POHEM — a framework for understanding and modelling the health of human populations

Michael C. Wolfson

Introduction

The POpulation HEalth Model (POHEM) is both an idea and a computer simulation model. At the level of ideas, it is part of a framework for understanding and thinking about human health. In terms of both software and data, POHEM is a practical implementation of as many of the concepts and ideas as have proven feasible with the resources invested so far.

The genesis of POHEM lay in Statistics Canada's concern that its health statistics programme was weak. This diagnosis was not so much due to problems with the data being gathered by Canada's national statistical agency, but rather to large gaps in information in important areas. The most important information gap, or imbalance in the health statistics programme, concerned information on the health status of the Canadian population. There were far more data on resource inputs to the health care system, and its throughput in terms of visits to doctors and hospital stays, than on health outcomes. Data on individual health status were dominated by causes of death and clinical diagnoses for hospital stays.

Other major statistical concerns were the lack of coherence among the myriad bits of health-related data being collected and proposed, lack of common concepts and definitions across jurisdictions and institutions within Canada, and a narrowness of focus reflecting the dominance of clinical medicine relative to other perspectives such as those of public health, health promotion, and appreciation of the social determinants of health.

These statistical concerns reflect deeper substantive concerns with the directions of Canadian health policy. Traditional (for high-income countries) health (actually illness) care is increasingly seen not only as very expensive, but also as having questionable efficacy. There is no denying a core of modern health care interventions that, from a broader historical perspective, are nothing short of miraculous. The concern, rather, is with a margin perhaps as large as 25% where otherwise useful interventions are applied appropriately, and thus have negligible benefits.

Moreover, even for medical interventions that offer some non-negligible benefit, more stringent requirements are being considered and increasingly demanded as a prerequisite for public funding — namely that the intervention rank highly according to some sort of benefit or cost-benefit criterion (3). In addition to costs, such benefit analysis requires at least 2 major elements. One is a broadly agreed-upon measure of benefit that appropriately reflects impacts of alternative health-affecting interventions on population health status. The other is a body of theory of the determinants of population health such that the implicit "what if" question of cost-benefit analysis can be credibly answered — what would the change in the population's health status be if the specified intervention were funded and implemented?

Finally, there is a sense in Canada that it is time for the pendulum to swing back somewhat from the current emphasis on medical interventions. Increasingly, evidence on the social determinants of health has entered public debate. In a somewhat unholy alliance, there now appears to be greater common cause among fiscal conservatives who wish to cut back on what is seen as a bloated medical establishment, and social progressives who wish to increase funding for programmes in areas like early childhood development and home care for seniors. Irrespective of the merits of these views, they clearly call for a much broader perspective on the field of health information, and on the range of health-affecting interventions to be considered in both a theoretical and a cost-benefit framework.

Our population health model is being developed with these statistical and substantive concerns very much in mind. POHEM began as a conceptual effort, forming a central part of a proposed system of health statistics (4). These efforts within Statistics Canada have also become integrated with a larger review of health information in Canada (5).

---

Statistics Canada and Canadian Institute for Advanced Research (CIAR). 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. I am deeply indebted to my Statistics Canada colleagues for their many contributions, and to the CIAR for providing a unique intellectual milieu. This is a substantially revised version of Ref. (1). I of course remain responsible for any errors or omissions.

The recent experience in the state of Oregon suggests considerable difference of views on the extent to which costs should be considered when ranking health interventions, but clear agreement on the need to measure benefit. See also Ref. (3).
POHEM subsequently moved through a prototype phase, and is now beginning to be used in applied analyses. While much of the value of the model lies in the kinds of analyses it can support, it should be borne in mind that the genesis of POHEM was and continues to be as a core element in a new system of health statistics.

POHEM builds on a foundation of computer-intensive techniques — both in its realization as a simulation model, and in its use of inputs drawn from rich and highly multivariate data. While this may make it sound expensive, it need not be. POHEM runs on standard personal computers. Since it is "upwardly compatible" from multi-state life table styles of analysis, it can build on data already collected for such purposes. For example, it can nest, as a special case, estimates of disability-free life expectancy (DFLE) or disability-adjusted life years (DALYs) /6,7/.

This article begins with the motivation behind POHEM at the conceptual and theoretical level, then describes microsimulation methods generally and the population health model specifically, and concludes with a few brief examples drawn from analyses using POHEM.

The discussion which follows makes reference to conceptual frameworks, statistical systems, and POHEM itself. These terms are closely interrelated but different. A conceptual framework is a set of general ideas. Specific reference will be made to the conceptual framework developed in the Template for Health Information /8/ — both for its ideas and for the pedagogical benefit of some of its graphic images. The Template in turn provides a sketch of the beginnings of a system of health statistics. An actual system of health statistics does not exist in Canada, but is currently the objective of renewed efforts, following the impetus of the recent National Task Force on Health Information /3/. Finally, POHEM is being developed as one part of the activities involved in creating a system of health statistics.

Motivations, concepts and "theories of health"

As noted in the introduction, much more statistical effort is spent measuring the inputs to and throughputs of the health care system than in measuring how healthy the population is. Thus one central motivation for POHEM is the development of population health status measures to help remedy this imbalance.

A desire for coherence is the second motivation. Health certainly rivals the economy in importance, yet in comparison the statistical base is confused, fragmentary and incoherent. The System of National Accounts (SNA) provides a coherent statistical framework for economic information. The coherence of the SNA derives from the fact that its data elements 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 generally does not exist for health data. Coherence in health statistics is beneficial for two reasons. First, at a conceptual level, it aids understanding of the interrelationships among various data elements. Second, at a statistical level, basic arithmetic identities, when combined with redundancy in data sources, provide an ongoing check on data quality (a point that is illustrated later).

In so far as POHEM is part of a larger process of statistical development, it must at least implicitly have some theoretical foundation. Theory and measurement are closely interrelated in all scientific endeavour. For example, theories in astrophysics explicitly guide the development and construction of specific kinds of radio-telescopes, particularly for the strengths and frequencies of signals to be detected. Of course, there is no uni-directional causality here — from concepts and theory to a measurement system. Rather, the history is one of iteration back and forth between theory and concepts on the one hand, and observation and measurement on the other.

In the case of human health, a large portion of current statistical measurement is based on a disease-oriented theory. However, there is enough accumulated evidence on the broad range of determinants of health, and concern about the sequelae of disease processes, to signal the need for major changes in theoretical and conceptual perspective, and therefore changes in approaches to observation and measurement.

To give some flavour to these empirical results and findings, we might mention that, for example, there is widely accepted evidence that serum cholesterol is connected to heart disease — but also to many other important metabolic pathways such as serotonin. Similarly, there is strong evidence that social support and incomes are connected to longevity — including associations with decades-long latencies; and there are specific "windows" in early childhood when key opportunities for physiological development (e.g., vision) are present. Yet other evidence suggests major roles for genetic and myriad environmental factors. These kinds of diverse empirical nuggets all have some superficial connection with one another — for example they are all pertinent to human health.

However, there are no grand theories that knit these diverse empirical findings together; and the piecemeal character of current partial theories tends to generate piecemeal empirical results. There is need at least for a theoretical structure which can hold and index all the piecemeal results, much as a library holds books on diverse topics. Moreover, it would be even better if the "index-
ing" in this library of health-related empirical findings connected the elements in a logical and coherent manner. To give a simple example, in addition to indexing tobacco smoking alphabetically under "T", it could also be connected to lung cancer and heart disease, and it could have connections coming from television advertising and peer pressures.

A theoretical structure or conceptual framework of this sort would not only be broad enough to encompass all the diverse kinds of phenomena associated with human health; it could also play an integrating role, particularly by fostering observations and empirical work that seek to knit together various piecemeal results. Such a conceptual framework was developed as part of the National Task Force on Health Information (5), specifically the Template for Health Information (8) which in turn built on Evans and Stoddart (9).

From a theoretical perspective, the basic structure proposed in the Template is in 2 parts — descriptions of the variables of interest, and descriptions of how they evolve over time. In dynamic systems, these are referred to respectively as the "state space" and the "laws of motion". In human health, the "laws of motion" are generally either unknown or contentious. But the set of variables of interest, the "state space", is more widely accepted. Thus, current knowledge can support a theoretical structure which has a place for all the variables of interest, and provides mechanisms for using or trying out various "laws of motion" as they are proposed and refined through careful research.

**Fig. 1**

Fig. 1 provides an image from the Template (actually a monochrome "still" drawn from the animated graphics in the Template software) that can serve as the basis for describing such a structure — a comprehensive state space for population health.

The image in Fig. 1 divides the health field into 3 broad domains. At the centre of the image (deliberately so) is a representation of an individual's life cycle. The array of cubbyholes is a visual metaphor for a hypothetical individual’s biography of events and states. The shading illustrates an animation sequence for a hypothetical individual’s biography. It starts with a blank matrix of cubbyholes or cells. The cells are then filled in from left to right, representing the evolving biography of the individual as he or she passes through the life cycle. For example, there are events like entering school, getting married, being exposed to risk factors, having a heart attack, partial recovery, but then declining health and eventually dying.

Surrounding but not completely enveloping the individual in Fig. 1 (visually and metaphorically) is the "External Milieu" — a major source of influences on individuals’ health. The external milieu has in turn has been classified into 4 kinds of environments.

The popular media tend to identify "the environment" only with the first of these, the physical—

---

*While the time dimension is explicit for the individual in the centre of Fig. 1, it is only implicit for the external milieu and for health-affecting interventions.*
chemical environment. However, there is strong evidence that socio-cultural and economic exposures, the second and third kinds of environment, are at least as important to human health as risks derived from physical-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).

However, the bulk of society's health-related activity is shown in the vertical bar to the right of the basic Template image. The visual nuance is that while the individual in the centre of the image is surrounded by the "External Milieu", an individual's health is not completely out of his or her control at the whim of various environmental exposures. Health is also influenced by individuals' and society's consciously intended actions — a large collection of "Health-Affecting Interventions". This phrase was deliberately chosen to be broader than "health care", since there is much more that influences the health of individuals in society than opportunities to visit health care providers.

The Template sub-divides this broad domain of interventions into 2 groups. The first, "individual" health-affecting interventions, act on us as individuals typically in one-on-one settings — for example encounters with providers of various health care services like a dental visit or hospitalization for an appendectomy. The second is "collective" health-affecting interventions. These interventions take the form of government programmes and regulations that act on us collectively, although indirectly (and often inadvertently) through the external milieu. Examples include regulations of food quality, tobacco advertising, pharmaceuticals, water quality and particulate emissions, or expenditures on traffic safety and headstart-type programmes.

Returning to the central portion of the image, the individual biography is represented much like the record layout in a computerized longitudinal microdata set. Each row corresponds to a group of variables or fields in the record layout, and thus to a set of state variables in a dynamic system. Each column refers to a time interval with age along the horizontal axis. The main groups of variables describing this hypothetical individual's stylized biography are:

- socio-economic status (SES) attributes — variables known to have strong statistical and likely causal associations with health;
- various risk factors including genetic predispositions, physical factors (e.g., serum cholesterol, obesity, hypertension), lifestyle factors (e.g., smoking), and risks deriving from environmental exposures (including noxious physical-chemical agents, and socio-cultural factors such as the availability of social support);
- clinically-defined diseases like ischaemic heart disease, lung cancer, osteoarthritis and Alzheimer's (classified by the International Classification of Diseases-ICD);
- vernacular health problems — such as not having sufficient lower limb or cognitive capacity to get around without a wheelchair, and chronic pain (classified by the impairment and disability portions of the International Classification of Impairments, Disabilities and Handicaps-ICIDH);
- direct costs of health services used; and
- finally the bottom row, as well as the "bottom line" — a summary health status value — a number between zero = dead (shown as black in the figure) and one = fully healthy (shown as white) summarizing each individual's overall health status for the year.

Note that environmental exposures (i.e., contacts with the external milieu) are largely implicit in risk factor exposures, the evolution of SES attributes like income from paid work, and in the case of handicap the interaction between vernacular health problems (i.e., impairments and disabilities) like being wheelchair-bound, and the physical and social environment. Also, handicap in the sense of the ICIDH is not considered an intrinsic individual attribute. Rather it is captured or "expressed" in the individual's interactions with the external milieu.

The dynamics or laws of motion for these state variables in an individual's biography are not explicit in this image. All that is shown is the top level of a classification structure. However, descriptions of the dynamics of various health-related processes are an essential part of the "theory" and health information that should be encompassed by this framework. For example, at school age, the socio-cultural environment of school, family and television may influence attitudes toward healthy behaviours. Later, these behaviours may influence diet, which then over many years may influence susceptibility to heart disease. One well-known marker could be serum cholesterol levels, but other factors like chronic stress may also be significant to the individual's health history. The Template image is agnostic about which specific causal stories are correct. The key point is that it provides a framework within which such stories, or better still scientific evidence if it is available, can be codified, quantified, and accumulated.

---

\* To extend the analogy with dynamic systems to the ideas of mathematical control theory, health-affecting interventions constitute the "control variables" for influencing the trajectories of individuals both directly, and indirectly via the trajectories of the environments of the external milieu.

\* Indirect economic costs like foregone earnings can also be considered, but they must be measured by a simulated comparison of earnings to a hypothetical scenario where a given disease or health problem is "deleted".

Rapp. trimest. stat1st. sanit. mond., 47 (1994)
For example, the shading in Fig. 1 indicates (illustratively) the levels of scalar values for a set of individual attributes over age intervals in the individual's life course. The dynamics representing various theories or knowledge relating to health then consist of those rules, mapping systems or formulae that allow any 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 on the individual's history to the left, plus any relevant exposures or influences from the external milieu and health-affecting interventions.

In reality, these dynamic processes involve a heterogeneous population. Thus, even moderately realistic statistical descriptions will likely be complex, multi-level and multivariate. They will inevitably have substantial stochastic components, not least to reflect myriad factors that have not been explicitly included in the analysis. The (various and partial) theories comprehended by the conceptual framework are embodied in these explicit descriptions of the dynamics of the various individual attributes.

For example, the fertility "theory" that underlies many 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-specific fertility rates. Of course, this is a very simple theory of the determinants of fertility. Other variables like educational attainment, marital status, prior birth spacing, ethnicity and labour market experience are omitted even though analysis of appropriately rich data would likely show them to be significant. Knowledge in a case like this is limited by the availability of multivariate longitudinal microdata.

The framework in Fig. 1 is open in this regard. As long as independent or determining variables like educational attainment are part of the state space, they can be incorporated as inputs to a (recursive) 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 heart disease; there is continuing debate in the epidemiological literature about the magnitudes of these effects, the form of interactions if any, and the range of other variables that are also significant determinants of coronary events. The framework in Fig. 1 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.

More broadly, there is clear evidence that the socio-cultural milieu and economic circumstances have a profound effect on health (10-13). 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 (14). Again, the framework in Fig. 1 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 as a well-defined function of a set of individual level variables (e.g., the neighbourhood's poverty rate) or are separately collected (e.g., ambient air pollution levels or crime rates).

**General role of microsimulation**

The discussion so far has set out our appreciation of the basic situation with regard to statistical information in the health field, and has sketched a broad conceptual framework based on the Template. In turn, this conceptual framework can guide the construction of a system of health statistics which both reflects current theory and understanding, and supports further development of new theory and understanding. This conceptual framework and proposed system of health statistics also serve as the basis for a theoretical structure expressed as a computer simulation model. This idea is developed in 3 stages. First, we describe the utility of computer simulation models as one central aspect of a statistical system like that proposed for health information. Second, the importance of microsimulation modelling methods is developed. Then, in a later section, the current version of POHEM as one specific instance of a microsimulation model is described, along with the analyses which it supports.

One basic reason for reliance on simulation in the health area is to "observe" life expectancy (LE). LE is a synthetic indicator that requires observation (populations at risk and mortality rates) plus the simulated answer to the specific "what if" question, "How long would a birth cohort live on average if everyone were exposed to a given series of age-specific mortality rates?" In turn, "statistics" that are generalizations of LE also require numerical simulation to be "observed". More concretely, we are interested in a family of generalizations of LE, including "cause-deleted" LE, as well as individuals' (in a given population) expected durations in various health states. Our premise is that such indicators are fundamental to creating good summary population health status statistics, and thereby remedying the imbalance of current health statistics. As a result, some sort of simulation modelling structure is an essential component of a system of health statistics.
Simulation models are also of central strategic importance because they give coherence. Without some sort of integrating analytical framework like the network of arithmetic identities in the System of National Accounts, or a simulation model, data series risk being a hodgepodge, as is the case currently in the health area. Turning this point around, it is relatively easy to consult with various data-using constituencies and then compile a wish list of needed data, as was done in the course of the Task Force consultations (5). But if the wish list is at least partly framed by, or can be mapped into a coherent structure like a simulation model, the whole may well be greater than the sum of its parts. The various constituent data series will be of interest in and of themselves. They can also be “indexed” by the overall conceptual framework (e.g., the classification structure embodied in the Template), by having each kind of data associated with a specific element or part of the framework. But the real power of coherent data will only be fully realized when combined in an explicit quantitative simulation model.

There are 2 main reasons. First, simulations allow the posing and answering of rigorously constructed “what if” questions. Exploration of these kinds of hypothetical scenarios is fundamental to decision making and basic research. Second, with proper designed-in redundancy in the data feeder systems, simulation can enhance the generation of “data confrontations”. These are situations where the same concept or number can be estimated in 2 different ways. In principle, the results should be identical. However, they are typically different (as in the “residual error of estimate” in the System of National Accounts). Such “confrontations” serve as an invaluable check on statistical error, particularly non-sampling error which is far more difficult to assess (15). This kind of data confrontation is illustrated in the case of lung cancer data in the penultimate section on POHEM applications below.

The utility of microsimulation follows essentially from the heterogeneity of individuals and their behaviours. 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 — and in economics and demography for that matter. A microanalytic approach is not only computationally feasible, but also needed to begin reflecting realistic population heterogeneity. It is also needed to provide the common foundation for the generalizations of LE to be described below.

The data on individuals envisioned in Fig. 1 above comprise an ideal microdata set. Unfortunately, these data cannot be observed directly. The implied longitudinal household survey is impractical — not least for reasons of respondent burden, privacy and confidentiality concerns, and the century or so 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 behaviour, and then make extrapolations and predictions about their implications.

In this context, microsimulation modelling is a methodology for synthesizing a cohort of the requisite biographies using more practical, albeit fragmentary, data sources. Microsimulation modelling 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 woven together into a set of realistic — but synthetic — individual life cycle histories. We sketch the details of this synthesizing process below. It must be supported by a large effort not only of meta-analysis, as increasingly undertaken in the epidemiologic literature, but more generally a process of meta-synthesis. In meta-analysis, a number of cohort studies considering the same phenomenon, say the relationship between serum cholesterol and heart disease events, are examined and efforts are made to pool the results so as to increase the effective sample size. Microsimulation modelling (and large scale modelling generally) 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 Fig. 1. It is also a simulation model capable of constructing hypothetical alternative versions of this data set where one or more factors have been changed. A much more familiar example of this kind of hypothetical simulation experiment is cause-deleted LE — how long could we expect to live if there were no mortality at all from a particular kind of cancer, for example. Microsimulation modelling is able to construct generalizations of this kind of indicator, and to base the computations on explicit models of

---

6 Note that these kinds of 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 heart disease, chronic obstructive pulmonary disease, etc. Such underlying factors causing elevated risk are completely ignored in the usual cause-deleted life table estimates.
the complex interactions among risk factors and co-morbidities.

In addition, by using microsimulation modelling methods, the underlying (synthetic) longitudinal population data are always available. Thus, other kinds of related statistics can be readily computed. Moreover, these synthetic individual biographies are always available for inspection, in turn allowing the plausibility or "face validity" of the simulation results to be assessed in much greater detail. This is not possible with conventional (partially aggregated or cell-based) multi-state life table methods, which invisibly yield life paths which are highly implausible.

The hypothetical alternative versions of the basic longitudinal microdata set of Fig. 1 generated by microsimulation modelling can provide the basis for policy analysis and decision making. For example, cause-deleted LEs have often been used to form a "league table" of the most "important" diseases — with heart disease in first place, and cancers second, since they typically account for the largest expectations of years of life lost in high-income countries. This ranking in turn influences the allocation of resources for health care and health research. However, this is not the best indicator for setting priorities for resource allocation.

First, the ranking measure should account for more than mortality effects, which are all that affect increments in LE due to the hypothesized elimination of a given disease. It should also take account of morbidity and disability while alive. The core idea is to adjust measures based on LE for how ill or unhealthy people are year by year. A wide variety of health-status-adjusted life expectancy measures have been proposed. For convenience and simplicity, we shall refer to such adjusted LE measures as healthy life expectancy (HLE). One consequence of adopting such adjustments is that in league tables based on cause-deleted HLEs rather than LEs, chronic disabling diseases which are not typically fatal like arthritis and dementias will rank as much more important.

Second, our thinking should be broadened so that diseases are not the only category of health-related phenomenon that can affect HLE. For example, tobacco smoking behaviour has just as much claim to be considered a "cause" in cause-deleted HLE as conventional diseases classified by the ICD. Generally, the capacity to treat as wide a range as possible of risk factors and health-affecting interventions as "causes" is desirable in generating estimates of cause-deleted HLEs. Then, not only would arthritis and dementia rise in the resulting league table ranking, so too would smoking, and perhaps chronic stress. Moreover, health care interventions like coronary artery bypass graft surgery could also be placed in the same ranking. Indeed, items currently outside the usual health care discourse like "neighbourhood cohesiveness" might enter the league table. Of course, this is all contingent on having plausible causal stories represented algorithmically, for example by formally-estimated stochastic processes.

In effect, HLE thus becomes the common standard for measuring population health status. It is also a simple summary index of population health whose regular publication would meet the basic concern expressed at the outset regarding imbalance in health statistics.

Microsimulation modelling can play a role beyond policy analysis and the construction of summary indices. It is also significant for basic research and the development of statistical priorities. Having microsimulation model-based indicators 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. A microsimulation model is built to estimate HLE. In a number of areas, data are very limited or even nonexistent. In these areas, "guessestimates" are made, and then a sensitivity analysis is undertaken to see which of the various guessestimates is most important to estimates of HLE. The result should be increased priority to the collection of data relevant to the most important guessestimate.

Even if such sensitivity analysis to various guessestimates does not feed back immediately to data-collection priorities, the process has value to basic social science research. Much of this research consists of conjectures about possible causal pathways and their magnitudes. For example, how important is unobserved heterogeneity with respect to some kind of innate "frailty" (16); 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 currently know how to collect the data (e.g., innate frailty), or because it is
logically impossible (dependent competing risks (17)). The alternative is to use a numerical simulation model.

**How to microsimulate**

Such is the conceptual background for a proposed system of health statistics, in which numerical microsimulation models can play a key role. A microsimulation model can serve as a repository for some of the requisite information, and provide a method for constructing of a family of HLE-based indicators. In so doing, the microsimulation model "solves" or draws out the implications of the latest observations of the population's status, using empirically-based inferences for the various "laws of motion".

In this section we illustrate the microsimulation modelling method using a simple life table example. The illustration appeals to the structure shown in Fig. 1. The rows in the individual's biography 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. It is also implicit that there is a series of "slices" or planes going back into the page representing a population of individuals.

Microsimulation (at least in a discrete time version) synthesizes these data sets in a series of 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 an 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.

Life table analysis can be nested as a special case of microsimulation modelling. It is not as efficient computationally to use microsimulation model instead of life table methods where the latter are feasible, and the only desired results are summary statistics like LE. However, microsimulation modelling is much more readily generalized than life table analysis. A pertinent example is the estimation of disability-free life expectancy (DFLE) by increment-decrement multi-state life tables (18). In microsimulation modelling terms, this life table analysis corresponds to a very much simplified version of the biographical record layout in Fig. 1. At most, 3 rows are needed from the image — one for alive or dead, one for healthy or disabled, and one for the health status value of each life-year. In other words, an individual's state space in each year of life consists of a three-tuple or a three element column vector.

These three-tuples are simulated using microsimulation one individual at a time, starting at birth and moving from left to right in the sense of Fig. 1 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.

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 is simulated to die, or advance one year in age, possibly becoming disabled or getting better. To represent this (discrete state, discrete time) dynamic process, assume that an individual's biography has been synthesized up to age \( a-1 \). Empirical observations provide the basis for 2 sets of transition probabilities. These are looked up from the model's input data — tables of \( m(i,a) \) 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.\(^{k}\) We thus have completely synthesized a population of complete life-cycle biographies with 2 of the 3 elements in the state space — alive/dead and healthy/disabled. Lastly,

---

\(^{3}\) The order of the 2 outermost loops can be reversed. In other words, the simulation could proceed individual by individual within a year 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).

\(^{k}\) 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.
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.

In this example, using a microsimulation model 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 LE. The same summing and averaging applied to the third health status row gives DFLE. As shown in Wolfson & Manton, a very wide range of models of disability, risk factor, and disease processes can be expressed in terms of the state space of Fig. 1 combined with Monte Carlo microsimulation. Microsimulation modeling clearly nests conventional multi-state increment-decrement life tables as a special case. Similarly, microsimulation modeling should be able to nest the recent World Bank DALY analysis (19, 20).

The multi-state increment-decrement life table model just reviewed, and cell-based models more generally, tend to embody strong simplifying assumptions. These include independence among processes, and processes represented by transition probability functions with only a few simple independent variables. The practical reason is that more involved process representations entail a combinatorial explosion in the size of the state space, and hence in the number of cells.

However, it is more realistic to consider many processes as interdependent and simultaneous. For example, getting married, buying a house, finishing school, and entering the labor market are decisions or socio-economic transitions occurring in early adulthood that are often jointly determined. Such interaction can be included in microsimulation modeling. The major problem is not any constraint imposed by the microsimulation modeling methodology. Rather, it is the set of difficult empirical questions raised in gathering the appropriate data and estimating any interactions among hazard or transition probability functions.

The POHEM microsimulation structure
POHEM is one particular instance of a microsimulation model, and is very much a work in progress, with a number of areas currently under active development. 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 prospective remarriage partners are explicitly included in this family structure or "case". Thus, individuals are simulated in close to a nuclear family context (i.e., with other individuals who will be part of a given individual's nuclear family at some point in his or her lifetime). The full life cycle of each case is simulated, not just one single 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 DFLE life table model just described in order to illustrate the microsimulation modeling 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. Each state variable is listed below, along with an indication of the method used to set the variable, or the variables drawn upon as inputs to the associated transition probability function.

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), labor force history, CLU history, and pre ordained marriageability (for unobserved heterogeneity, see Rowe and Wolfson).
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.

——


References
165
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.°

child leaving home — probability based on age, sex and birth order of child.

remarriage — probability based on age, sex and whether divorced or widowed.

second spousal age difference — based on marrying person's age at marriage, sex, and prior marital status, then 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 (22).

labour market earnings — dollar level each year based on an auto-regressive stochastic process with parameters based on age, sex, and strength of labour force attachment.

**Risks**

smoking, cholesterol, blood pressure, and obesity — 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 (23). Tomiak and Berthelot have updated the analysis to make use of the recent round of provincial Heart Health Surveys.

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

ages of females at menarche and menopause — based on a sample of 90,000 women from the National Breast Screening Study (24).

**Diseases**

first coronary event incidence — a multivariate risk function from the Framingham study and Merck model (25,26).

heart disease progression, types of events (cardiac arrest, myocardial infarction, angina pectoris, or some combination) and case fatality — based on Weinstein et al. (Ref. (27) and personal commu-
medical oncology advice and 1988 unit costs (34, 35) (Fig. 7). The heart disease treatments in Weinstein et al (27) have not been used. A detailed breast cancer costing module is under development along similar lines to that for lung cancer.

Health
The plan is to use multi-attribute value and utility scales (33). Currently, preliminary weights for disease based on Torrance et al. have been implemented.¹

It should be evident from these very brief sketches of the modules in POHEM that the effort is seriously constrained by available data. The general philosophy has been to push on anyway. Where critical data are needed, we have sought the best possible approximation via expert opinion, consensus panel, or by approaching researchers with relevant data which can be re-analysed to generate the required statistics. The cooperation of many researchers in this endeavour has been invaluable.

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 DFLE 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 woven together from diverse and partial descriptions of the dynamics of health and health-related processes.

One general caveat is in order. The processes in POHEM as just sketched are quite disease-oriented. This is in large part by necessity. While we certainly wish for a much broader perspective on health and health-related processes, the vast majority of existing data and epidemiological studies are disease-based. Extensions, for example, in the directions of the roles of social determinants of health and disability outcomes, must await new data.

Toward a coherent system of health statistics
One method for achieving coherence among a set of statistical indicators is to define all the indicators as projections or transforms (in the mathematical sense) of the same underlying data. POHEM achieves this by creating a common underlying synthetic longitudinal data set representing the full life cycles of a birth cohort. Each of these individual biographies, graphically represented in Fig. 1 by a “rectangle” (state space along the vertical axis, age along the horizontal) can be stacked (with individuals going back into the page along a third dimension). This results, with some poetic license, in a “data cube” like that shown in Fig. 2 (an image also drawn from the Template).

This data cube provides a common (synthetic) microdata foundation and hence, a coherent basis for a variety of derived statistics and graphs, including the HLE family of health outcome measures. For example, the health status information along the bottom row of each individual’s biography 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 ones; 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 historic increase in chronic disease. It is more valuable to have indicators of how healthy people are while they are alive, as well as the expected distribution of their life lengths. To capture this notion, a somewhat different algorithm can be applied. Specifically, 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 — is not thrown away.

The algorithm again extracts the bottom plane of the data cube. But this time, the series of shades of grey is used to represent varying degrees of severity of illness. The conventional survival curve is constructed as before. But we now begin shading the area under the curve. This starts in the lower-left with a pure white area representing life-years spent in full health. Then for each of a series of threshold levels $x$ of less than full health (e.g., 0.9, 0.8, 0.7, ...), the durations of intervals (there may be more than one) in each individual’s biography where he or she was alive and had a health status index value above this threshold $x$ are cumulated. 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 grey.

Fig. 3 illustrates this process using another image from the Template. The top portion shows the bottom plane of the data cube turned up on its edge. A small set of individual health status vectors are shown, with the varying grey shade levels corresponding to varying states of health (in the $[0, 1]$ range).

---

interval). The bottom portion shows the resulting survival curve plus contour plot assuming a num­ber of levels of disability.

The area under the conventional survival curve is simply LE. The heavy "disability-free survival" line illustrates an alternative survival curve based on an arbitrary threshold dividing health states between "healthy" and "disabled". The area un­der this curve is precisely DFLE. More generally, the "weighted" area under the survival curve, with weights corresponding to the health status values represented by shades of grey, is Healthy Life Expectancy (HLE). Of course, HLE could be comput­ed directly from the data cube by summing over.

Rapp. trimest. statist. sanit. mond., 47 (1994)
the entire bottom plane and then dividing by the sample size.

HLE 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 heart disease or arthritis. It could also be broken down by age interval, for example to derive the portion of the discrepancy between HLE and LE attributable to women suffering from arthritis after age 65 (analogous to a disaggregation of the "all items" consumer price index into components for food, transportation, etc.).

Of course, POHEM and this synthetic birth cohort data cube are only part of the proposed system of health statistics. This system should also include a variety of longitudinal microdata sets on real individuals. These real data sets provide the basis for estimates of the various transition probability functions used as inputs to POHEM. For reasons already noted, however, these data sets will inevitably be partial — both in their range of variables and length of follow-up. Still, these data sets cease being a hodgepodge and become coherent to the extent they are built with common concepts and definitions, and become systematically related by feeding into the construction of POHEM.

Thus, a model like POHEM aids statistical coherence in two ways. First, it clarifies the interrelationships among diverse data sources by an explicit set of processes whereby empirical patterns derived from these data sources are woven together. Second, it creates a common synthetic core of longitudinal microdata from which a variety of health indicators, particularly the HLE family, are all derived.

**Illustrations of POHEM-based analyses**

We turn in this penultimate section to a few brief illustrations of POHEM simulations. These results highlight population health status measures, chronic disease burdens, statistical coherence, health interventions, and health research applications.

**Burdens of chronic disease**

To begin, Table 1 shows summary results from a "baseline" POHEM simulation. The morbidity processes associated with 3 major diseases have been modelled: ischaemic heart disease, lung cancer and osteoarthritis, though only the first 2 can be fatal. Each time an individual in the simulation suffers from one of these diseases, there is an age of onset. This may be followed by death from the disease. Individuals may die of other causes, where only mortality has been explicitly modelled, based on rates from vital statistics.

Table 1 shows, for example, that 4.2% of females can expect to have an incident case of lung cancer, at an average age of 68.2 years. Most will not survive, with 3.4% dying. Over half of all individuals can expect to have some form of heart disease during their lifetimes, where this is fatal in about half the cases. Heart disease onset is about 10 years earlier for men than for women. Over three-quarters of all women can expect to suffer from some form of osteoarthritis, beginning on average in their early 50s. The overall result is life expectancies of 80.0 and 73.9 years; but when account is taken of years lived but spent in less than full health due to the 3 major diseases explicitly modelled, healthy life expectancy is from 3 to 5 years less.

### Table 1

**Summary results from a baseline POHEM simulation**

<table>
<thead>
<tr>
<th></th>
<th>Females</th>
<th>Males</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average age (Age moyen)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer onset</td>
<td>68.2</td>
<td>66.8</td>
</tr>
<tr>
<td>Lung cancer death</td>
<td>68.5</td>
<td>69.3</td>
</tr>
<tr>
<td>Heart disease onset</td>
<td>74.8</td>
<td>64.4</td>
</tr>
<tr>
<td>Heart disease death</td>
<td>80.5</td>
<td>73.5</td>
</tr>
<tr>
<td>Arthritis onset</td>
<td>52.7</td>
<td>64.3</td>
</tr>
<tr>
<td>Other causes of death</td>
<td>80.5</td>
<td>74.7</td>
</tr>
<tr>
<td>Life expectancy (LE)</td>
<td>80.0</td>
<td>73.9</td>
</tr>
<tr>
<td>Healthy life expectancy (HLE)</td>
<td>75.0</td>
<td>70.5</td>
</tr>
</tbody>
</table>

*Wld hth statist. quart., 47 (1994)*
Fig. 4

Joint distribution of years with arthritis and years with heart disease (females)

Distribution conjointe des années "avec arthrite" et des années "avec cardiopathie" (femmes)

Figs. 4 & 5 show quite different "views" or sets of results from the same POHEM simulation, focusing on heart disease and arthritis. Fig. 4 shows a scatter plot of durations of time individuals spend burdened by each disease. Since the underlying simulation actually generated 100,000 cases, the plot only shows a random sub-sample of 2,500 cases. This scatter plot indicates a considerable amount of co-morbidity (the dots that are in the positive orthant and not on either of the axes), and suggests that more years are likely to be spent with arthritis than with heart disease. The very ability to construct the plot illustrates the microanalytic aspect of the model.

Fig. 5 summarizes the data represented by the scatter plot into a few average expected durations for the hypothetical birth cohort generated by this POHEM simulation. Given survival to age 15, as was assumed in this simulation, the figure shows an overall life expectancy of 65.0 years (i.e., to age 80.0). Of this total expected life length, 21.1 years or about one quarter can be expected to be spent with some form of arthritis, of which 4.0 years will be with more severe disabling arthritis, and 4.1 will be co-morbid with heart disease. 1.1 years can be expected with both disabling arthritis and heart disease. Looking at these disease burdens from another perspective, 5.8 years can be expected with heart disease, of which 4.1 will be co-morbid with some form of arthritis.

Methodologically, these estimates of co-morbid health state expectancies (sojourn times in multivariate health states, in stochastic process parlance) are simply tabulated results from the "data cube" underlying a POHEM simulation. At the same time, they reflect quite detailed descriptions of the underlying disease processes, and by virtue of the microanalytic method they capture a considerable amount of the heterogeneity of the actual population. (These estimates would be practically infeasible using life table methods, not least because of the complexity of the state space.)

Substantively, considerable co-morbidity is evident. However, by their nature, heart disease and arthritis are independent disease processes. More precisely, they are conditionally independent, since they are both functions of age. It is this common dependence on age that results in co-morbidity rates higher than might otherwise be expected. In turn, these results open the possibility of exploring whether or not there is any interaction between heart disease and arthritis. The hypothesis represented by this POHEM simulation is that they are conditionally independent (on age). If we were to observe significantly more or less co-morbidity in a cross-sectional survey for a given older age group than is generated by the simulation, this would be suggestive that the conditional independence assumption is wrong.

Toward coherence in health statistics

A central objective of POHEM has been to bring coherence to health statistics. One kind of coheren-
Incidence has just been illustrated—Table 1 and Figs 4 & 5 are coherent in so far as they are all "projections" or views or aspects of the identical underlying data cube.

A second kind of coherence arises when statistics are assembled from diverse sources into a systematic framework which can then be used to generate "data confrontations". Two different sources or approaches are used to estimate a given magnitude, where in principle the results should be identical. This provides a method for revealing and assessing the extent of overall error in the many steps of the statistical process.

The example presented here focuses on the lung cancer module. Lung cancer incidence, progression and case fatality have been simulated with each of these 3 variants. For each simulation, steps of the statistical process.

The first and simplest uses only age/sex-specific mortality rates by cause, with lung cancer as one cause of death (the "mortality only" scenario). The data come from vital statistics and the census. They take no account of incidence and morbidity, so lung cancer is effectively modelled as having an infinitesimally short morbid phase always followed by death, i.e., 100% case fatality.

The next variant ("incidence and survival") explicitly distinguishes incidence and case fatality. Incidence is based on Canadian 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 modelled based on a detailed literature review and expert clinical judgement (35).

The most detailed variant ("relative risk") builds on the second. It is the same except that incidence rates are adjusted for risk factor exposures. Data from a survey of 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 (29) is used to scale each individual’s risk of contracting lung cancer as a function of his or her personal risk factor history. Given an incident case, lung cancer progression and case fatality are modelled in the same way as the second variant.

Table 2 shows the results of POHEM simulations with each of these 3 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 2 rates, which is the proportion of all deaths attributable to lung cancer, should also be identical across all 3 scenarios. But the simulations suggest an inconsistency, with registry-based cancer incidence rates yielding lung cancer death rates that are about 0.7 to 1.1 percentage points lower than those coming from mortality rates based directly on death certificates and the population census. Two factors that might account for these discrepancies are under-coverage of lung cancer cases by the cancer incidence registry on the one hand, and classification of some uncommon lung tumours with sites in the lung as "lung cancer" on death certificates. Such tumours should be excluded, since they are managed differently than "typical" lung cancers. One or other of these factors is suggested by the 3.5% mortality rate for females in the first scenario, which is greater than the 3.3% incidence rate from the incidence registry in the other two scenarios.

*All digits shown are significant with respect to Monte Carlo variability. — Tous les chiffres indiqués sont significatifs au regard de la variabilité de la méthode de Monte Carlo.
While the inconsistency between the vital statistics and cancer incidence registry data are clear for females, and this complex POHEM simulation is not necessary to reveal it, some kind of modelling would be required. Moreover, the inconsistency for males would not be revealed by such a simple comparison. In sum, this is an example of data confrontation from otherwise disparate data sources that has been rendered feasible by the coherent structure of POHEM.

**Lung cancer treatment**

The lung cancer module is currently the most detailed in POHEM, especially in so far as it includes a detailed description of care processes. This, in turn, has enabled the simulation of benefits and costs for specific kinds of therapeutic interventions. As an example, Fig. 6 shows the standard "treatment schema" for Stage III and IV non-small cell lung cancer.

Based on detailed 1988 costs, this treatment schema and corresponding schema for other kinds of lung cancer suggest that the total direct costs of lung cancer treatment in Canada in 1988 were $325 million, with an estimated average five year cost per case of $21,500. This latter figure is an average, and ranged from $15,000 to $25,000 for a case of Stage IV non-small cell lung cancer (NSCLC) to $29,900 for limited small cell lung cancer (SCLC).

It is naturally difficult to obtain detailed data on the course of lung cancer in the absence of treatment. Nevertheless, given estimates of such "untreated" survival curves, POHEM simulations suggest that compared to a baseline of no lung cancer treatment at all, the estimated cost per life-year gained was about $11,000 for NSCLC and $19,600 for SCLC (35).³

POHEM has also been used to estimate the costs per life-year gained of several kinds of neo-adjuvant chemotherapy. For example, Evans et al. have estimated that the use of this chemotherapy as part of combined modality therapy involves costs between $2,900 and $5,000 per life-year gained.²

**Impacts of cholesterol-lowering interventions**

We turn next to estimated impacts of a quite different kind of health intervention, cholesterol lowering strategies (36). The starting point is Frank et al.'s re-analysis of the Honolulu Heart Study data (37). They replicate the usual elevated mortality risk from heart disease in the case of high serum cholesterol levels at baseline. However, their analysis goes beyond the common heart disease end point to consider mortality from other causes. As shown in Table 3, their results indicate elevated mortality risk both for high and for low levels of serum cholesterol. One important corollary of these data is that it is not clear a priori that a cholesterol-lowering programme will be beneficial for an entire population when all causes of death are considered.

---


### Table 3

<table>
<thead>
<tr>
<th>Cholesterol range (mg/dl)</th>
<th>Cancers (including lung)</th>
<th>Other circulatory diseases</th>
<th>Other diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>mg/dl</td>
<td>mg/dl</td>
<td>mg/dl</td>
<td>mg/dl</td>
</tr>
<tr>
<td>&gt;260</td>
<td>0.9</td>
<td>1.5</td>
<td>1.0</td>
</tr>
<tr>
<td>200-260</td>
<td>1.0</td>
<td>1.2</td>
<td>1.0</td>
</tr>
<tr>
<td>180-200</td>
<td>1.0</td>
<td>1.0</td>
<td>1.1</td>
</tr>
<tr>
<td>160-180</td>
<td>1.0</td>
<td>1.0</td>
<td>1.1</td>
</tr>
<tr>
<td>140-160</td>
<td>1.3</td>
<td>1.5</td>
<td>1.2</td>
</tr>
<tr>
<td>&lt;140</td>
<td>2.0</td>
<td>1.5</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Rapp. trimest. statist. sanit. mond., 47 (1994)
Algorithms to reflect these specific relative risk patterns were added to the baseline disease incidence modules already in POHEM. This is equivalent to the assumption that the observed relationships in Table 3 are causal, and not just associations. Then 3 POHEM simulations were run — a base case and 2 health-promoting intervention scenarios. The first intervention was a dietary policy which had the effect of lowering everyone’s total cholesterol, whatever it was, by 5%. The second added to this dietary policy selective prescription of cholesterol-lowering drugs for those with higher risks. This latter scenario is actually a bit complex because it assumed drug prescription and use would depend on individual risk factor profiles as well as partial compliance by patients. The assumed impacts of both of these intervention scenarios on cholesterol levels follow the published literature in this area.

Table 4 shows results for males. Everyone “departs” by age 75, either by one of 3 causes of death, or by surviving to age 75. Lowering cholesterol does lower death from heart disease. But the increased mortality from other causes, assuming the Honolulu Heart Study relative risks are causal, leaves the proportion surviving to age 75 under the first dietary intervention essentially unchanged. The second, more aggressive intervention results in almost a 10% reduction in the proportion of deaths from heart disease before age 75 (-1.28% against 14.82% in the base case). But this intervention also raises deaths from other causes, so that half of the reduction in heart disease deaths is matched by increases in other causes of mortality. It should be emphasized that these are not necessarily rigorous results (see Ref. (41) for alternative analyses). However, they highlight the importance for health policy of a comprehensive, population-based perspective and a concern with all endpoints, not just those of interest to a specific disease or research group.

**Explaining socio-economic status gradients in mortality**

In our final example, POHEM has been used to explore the sources of the clearly observed positive association in Canada of a variety of socio-economic status (SES) measures and health. One possibility is the observed inverse correlation of smoking with educational attainment. If smoking is more prevalent in lower SES groups, this factor alone might account for a very large portion of the observed gradient in mortality by SES.

The base case scenario in POHEM has smoking independent of any SES variables. (Smoking does depend on age and sex, and is correlated as observed in the 1978 Canada Health Survey with prior smoking, and other risk factors like obesity and hypertension.)

To explore this question, an alternative scenario was constructed where smoking prevalences were adjusted to reflect observed patterns by educational attainment (39). The question was then how much of a mortality gradient, using educational attainment as a marker for SES, could be generated simply by this observed correlation of SES and smoking. As points of comparison, 3 other scenarios were also constructed: one where no one ever smoked, one where everyone was a heavy smoker from age 15 onward, and one where smoking was as in the base case but there was no death at all from lung cancer. The results, in terms of HLE, are shown in Fig. 7.

The bold horizontal lines in the two graphs represent overall HLE (not LE), given survival to age 25 and assuming the distribution of smoking patterns observed in Canada independent of SES — 70.6 years for men and 74.9 years for women. Next, the top two horizontal lines represent hypothetical scenarios where either no one smoked at all, or smoking was as usual but no one ever contracted lung cancer. These two scenarios have very similar impacts, and raise HLE by 0.5 and 0.6 years respectively for women, and 0.9 and 0.7 years for men. They show that a fully successful “war against smoking” would have at least the same payoff in terms of HLE gains as instantly eliminating all lung cancer mortality, taking account only of the effects of smoking on lung cancer and ischaemic heart disease (IHD), about half the tobacco-sensitive causes of death.

At the other extreme, the bottom line represents a hypothetical scenario where everyone was a
heavy smoker all their lives. This heavy smoking scenario has a greater impact on men, lowering their HLE by 2.0 years, and that of women by 1.1 years. Overall, the two smoking extremes account for a range of about 3 years in HLE for men, and almost 2 years for women. This smaller range for women is associated with their lower (inferred) average age-specific incidences for lung cancer and IHD, holding smoking rates constant.

Finally, the stepped line in the middle gives the scenario where each of three educational attainment groups smokes at the distribution of rates actually observed. For women, the larger effect is for the lowest educational group, lowering HLE by about 0.2 years, while for the middle and upper educational status groups, HLE is about 0.1 years greater than average. For men, we see a more pronounced gradient, with a range in HLE of 0.7 years from the lowest (70.2 years) to the highest (70.9 years) educational attainment group.

The rough conclusion was that about one-fifth of the observed gradient in mortality by SES in Canada might be “explained” by the smoking-SES gradient working through lung cancer and IHD.

**Conclusion**

This article has described the origins and use of the POHEM microsimulation model in the context of a number of broad statistical and health science issues.

Numerical microsimulation models like POHEM can play several key roles in health statistics. One is to produce summary indicators of population health, by drawing on such generalizations of life expectancy as healthy life expectancy (HLE). In this role, POHEM can help remedy the imbalance in much of current health statistics where far more data are available on inputs and throughputs of the health care system than on population health outcomes.

A second role for a simulation model like POHEM is to provide coherence to health information, in a context where many data series are little more than a hodgepodge. Such a coherent structure, combined with sensitivity analyses, can also help guide development of new statistical sources.

A third role for POHEM is to support decision-making, particularly with regard to resource allocation for health-affecting interventions. POHEM can be used to model the impact of possible interventions, and then generate quantitative estimates of costs and benefits, where benefits are measured in terms of health expectancy.

A fourth role is as a tool for basic health science research. As shown in the co-morbidity and SES gradient examples above, POHEM can be used to obtain indirect, quantitative estimates regarding the importance or character of various health-related phenomena.

POHEM is a work in progress. Among its weaknesses are that it is still primarily disease-centred; it makes extraordinary demands on data; and it presumes far more knowledge than currently exists regarding the causal pathways influencing human health. At the same time, this ambitious character may help spur the data development and basic health science research POHEM requires, thereby facilitating advances in new areas of health science and health policy analysis.

**Summary**

A variety of developments have come together to serve as both an impetus to and foundation for the development of a new POPulation HEalth Model (POHEM) at Statistics Canada. Part of the impetus is statistical and derives from weaknesses in Canada’s health statistics programme — particularly the lack of balance between information on health outcomes and health care resource consumption, and the absence of a coherent statistical structure. The other major impetus is the need for rational processes for managing and allocating resources to improve the health of Canadians. The foundation for the development of this model has come from the revolution in computing. Dramatic improvements have opened up new methodological opportunities, particularly sophisticated simulation modelling and detailed analyses of large volumes of microdata. POHEM is designed to build on these increasingly powerful methods in order to meet health statistical and policy needs. At this time, POHEM is like a partially-completed model.
building. This article reviews its motivation, the overall architectural plan, and the portion of the structure already completed. A major portion of POHEM is devoted to the explicit modelling of chronic disease processes, using Monte Carlo microsimulation methods. The article concludes with illustrations of a few recent applications, focusing on the joint patterns of smoking, cholesterol and heart disease, osteoarthritis and lung cancer morbidity. While POHEM has been developed in a Canadian context, work is under way to create a version that can be used in other countries.

Résumé

Plusieurs faits nouveaux ont contribué à stimuler l’élaboration par Statistique Canada d’un modèle de santé des populations (POPopulation Health Model: POHEM). L’inspiration en est en partie d’origine statistique et trouve plus précisément sa source dans les lacunes du programme de statistiques sanitaires du Canada, en particulier le déséquilibre entre les données sur les résultats sanitaires d’une part, et les données sur la consommation des ressources destinées aux soins de santé d’autre part et, en plus, l’absence d’une structure statistique cohérente. L’autre motivation principale est la nécessité de disposer d’un processus rationnel de gestion et de répartition des ressources en vue d’améliorer la santé des Canadiens. C’est la révolution dans le domaine de l’informatique qui a permis de mettre au point ce modèle. Des améliorations spectaculaires ont ouvert la voie de nouvelles méthodologies, en particulier une modélisation perfectionnée par simulation et l’analyse détaillée d’un grand volume de microdonnées. Le POHEM est conçu pour tirer parti de ces méthodes de plus en plus performantes afin de répondre aux besoins en matière de politiques et de statistiques sanitaires. À l’heure actuelle le POHEM évoque un bâtiment encore inachevé. Le présent article décrit sa motivation, le plan architectural global et le partie de la structure déjà achevée. Une part importante du POHEM est consacrée à la modélisation explicite de l’évolution des maladies chroniques, avec utilisation des méthodes de microsimulation de Monte Carlo. L’article conclut en illustrant quelques applications récentes. Elles sont axées sur le tableau commun de la morbidité due au tabagisme, au cholestérol et aux cardiopathies, à l’arthrose et au cancer du poumon. Le POHEM a été conçu dans un contexte canadien, mais des travaux sont en cours pour en créer une version susceptible d’être utilisée dans d’autres pays.

References/Références


34. Evans, W.K. et al. Diagnostic and therapeutic approaches to lung cancer in Canada and their costs (in review, 1994).


