance Claims in Canada,* **Miles Corak**

48. *The Dynamics of Firm Turnover and the Competitive Process,* **John Baldwin**

49. *Development of Longitudinal Panel Data from Business Registers: Canadian Experience,* **John Baldwin, Richard Dupuy and William Penner**

50. *The Calculation of Health-Adjusted Life Expectancy for a Multi-Attribute Utility Function: A First Attempt,* **J.-M. Berthelot, R. Roberge and M.C. Wolfson**

51. *Testing The Robustness of Entry Barriers,* **J. R. Baldwin, M. Rafiquzzaman**

52. *Canada's Multinationals: Their Characteristics and Determinants,* **Paul K. Gorecki**

53. *The Persistence of unemployment: How Important were Regional Extended Unemployment Insurance Benefits?* **Miles Corak, Stephen Jones**

54. *Cyclical Variation in the Duration of Unemployment Spells,* **Miles Corak**

55. *Permanent Layoffs and Displaced Workers: Cyclical Sensitivity, Concentration, and Experience Following the Layoff,* **Garnett Picot, Wendy Pyper**

56. *The Duration of Unemployment During Boom and Bust*,* **Miles Corak**

57. *Getting a New Job in 1989-90 in Canada,* **René Morissette**

58. *Linking survey and administrative data to study determinants of health,* **P. David, J.-M. Berthelot and C. Mustard**

59. *Extending Historical Comparability in Industrial Classification,* **John S. Crysdale**

60. *What is Happening to Earnings Inequality in Canada?,* **R. Morissette, J. Myles and G. Picot**

61. *Structural change in the Canadian Manufacturing Sector, (1970-1990),* **John Baldwin and M. Rafiquzzaman**

62. *Unemployment Insurance, Work Disincentives, and the Canadian Labour Market: An Overview*,* **Miles Corak**

63. *Recent Youth Labour Market Experiences in Canada,* ***Gordon Betcherman, René Morissette***

64. *A Comparision of Job Creation and Job Destruction in Canada and the United States,* ***John Baldwin, Timothy Dunne, John Haltiwanger***

65. *What is Happening to Weekly Hours Worked in Canada?,* ***René Morissette and Deborah Sunter***

For further information, contact the Chairperson, Publications Review Committee, Analytical Studies Branch, R.H. Coats Bldg., 24th Floor, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, (613) 951-8213.