

HealthPaths Dynamics 1
Estimating Multivariate Trajectories of
Functional Health over the Life Course

Geoff Rowe & Michael Wolfson, Statistics Canada

International Microsimulation Association, Ottawa, June 8-10, 2009

Abstract

The Health Utilities Index (HUI) is a widely used summary index of an individual's functional health, computed as a non-linear function of eight distinct categorical attributes – vision, hearing, speech, mobility, dexterity, cognition, emotion, and pain. The dynamics of HUI in a representative sample have been observed with Statistics Canada's National Population Health Survey every two years since 1994. These dynamics are complex, not least because they reflect at least two broad types of health change: accidents and acute events on the one hand, and gradual or progressive changes typically related to various chronic conditions on the other. Inferring the dynamics of HUI and its component attributes is further complicated by the myriad determinants of health, resulting in both observed and unobserved sources of heterogeneity.

This paper will describe estimation of the dynamics of HUI based on a system of equations for each of its component attributes plus mortality. The many determinants of health will be taken into account both explicitly, by including Education as well as time-varying covariates: obesity (measured by body mass index or BMI), Daily Smoking, a psycho-social covariate termed 'Coherence', and implicitly by allowing for unobserved (occult) heterogeneity. More generally, the estimation process is designed for subsequent analysis using microsimulation to synthesize a representative sample of realistic functional health life course trajectories.

1. Introduction

A model of health dynamics is necessarily complex. The 'web of causation' (Krieger, 1994) – essentially a path diagram representing the interlaced causal relationships among social, environment, biological etc. factors and health outcomes – is complex to begin with, but often presents an essentially static view. The way in which the interplay of these relationships could evolve over the course of an individual's life suggests yet another level of complexity - with the influences expressed in the path diagram being iterated dynamically, many times over a single life. The cumulative impact of these iterations defines an individual's health trajectory (HealthPaths). However, some properties of health trajectories may not be directly represented in the path diagrams at all. For example, random path effects (individual differences: Orpana, et.al, 2009) can have an iterated impact that is not at all clear from the diagram itself. Moreover, there need be no unique way of representing the relationships, if they encompass a mix of direct exogenous and lagged endogenous pathways (Zellner and Palm, 1974)¹. For these reasons, we

¹ Using algebraic re-expression of systems of linear time-series equations, Zellner and Palm demonstrate as many as five distinct forms of equation systems, each expressing the same information in different ways. The five forms of equations exploit the fact that, in dynamic systems, lagged endogenous variables may substitute for combinations of lagged or time-invariant exogenous variables, and vice versa.

propose that microsimulation models may provide a feasible approach to the empirical exploration of the complexity of the ‘web of causation’ and of the dynamics arising from it. In effect, a new role for microsimulation is that of an essential analytical tool, on par with graphical techniques, for exploratory analysis of data.

This paper is the first of a two part description of our early work on HealthPaths. It will focus on the specification and estimation of a system of dynamic health equations; their implementation in the HealthPaths computer model and the initial validation of the output from that model. Our simulation model uses Statistics Canada's Modgen technology (details available on the Statistics Canada website: www.statcan.gc.ca/spsd/Modgen.htm), which greatly facilitates implementation.

2. Functional Health Dynamics

Our analysis focuses on the health dynamics of individuals as measured by the Health Utility Index or HUI (Feeny, et.al, 2002; and www.healthutilities.com/HUI.htm). This index provides a description of an individual's overall functional health using eight attributes: vision, hearing, speech, mobility (ability to get around), dexterity (use of fingers and hands), cognition (memory and thinking), emotion (feelings), and pain. Based on a standard set of questions, the HUI provides a summary functional health score between -.360 and 1.000. For instance, an individual who is nearsighted, yet fully healthy on the other seven attributes, receives a score of 0.973. On this scale, the most preferred health level (perfect health) is rated 1.000 and death is rated 0.000, while negative scores reflect health states considered worse than death.

The HUI score is derived utilizing a weight that has been assigned to each level of the eight attributes. The weights embody the views of society concerning health status. These views are termed “societal preferences”, and are based on preferences about various health states as elicited from a representative sample of individuals. The HUI was developed by McMaster University's Centre for Health Economics and Policy Analysis, and was derived using societal preferences from a random sample of 500 people within the boundaries of the City of Hamilton, chosen from a list obtained from the Planning Department of the Regional Municipality of Hamilton-Wentworth, Ontario, Canada. The algorithm mapping the questions to the HUI is itself the property of Health Utilities Inc.

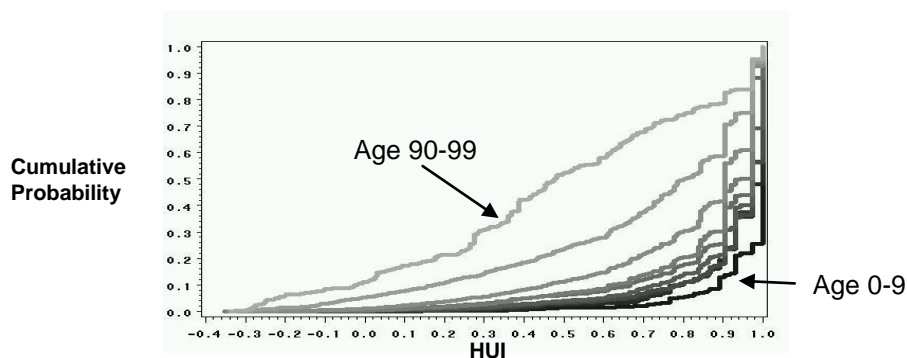
Our analysis of longitudinal HUI data draws on Statistics Canada's National Population Health Survey (NPHS) (Statistics Canada, 1999). The NPHS is a panel survey of self-reported health based on interviews conducted biannually over more than a decade. The initial sample comprised over 17,000 respondents, with more than 11,000 providing a full response in all of the seven cycles available to us. For purposes of variance estimation, NPHS data files are disseminated with 500 sets of bootstrap weights (Yeo, et.al, 1999). In this study, we have used only the Household component of the NPHS; but, for longitudinal purposes, this also includes information on respondents who were initially sampled in a private household and were subsequently institutionalized. A more complete analysis would also include those who were initially sampled in an institution.

Wolfson and Rowe (2004), as well as Rowe and Binder (2008) have shown that health dynamics, as reflected in HUI, can be complex: for example, periods of stability might be followed by abrupt temporary changes in state (e.g., accidents) or by spells of gradual change. The former, at any rate, was the interpretation of relatively frequent changes from perfect health (HUI=1.0) to less than perfect health and back to perfect health over the course of four years. It is difficult to develop a coherent model incorporating both dynamics of that sort and dynamics that arise from the declines in health associated with age. It is even harder to account for such dynamics by

modelling HUI as a continuous variable, because some features of the dynamics appear extremely non-linear. As such, the approach that we will take in this study will focus on the discrete dynamics underlying the HUI itself (i.e., dynamics of vision, hearing, etc.) and then to utilize the power of microsimulation to reassemble simulated values of HUI leading to summary indicators of its dynamics.

2.1 Functional Population Health by Age Group

The health of a majority of children, as assessed by the HUI, is characterized by perfect or near perfect health. At succeeding ages, the proportion at or near perfect health declines and the range of HUI over which the remainder of the population is distributed increases. These basic facts can be seen in the empirical distribution functions in Figure 1 which display cumulative probabilities versus corresponding HUI values for each of ten 10-year age groups. In this chart, HUI appears to provide a highly plausible description of the affect of aging on population health. However, our concern in this paper will be with the problem of providing a realistic description of individual health trajectories.



**Figure 1: Empirical HUI distribution functions by age group:
10-year groups ordered youngest (black)-to-oldest (light grey)**

2.2 Components of Functional Health

There are two essential elements that make up an HUI score. The first step in the calculation of HUI involves classification using a standard set of categories on the eight dimensions of functional health:

Vision

1. No visual problem
2. Problem corrected by lenses
3. Problem seeing distance - not corrected
4. Problem seeing close - not corrected
5. Problem seeing close and distance - not corrected
6. No sight at all

Hearing

1. No hearing problem
2. Problem hearing in group - corrected
3. Problem hearing in group and individual - corrected
4. Problem hearing in group - not corrected
5. Problem hearing in group and individual - individual not corrected

6. Cannot hear

Speech

1. No speech problem
2. Partially understood by strangers
3. Partially understood by friends
4. Not understood by strangers
5. Not understood by friends

Mobility

1. No mobility problem
2. Problem - no aid required
3. Problem - requires mechanical support
4. Problem - requires wheelchair
5. Problem - requires help from people
6. Cannot walk

Dexterity

1. No dexterity problem
2. Dexterity problem - no help required
3. Dexterity problem - requires special equipment
4. Dexterity problem - requires help with some tasks
5. Dexterity problem - requires help with most tasks
6. Dexterity problem - requires help with all tasks

Emotion

1. Happy and interested in life
2. Somewhat happy
3. Somewhat unhappy
4. Very unhappy
5. So unhappy that life is not worthwhile

Cognition

1. No cognition problem
2. A little difficulty thinking
3. Somewhat forgetful
4. Somewhat forgetful/a little difficulty thinking
5. Very forgetful/great deal of difficulty thinking
6. Unable to remember or to think

Pain

1. No pain or discomfort
2. Pain does not prevent activity
3. Pain prevents a few activities
4. Pain prevents some activities
5. Pain prevents most activities

Then, given an individual's eight dimensional classification, a summary HUI score is derived utilizing the weight that has been assigned to each of the categories. The weights, as described above, embody the views of society concerning health status. Our challenge, in the work reported

here, is to accurately simulate the dynamics of HUI by reconstructing HUI scores from the simulated components.

2.3 Covariates of Functional Health

The starting point for our implementation of HealthPaths will be the assumption that everyone begins – at birth – with perfect health (i.e., Vision=1, ..., Pain=1 => HUI=1). Thereafter, the health trajectory dynamics will be modulated by:

1. intrinsic relationships among the components of functional health (for example, some underlying health processes involving Dexterity problems in one time period will imply both Mobility and Dexterity problems in the next time period), and
2. extrinsic influences of covariates on individual components of functional health.

The following four sections identify the four main extrinsic covariates employed in the current version of our HealthPaths model. Age also plays a vital role in our model, but it is obviously another intrinsic feature of HealthPaths (i.e., survival and the passage of time).

2.3.1 Socio-Economic Factors: Education

We use education variables to introduce meaningful socio-economic diversity into our simulations. Two education attainment variables are used: 1) Less than Secondary School Graduate (LtSSG) and 2) University Graduate (BA+) [i.e., Secondary School and Community College Graduates are the reference category]. We chose to implement Education by imputing a lifetime attainment level to each simulated individual at birth and by initializing their time-varying education level to LtSSG. Then, if appropriate, the time-varying education level is incremented on the 18th and 22nd birthdays – representing Secondary School and University graduation, respectively. The data employed in the imputation of lifetime attainment levels is the corresponding set of proportions drawn from the 2006 Census of Canada. In the simulation, this imputation depends on Sex and Birth Cohort only.

2.3.2 Behavioural Factors: Daily Smoking

Smoking is a key risk factor for many health outcomes and may act as a proxy for still other behavioural factors. Our implementation of Daily Smoking involves generating a single random waiting time until a ‘Start Smoking’ event and then, as part of that event, generating a single random waiting time until a ‘Stop Smoking’ event. In effect, simulated individuals are limited to one smoking period over the course of their lives and Years of Daily Smoking are accumulated during that period. Other parts of the simulation model make direct use of the time-varying classification of persons into Never, Current and Former Daily Smoker categories, as well as the Years of Daily Smoking covariate.

The Daily Smoking model utilizes NPHS data on (possibly censored) age at first daily smoking from respondents age 25 or older and (possibly censored) duration of daily smoking from respondents age 50 or older. We fit a Weibull distribution to this censored waiting data employing Sex, Lifetime Education and Birth Cohort as time-invariant covariates. Only the final valid observation from each eligible NPHS respondent was used in these estimations.

2.3.3 Lifestyle Factors: Body Mass Index

The body mass index (BMI) is another key risk factor for many health outcomes and may also act as a proxy for yet other lifestyle factors (work conditions, etc.). Our implementation of BMI imputation involves making a random draw from a parametric distribution at least once a year during the simulation. The parametric distribution has three separate equations that express relations between covariates and 1) the median, 2) the dispersion and 3) the asymmetry of the distribution (see Section 2.5 for more detail on this specification). Each of the three equations has covariate terms comprised of Age, Years of Daily Smoking, Time-varying Education, HUI, and Sex. The random component in these equations is Gaussian with two lagged correlations. The latter ensures that a simulated individual's rank in the BMI distribution will drift gradually over their lifetime.

2.3.4 Psycho-Social Factors: Coherence

Coherence is a 13-item version of the sense of Coherence scale developed by Antonovsky (1987). It denotes the extent to which individuals perceive events as comprehensible, manageable and meaningful. Within the simulation model, it is imputed in a similar fashion to BMI and its system of three imputation equations has the same covariates. However, a simulated individual's rank in the Coherence distribution is determined at birth and remains fixed. Nevertheless, imputed Coherence values vary over time, as determined by the selected time-varying covariates.

Use of the Coherence covariate imposed constraints on the overall model. Not only was it reported only in the first two NPHS interviews, but it was limited to respondents who were 18 years of age or older. Thus, to avoid problems of a model seam at age 18, all of our estimates of health and covariate dynamics are based on older (18+) respondents. As such, simulations of health dynamics below age 18 are entirely extrapolations.

2.4 Modeling Change in the Functional Health of Individuals

Specification of the equations in our model of functional health change is motivated by two key observations: 1) that the categories on each of the eight dimensions are ordered and 2) that the interval between NPHS interviews (two years) is large enough for multiple (unobserved) events to have occurred. These two observations suggest a conservative use of the data might be to identify two types of event:

$$\begin{aligned} \text{Better Vision: } & \text{Vision}_{t+2} < \text{Vision}_t \\ \text{Worse Vision: } & \text{Vision}_{t+2} > \text{Vision}_t \end{aligned}$$

where Vision_t and Vision_{t+2} are the categories obtained from responses in two successive interviews. A jump of two levels of Vision between interviews is modeled as if this change involved two distinct steps, one of which was unobserved and as if health always shifts up or down a functional health status continuum one-step-at-a-time. Consequently, we estimate two separate hazard equations corresponding to the two types of events for each HUI component (e.g., Better Vision and Worse Vision) for a total of 16 equations. From the perspective of accuracy of estimation, the practical consequences of this formulation are that: more observations contribute to estimation of each equation than would be the case for direct transitions between each pair of states, and that the ordinal character of the variables is respected.

Our implementation of these equations in the simulation model capitalizes on the fact that, using Modgen, events are simulated in continuous time; therefore, in the course of simulation, control passes instantaneously from one set of equations to another as the 'current' simulated, functional health state changes. Thus, successive moves up and down the functional health status continuum

are generated in continuous time, which, we hope, will approximate the multi-year, multi-step jumps observed in the NPHS.

It is likely that event probabilities vary with the duration in a given functional health status, but, in the NPHS, sojourn times in functional health categories were not reported. Thus, we estimate event hazards conditional on lagged (t-2) and current (t) functional health categories, which – together with the additional observation needed to identify change (t+2) – requires the longitudinal data to be arranged into sequences of three interviews. With seven waves of the panel survey in hand, we can identify as many as five such sequences from each respondent (totaling more than 50,000 such sequences). However, having made use of the multiple responses collected from most individual respondents, we must also be prepared to add person-specific terms to the equations to account for the correlation between observations drawn from the same respondent. These terms will represent otherwise unobserved person-specific factors (i.e., those that remain fixed from time-period to time-period) and, as such, reflect the effects of some omitted variables.

Ultimately, our estimation approach assumes a conditional Poisson probability (Lawless, 1987) for binary events y (i.e., representing either better or worse functional health events) jointly with an unobserved, person-specific, Gaussian random effect z . In the following expressions, ‘ i ’ denotes a respondent and ‘ t ’ a time-period; while PY_{it} represents approximate person-years-at-risk (i.e., two years for respondents who experienced no ‘event’ and one year for respondents who reported any change in health between interviews: an approximate adjustment for differences in exposure that allows for competing risks).

Likelihood Equations

$$y_{i,t} = PY_{i,t} \exp(\beta_{i,t} + \sigma z_i) \varepsilon_{i,t}$$

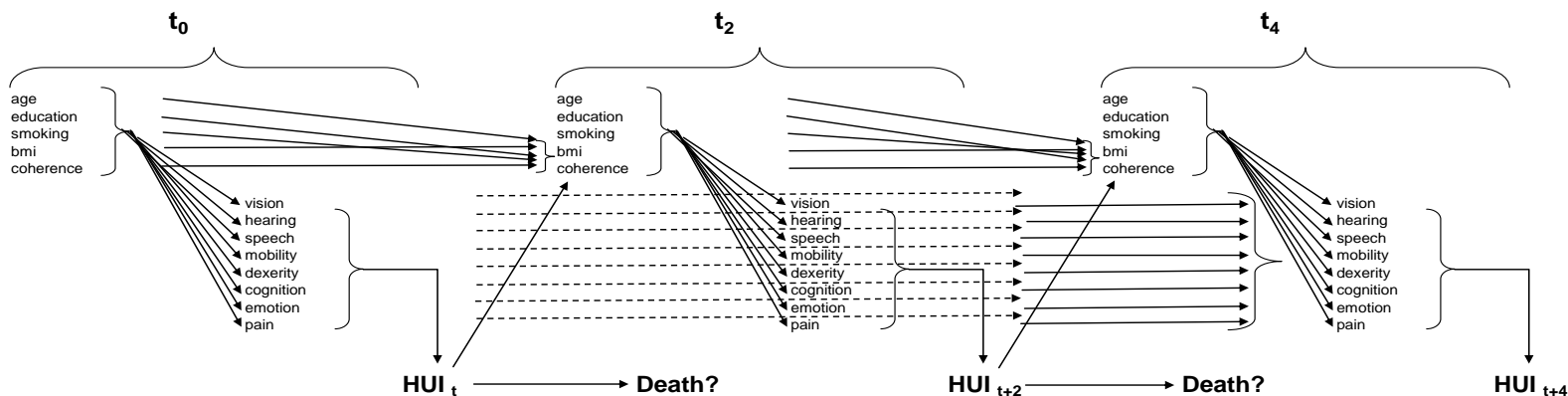
$$p(y_{i,t} | PY_{i,t}, X_{i,t}, \beta, \sigma) = \int p(y_{i,t} | PY_{i,t}, X_{i,t}, \beta, \sigma, z_i) p(z_i) dz_i$$

where $y_{i,t} | PY_{i,t}, X_{i,t}, \beta, \sigma, z_i \sim \text{Poisson}(PY_{i,t} \exp(\beta_{i,t} + \sigma z_i))$
and $z_i \sim \text{Gaussian}(0, \sigma^2)$

$$\hat{\beta} \approx \arg \max_{\beta} \left[\prod_{i,t} p(y_{i,t} | PY_{i,t}, X_{i,t}, \beta, \sigma) \right] \text{ subject to } \left(\sum_{j=1}^J |\beta_j| \right) < s$$

In practice, however, we first maximized a simple conditional Poisson distribution (i.e., holding z equal to 0.0 – it’s expected value) employing a procedure called Forward Stagewise Regression to estimate β subject to the constraint on the absolute values of the estimates. This technique – described in detail below – simultaneously selects a model and estimates its coefficients. We then used the SAS procedure NLMIXED to estimate the random effect parameter σ minimizing the Poisson-Gaussian log likelihood conditional on the selected model (Pinheiro and Bates, 1995). Constrained estimation permitted specification of much larger candidate equations than is normally feasible. In this application, $J=1071$ terms (not including the intercept) were estimated for each of the 16 hazard equations (for details on the equation specification - see Figure 2, below).

Figure 2: An approximate schematic of direct and lagged relationships estimated from NPHS data



Covariates: Education depends on Sex & Birth Cohort only; while Daily Smoking depends on Sex Education & Birth Cohort only.

HUI Component Equations (each of 16 equations):

- Concurrent and Lagged Component Category Indicators - 81 terms
- BMI, Smoking & Coherence Variables - 3, 8, & 2 terms
- Age & Age Interactions with each Category Indicator & the other variables - 174 terms
- Interactions of the 1st 268 terms with Sex - 268 terms
- Interactions of the 1st 268 terms with Education Less Than Secondary School - 268 terms
- Interactions of the 1st 268 terms with Education BA+ - 268 terms

BMI & Coherence Equations:

- Age - 6 terms
- Daily Smoking - 9 terms
- Sex - 3 terms
- Education Less Than Secondary School - 3 terms
- Education BA+ - 3 terms
- HUI - 9 terms

TOTAL

- 1072 terms

2.4.1 Estimation of β : Forward Stagewise Poisson Regression

The risk of over-fitting is a well-known problem in construction of useful regression models. It is even more of a problem in large models for which estimation is difficult to begin with and where commonly available techniques for model selection may be impractical. The introduction of a constraint on the absolute values of coefficients (see Likelihood Equations above) represents one step towards making model selection with a large number of candidate variables simple. The technique was first introduced by Tibshirani (1996, 1997) and termed the LASSO (Least Absolute Shrinkage and Selection Operator). Tibshirani noted that the absolute value constraint functioned like Ridge Regression in shrinking coefficients closer to zero than they would be in unrestricted estimation; but, that it was far more likely than Ridge Regression to result in a shrunken estimate of a given coefficient exactly equal to 0.0 and thereby also functioned like a model selection technique.

A second step in simplifying selection/estimation of large models was recognition of the close connection between the LASSO and Forward Stagewise Regression (further detail is available in Freidman, 2001, Hastie, et.al, 2001, Efron, et.al, 2004). The Forward Stagewise Regression algorithm frequently leads to a close approximation of the LASSO and, in the event of a divergence between the two, may provide an improvement on the LASSO in large models (Hastie, et.al, 2007). The algorithm has the following form (for regression estimates that maximize likelihood L):

Generalized Forward Stagewise Regression Algorithm

1. Start with $\beta_0, \beta_1, \beta_2, \dots, \beta_p = 0$ and with centred/standardized covariates X
2. Find the index 'm' corresponding to the largest score $[|\delta \ln(L) / \delta \beta_m|]$ evaluated at the current predictor μ
3. Update $\beta_m = \beta_m + \varepsilon * \text{sign}(\delta \ln(L) / \delta \beta_m)$, where ε is small (0.005 in our work)
4. Update the intercept β_0 and the predictor μ
5. Repeat steps 2-4 many times and test each time for termination
6. On termination the estimates $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are back-transformed to adjust for covariate standardization.

(adapted from Hastie, et.al, (2007)). Since we used a simple Poisson likelihood at this stage, the weighted scores took the form: $\sum_i [(Y_i - \mu_i) * X_{ik} * W_i]$, where W_i is the survey weight. Thus, at each updating step, the active coefficient 'm' is the one for which there is the strongest association between X_m and the current regression residuals. The Forward Stagewise Regression algorithm requires many passes through the data; but, at its core, only requires inspecting a vector of log likelihood gradients to locate the maximum absolute score and, as such, the number of candidate variables can be very large.

In application of the algorithm, it is generally good practice to standardize the covariates (average 0, standard deviation 1) in order to eliminate differences of scale among the covariates. Having done that, repeating K iterations of the algorithm will result in $\sum |\beta_j| = K * \varepsilon$, which corresponds to the coefficient constraint (see Likelihood Equations). The final problem to be resolved is deciding when to stop the updating process (i.e., choosing a terminal value K corresponding to the constraint parameter $s = K * \varepsilon$). We chose to use a cross-validation measure of out-of-sample prediction error as a stopping criterion (Rowe and Binder, 2008). During step 5 of the algorithm, we assessed the prediction error based on updated estimates and stopped the iterations (i.e., chose a final value of s) when prediction error began to increase [a sure sign of over-fitting]. Evaluation

of the prediction error involved use of bootstrap weights for estimation and unsampled observations for cross-validation. The use of bootstrap weights also permits evaluation of the combined effect of sampling and model selection on coefficient estimation.

To demonstrate the effectiveness of the algorithm, Figure 3 displays out-of-sample prediction error from an early modeling attempt that had few enough candidate covariates to permit unrestricted estimation of all of the terms (i.e., a conventional estimate produced by SAS Proc GENMOD). Figure 3 shows differences by iteration number between the updated prediction error and the final unrestricted error. The difference (excess error) is positive where the unrestricted model is superior and negative where the Forward Stagewise estimate is superior. In this example, fewer than 150 iterations ($\epsilon = 0.005$) were needed to find a model that gave better predictions than the unrestricted model. Between iterations 300 and 400, estimates were nearly equal in predictive power, while iterations beyond 400 show steadily worse predictions. Eventually, a very large number of iterations should converge on or near the unrestricted estimate (i.e., where excess error equals 0, again).

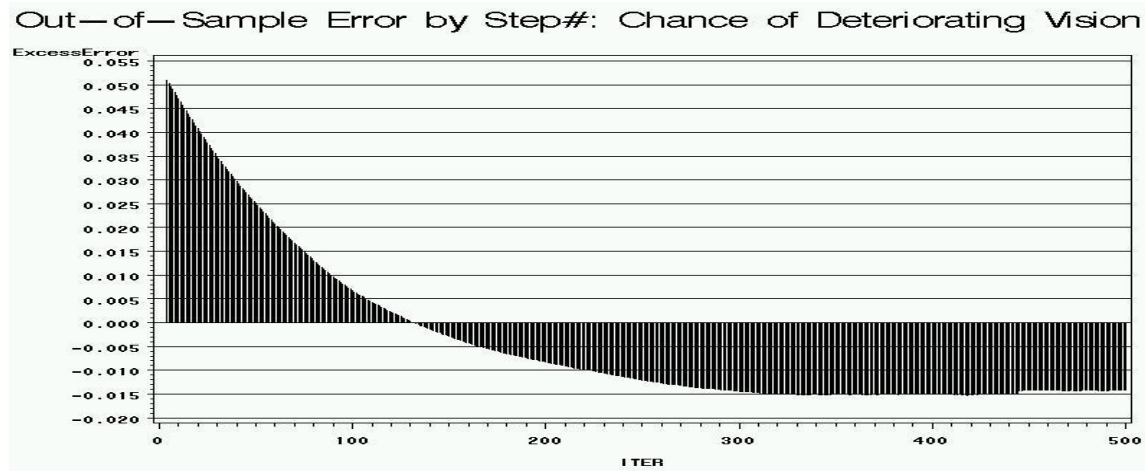


Figure 3: Evolution of Forward Stagewise prediction error over 500 iterations.

Forward Stagewise Regression has two main advantages: it is capable for identifying better predictive models than unrestricted stepwise regression methods, and it is capable of estimating very large models. The latter advantage was as important in our work as was the former. As indicated in Figure 2, our equations representing functional health change had 1072 candidate covariates (counting interaction terms as distinct covariates). One reason for considering equations of this size was that we did not know, a priori, whether to stratify on Sex or on Education or both or none. The question of whether to estimate separate equations for different strata can be difficult to answer unless the sample size is large. The situation with covariates like Education that have more than two strata can be even more problematic. (Note that the Forward Stagewise Regression Algorithm can handle models with substantially more covariates than observations).

But, in our application, the more important reason for considering the models of this size has to do with Age interactions. The assumption that a lagged relationship between two functional health states will have a fixed strength independent of age may be unrealistically simplistic. In concrete terms, the chances of a fracture leading to other health complications are likely much greater and recovery will tend to be much slower at older ages than for teenagers. But, specification of the interaction terms that would account for age mediated concurrent and lagged

influences quickly leads to models that are not readily handled by unrestricted estimation techniques, either because of sheer size or because they push limited data too hard. Forward Stagewise Regression provides an effective way to handle both problems.

2.4.2 Estimation of Residual Variance (σ^2): SAS Proc NLMIXED

Equation 1 specifies a mixed likelihood that includes regression slopes (β) and a person-specific random term (z) with variance parameter σ^2 . Explicit estimates of z_i can be obtained given multiple observations from most respondents. Such estimates resemble averaged respondent-specific residuals which, in the present case, were obtained conditional on the Forward Stagewise slopes. However, more simply - assuming that the z_i 's approximately follow a Gaussian distribution with variance σ^2 and are independent among respondents - a numerical-integration formula [Gauss-Hermite quadrature] provides the basis for estimation of σ^2 maximizing the integrated (marginal) likelihood that does not need explicit respondent-specific residuals [using the techniques implemented in the SAS procedure NLMIXED]. We used the general likelihood facility in NLMIXED so that survey weights could be incorporated in the Poisson-Gaussian likelihood (i.e., maximizing a marginal pseudo-likelihood). Since only a single parameter per equation was estimated in this step, it was relatively easy.

2.5 Modeling BMI and Coherence Covariates

As noted above (sections 2.3.3 and 2.3.4), imputation of both BMI and Coherence utilized a common set of covariates and a common form of equation (Y represents the observed BMI or Coherence, as appropriate):

$$Y \approx \begin{cases} A + B * \frac{\exp(G * u) - 1}{G}, & G \neq 0 \\ A + B * u, & otherwise \end{cases}$$

where u is a standard Gaussian random variable (average 0 and standard deviation 1.0) and where the terms A , B and G represent linear equations:

$$A \text{ is the median equation: } A = D \theta_A$$

$$B \text{ is a dispersion equation: } B = \max(0.0, D \theta_B)$$

$$G \text{ is an asymmetry equation: } G = D \theta_G$$

and, finally, where D is a design matrix of covariates. These equations specify a complete family of distributions (power transformations of a log-normal) with each of three properties of the distributions depending on covariates (D) via the parameters θ_A , θ_B , and θ_G . An extended version of this family of distributions which can also account for patterns of tail elongation (Hoaglin, 1985) has been recommended for use in random imputation (He and Raghunathan, 2006).

These regression equations may be thought of as smooth generalizations of quantile regression, since the random term u will reflect a simulated individual's rank (quantile) in the distribution. In the case of BMI, u follows an autoregressive process (i.e., $u_t = r_0 * v_t + r_1 * v_{t-1} + r_2 * v_{t-2}$, where the v 's are standard Gaussian) with parameters (r) estimated from empirical lagged correlations. However, in the case of Coherence, u is fixed at birth for each simulated individual. This was because the Coherence scale was only available from the first two waves of the NPHS and lagged

correlations could not be reliably estimated. Estimation was carried out using Proc NLMIXED to facilitate specification of u as standard Gaussian.

The method of dealing with over-fitting of these equations was more conventional than in the case of functional health change equations. A straightforward, backward elimination technique was applied in linear least squares regression of BMI and Coherence on the candidate covariates. This approach was motivated by the fact that the expected value of the dependent variable reflects influences from all three properties of the distributions of the data:

$$E(Y | D) = \begin{cases} A + B * \frac{\exp(0.5 * G^2) - 1}{G}, & G \neq 0 \\ A, & otherwise \end{cases}$$

In a succession of steps, candidate covariates were eliminated one-at-a-time, if the associated p-value was above 0.05. Estimation was carried-out using SAS Proc SURVEYREG, so that the significance tests took account of the complex design of the NPHS (i.e., stratification and clustering). But, as a means of selecting appropriate covariates for the median, dispersion, and asymmetry equations, this can only be a first approximation.

2.6 Accounting for Non-linear Covariate Effects

It would be a serious error, in a model as exploratory as this, to assume that all the covariates entered each of our equations in a linear fashion. That said, we did not want non-linearity to become the focus of this stage in our work. As such, we chose a relatively parsimonious - but flexible – functional form for each of our continuous covariates (Age, BMI, Coherence, Years of Daily Smoking and HUI). Following Hastie, Tibshirani, and Friedman (2001), a natural cubic spline with K knots can be represented linearly by K basis functions counting the intercept. [A cubic spline is a cubic polynomial on the interval between knots but is continuous and has continuous first and second derivatives at the knots. A natural cubic spline adds additional constraints so that the fitted curve will be linear at the endpoints (i.e., below the minimum knot and above the maximum knot). These properties produce flexible and smooth fitted curves that can also provide sensible extrapolations beyond the endpoints.] The two spline basis functions (involving three knots) that we incorporated in our equations were:

linear term ⇒ X

$$\text{non-linear term} \Rightarrow T(X) = \frac{(X - X_1)^2 - (X - X_3)^2}{X_3 - X_1} - \frac{(X - X_2)^2 - (X - X_3)^2}{X_3 - X_2}$$

for knots $X_1 < X_2 < X_3$

In the current work, we chose to use the same sets of knots for each covariate in all of the equations that we estimated: these sets of knots (selected based on exploratory scatter plots) are: Age – 30, 50 & 70; BMI – 20, 25 & 30; Coherence – 50, 60 & 70; Years of Daily Smoking – 10, 20 & 30; and HUI (in the BMI and Coherence equations) – 0.25, 0.5 & 0.8.

2.7 Relating Mortality to Health: Calibration to the 1960 Birth Cohort

The empirical basis for mortality was the set of age-sex-cohort-specific mortality hazard parameters estimated and projected as part of the LifePaths microsimulation model (details

available on the Statistics Canada website: www.statcan.gc.ca/spsd/LifePaths.htm). In addition, two other components were borrowed from the LifePaths model: relative mortality risks associated with institutionalization and additional relative mortality risks associated with HUI. The previous work with the LifePaths model had shown that the information available (mainly from the NPHS, but augmented with Census data) on the relative mortality risks associated with institutionalization and HUI was inadequate. For example, only 1552 of the 17,276 household respondents in the longitudinal part of the NPHS died in the follow-up period and only 1309 were ever institutionalized. When divided between the sexes and stratified by age, this sample was not adequate to provide completely reliable estimates of relative risks. Therefore, calibration of those components of the LifePaths model was found to be necessary: another example of the all-to-common problem of limitations in available data.

Direct implementation of these borrowed components in the HealthPaths model required recalibration. In this case, the mortality benchmark (see Figure 4) focused on comparisons of average simulated distribution of the 1960 birth cohort ages at death (conditional on simulated HUI) with the corresponding distribution implied by the input mortality hazard parameters. Uncalibrated simulations revealed that mortality levels were too high given the simulated HUI distributions. This discrepancy was eliminated by calibration, allowing for age-sex-specific calibration factors.

After combining the estimated relative risks with calibration factors, our mortality model implied a negligible difference, below age 80, between the input average hazard and the implied hazard for persons living in private households evaluated at average age-sex-specific HUI levels (Figure 4). However, examination of estimated relative risks (Figure 5) reveals anomalous ratios (greater than 1) for women between age 45 and 65. Clearly this module requires further development. At issue is the question of whether the coupling of mortality and functional health is properly identified.

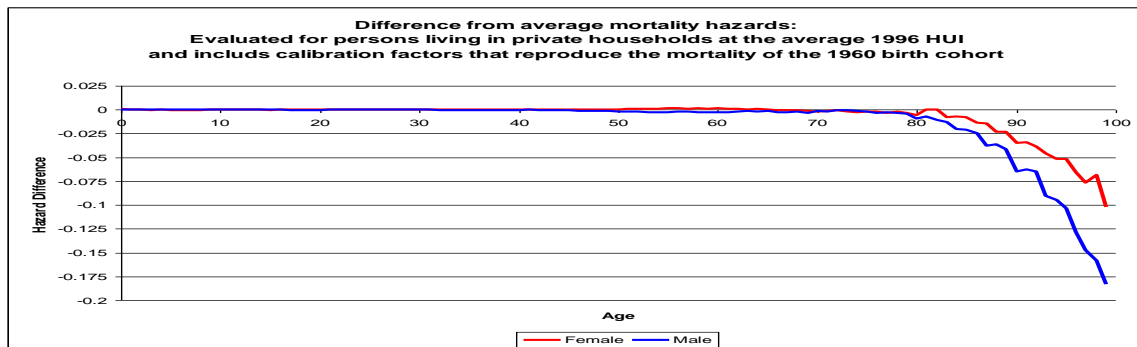


Figure 4: Difference from average hazards: non-institutional population with average HUI spreadsheet calculations using NPHS data and model parameters.

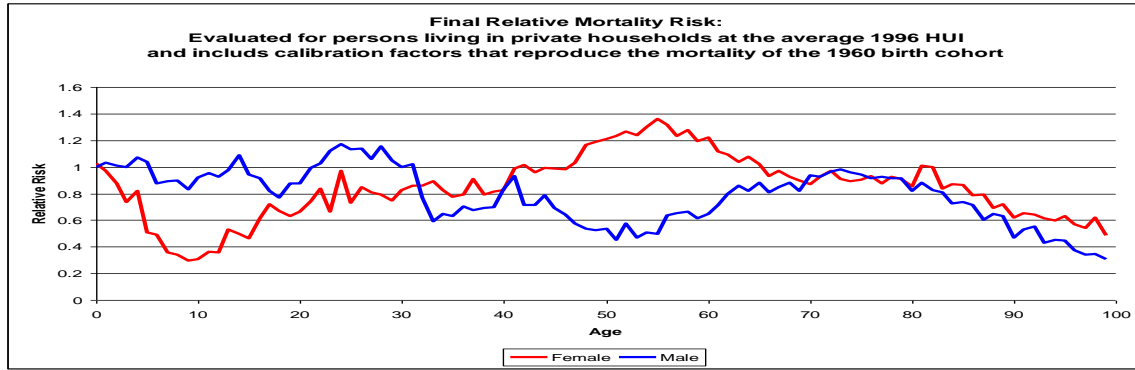


Figure 5: Relative mortality risks: non-institutional population with average HUI spreadsheet calculations using NPHS data and model parameters.

3. Validation

Before carrying out any experiments with this model, it was essential to perform a set of validation tests that are conditional on the adjustments to reproduce the mortality of the 1960 birth cohort. The purpose of these tests is to ensure that the baseline version of the models provides a reasonable approximation of the most salient properties of the processes involved, by comparing output directly to NPHS observations. For this purpose, the simulation model had to be constructed so that it could produce results representative of a broad range of birth cohorts. This is despite the fact that, for present purposes, simulation experiments will be conducted on a single birth cohort. Thus, differentiation among birth cohorts in the Education, Smoking and Mortality modules is intended only to aid validation.

The suite of charts in Figure 6 compare observed and simulated profiles of mortality, smoking, BMI, Coherence and HUI for a broad range of ages in the Canadian population as of 1996; the year in which there was a special (large) cross-sectional sample that augmented the longitudinal part of NPHS. As noted above, cohort differences in mortality and in smoking had to be calibrated. In addition, simple bias adjustments were made to the intercept and male coefficients of the BMI median equation. Otherwise, all simulation parameters were used as they had been estimated.

The bottom-line of the simulation is the accuracy of the HUI scores reconstructed from their simulated components. The average HUI appears to be accurate up to about age 50; thereafter, the simulations decline faster than is observed. But, subsequent charts reveal that the discrepancy is due to simulation of a smoothly increasing proportion in Poor health (HUI<0.5) after age 50. This contrasts with the observed trend which shows a persistently small proportion in Poor health until about age 70, after which there is an abrupt increase. Otherwise, the observed and simulated proportions in Good and in Fair health seem to be in close agreement.

Perhaps the discrepancy in the lower tail of the HUI distribution could be addressed by fine-tuning the degree of non-linearity in the Age terms of the various equations. Two other alternative explanations for the discrepancy might be considered: 1) a more complete analysis of the institutional population might be required – for example, by estimating equations that incorporate institutional interactions; and 2) birth cohort specific factors may be at play – for example, differences among cohorts in the relative mortality risks associated with HUI and/or differences in the influence of educational attainment levels on health (i.e., secondary school

graduation was less common historically, than is the case today). Ultimately, the explanation will probably turn out to be a combination of factors rather than one.

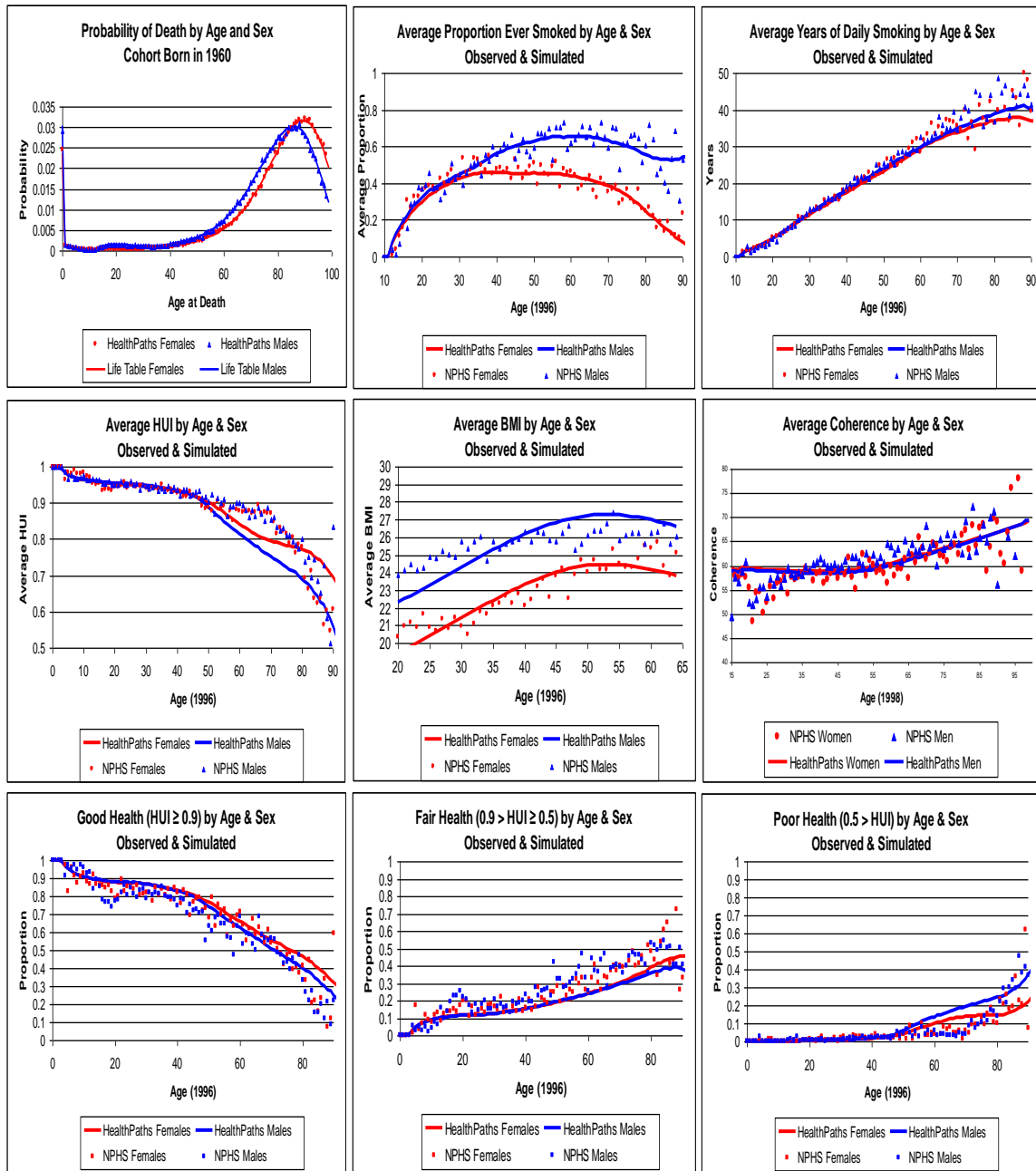


Figure 6: Suit of validation tests

4. Accounting for Model Uncertainty

4.1 Covariate Selection

Bootstrapping the model selection process accounts for intrinsic variability in the estimation of coefficients as well as the variability in choice of coefficients for estimation; thus, the procedure

produces a more comprehensive assessment of model uncertainty than is usually the available. The method is related to the so-called 'bagging' (Bootstrap AGGREGatING) technique (Breiman, 1996), which is itself related to the 'Bayesian Bootstrap' (Clyde and Lee, 2001). The main benefit of the method, as it relates to model selection, is improved identification of appropriate covariates, which is often needed because choosing whether to include or exclude a covariate from a model can be an unstable process. Ultimately, estimating a range of equations that variously include or exclude covariates may have an effect that is similar to the coefficient shrinkage exploited by the LASSO (i.e., bootstrap model selection will tend to include a given covariate with a certain probability, otherwise it will have a coefficient of zero; whereas, shrunken estimates will all tend to be drawn closer to zero than an otherwise unrestricted estimate).

One way by which to assess the overall contribution of covariates to prediction is to look at the frequency with which they are picked in the model selection process. Given the large number of interaction terms in the Health Equations, we provide a summary count of the number of terms that involve, for example, Age as either a main effect or an interaction, and so on, for each of the candidate covariates. This necessarily results in double counting. Table 1 shows the number of terms that involve each candidate covariate; this count totals 1600, whereas there are only 1072 terms in each equation.

Table 1: Number of terms involving each candidate covariate

Age	BMI	Smoking	Males	Education		Coherence
				LtSSG	BA+	
696	28	48	268	268	268	24

Table 2 displays the average frequency of selection of terms involving each candidate covariate (evaluated over 50 bootstrap model selection samples). It appears that Age will have a particularly strong influence given the number of times it was chosen (despite the fact that it ranks low relative to the potential number of terms). Coherence appears to be a surprisingly strong covariate considering its general unfamiliarity. In terms of frequency of selection, Coherence is on par with more familiar covariates like Smoking and BMI. Education has the most marked pattern; it plays an important role in 'Better' health events, but was never involved in 'Worse' health events. The interpretation of the Education results should be put in context: each estimated coefficient is a direct partial effect distinct from the other (correlated) covariates in the equations. Education attainment, as with Sex, will tend to be time-invariant over the range of ages most important to health change; and, given that, some of the impact that it might have had will inevitably be taken up by lagged functional health status effects.

Table 3 displays coefficients of variation in percentages (%CV) that correspond to the averages shown in Table 2. The minimum %CV is 25 and many are above 100; CVs as large as this indicate that the model selection process – despite being guided by minimization of out-of-sample prediction error – is quite variable from sample to sample. In most cases, that variability is greater than if selection were a simple random event (e.g., a Poisson distributed count varying from sample to sample).

Table 2: Average frequency of covariate selection: Health Equations
Blanks correspond to situations in which covariates were never selected.

Better Health

	Age	BMI	Smoking	Males	Education	Coherence	
					LtSSG	BA+	
Vision	11.7	1.8	2.3	7.7	3.1	5.0	2.7
Hearing	12.4	1.3	0.5	1.9	3.1	3.9	0.5
Speech	0.5	0.1	0.6	1.0	0.2	0.6	0.1
Mobility	10.2	0.6	0.6	1.7	2.2	2.4	0.3
Dexterity	2.2		0.6	1.7	0.8	3.5	0.0
Emotion	8.0	0.9	2.0	3.7	1.9	5.5	2.2
Cognition	6.3	0.0	3.0	4.5	1.4	4.9	1.1
Pain	10.8	2.3	2.4	5.5	3.1	8.3	0.6

Worse Health

	Age	BMI	Smoking	Males	Education	Coherence	
					LtSSG	BA+	
Vision	19.8	0.2	2.9	7.5			2.0
Hearing	28.8	1.6	1.4	11.4			0.5
Speech	2.2		0.0	1.6			0.1
Mobility	38.0	1.8	1.6	6.9			0.4
Dexterity	16.7	0.0	0.3	2.2			0.0
Emotion	24.0	0.8	4.1	7.2			4.5
Cognition	25.5	0.4	3.3	8.8			5.3
Pain	25.8	1.3	3.8	6.5			3.9

Table 3: Coefficient of Variation (%) of covariate selection counts: Health Equations
Blanks correspond to situations in which covariates were never selected.

Better Health

	Age	BMI	Smoking	Males	Education	Coherence	
					LtSSG	BA+	
Vision	47	49	64	35	69	51	52
Hearing	36	66	125	80	73	63	170
Speech	250	523	117	97	216	172	323
Mobility	43	140	125	103	95	77	214
Dexterity	136		86	112	158	90	495
Emotion	58	87	70	71	124	48	41
Cognition	66	495	50	50	111	50	69
Pain	44	54	77	54	73	39	144

Worse Health

	Age	BMI	Smoking	Males	Education	Coherence	
					LtSSG	BA+	
Vision	31	231	59	49			39
Hearing	45	83	81	49			161
Speech	237		707	133			400
Mobility	30	31	48	47			168
Dexterity	59	707	183	104			707
Emotion	25	100	33	35			26
Cognition	25	182	35	39			25
Pain	32	44	32	51			26

4.2 Simulating Uncertainty

Each of the sets of equations estimated from a given bootstrap sample can be used to simulate a complete set of results. Thus, the variability among sets of simulation results will gauge the usual Monte Carlo variability plus variability intrinsic to coefficient estimation plus model selection variability plus the variability due to random person-specific effects (occult heterogeneity). In all, this provides a more comprehensive picture of the variability than is typically the case.

A common practice is to estimate a set of regression coefficients from the all of the available data, and to treat the associated coefficient standard errors as measures of the intrinsic inaccuracy of the estimates. This is entirely appropriate when the specification of the equation is known and the intrinsic inaccuracy of the estimates is predominantly due to measurement error. However, in the present case, the relatively incomplete specification of our models implies the presence of heterogeneity that can be attributed, in large part, to omitted variables and interaction terms. These omissions may give rise, in our bootstrap model selection and estimation, to variability that predominantly reflects the genuine heterogeneity of the population, rather theoretically irrelevant measurement error.

Moreover, because many of the equations were estimated from the same survey (HUI component, BMI and Coherence equations were based on NPHS data), there will be a degree of correlation between equations estimated from the same bootstrap sample (akin to 'seemingly-unrelated equations', Zellner (1962)). Thus, the significance of differences in simulation output (e.g., comparing baseline to variant scenarios) will be most accurately reflected in the pair wise differences between separate bootstrap samples.

Another feature of this simulation strategy is that averaging the output across the results of simulations based on separate bootstrap samples leads to an estimate of the average of the predictive distribution. The latter roughly corresponds to a non-parametric Bayes distribution given a uniform (non-informative) prior (Efron and Tibshirani, 1998). This feature is important because we may think of the simulation as a non-linear, noisy transformation of imperfectly known simulation parameters (i.e., $T(\text{Parameters})$) that are assumed subject to intrinsic random variation (e.g., individual variation). In that case, the fact that $E(T(\text{Parameters})) \neq T(E(\text{Parameters}))$ can make a considerable difference to the simulation output and could be a general source of bias in simulations that ignore coefficient and/or model uncertainty (Stow, et.al, 2006). The difference is analogous to the difference between the mean and the median of a skewed distribution. By construction, our approach simulates the mean: $E(T(\text{Parameters}))$.

5. Conclusion

This paper provides an outline of the estimation work that went into construction of the HealthPaths model. It sets the stage for **HealthPaths Dynamics 2**, which will begin an exploration of parameter and simulation space and will reveal aspects of the dynamics associated with the 'web of causation' (Figure 2) that require microsimulation to delineate.

Nevertheless, some important conclusions may be drawn directly from the results presented here. Firstly, the 'web of causation' that we've chosen to study is probably too simple, rather than too complex. We have found considerable evidence of interactions between Age and other parameters in the equations. On average, 40% of the Age interaction terms that were detected are simple interactions between effects of current or lagged functional health state and linear or non-linear Age terms. This is not unexpected, since we know that bodily functions and repair mechanisms change with age. But, this finding implies that the absence of Age interactions – and

the concomitant assumption of age-invariant covariate effects – may represent a serious oversimplification common to many health models.

An additional source of complexity in our work has been the representation of variability due both to estimation error and to uncertain model selection, as well as that due to the residual heterogeneity among individuals. Thus far, the findings – even if the result of a preliminary exploration – serve as a demonstration of a genuine need for modeling tools that can accommodate high dimensionality, if we are to explore life course dynamics (and longitudinal data) deeply. Ultimately, these explorations will also likely require use of microsimulation, as an analytical tool, given the complexity of the models expected to emerge.

References

- Antonovsky, Aaron. (1987). *Unraveling the mystery of health*. Jossey-Bass. San Francisco.
- Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24, pp. 2350-2383.
- Clyde, M.A. and Lee, H. K. H. (2001). Bagging and the Bayesian bootstrap. In T. Richardson and T. Jaakkola, editors, *Artificial Intelligence and Statistics 2001*, pages 169-174.
- Efron, Bradley and Tibshirani, Robert (1998). The Problem of Regions. *The Annals of Statistics*, 26(5), pp. 1687-1718.
- Efron, Bradley, Hastie, Trevor, Johnstone, , Ian and Tibshirani, Robert (2004). Least Angle Regression, *The Annals of Statistics*, 32(2), pp. 407-451.
- Feeny,D., Furlong,W., Torrance, G., Goldsmithh, C.H., Zhu, Z., DePauw, S., Denton, M., and Boyle, M. (2002). Multi-attribute and single-attribute utility functions for the Health Utilities Index Mark 3 system. *Medical Care*, 40(2), pp. 113-128.
- Friedman, Jerome H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5), pp. 1189-1232.
- Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome (2001). *The Elements of Statistical Learning*. Springer Verlag, New York.
- Hastie, Trevor, Taylor, Jonathan, Tibshirani, Robert and Walther, Guenther (2007). "Forward stagewise regression and the monotone lasso". *Electronic Journal of Statistics* 1(1), pp.1-29.
- He, Yulei and Raghunathan, Trivellore E. (2006). Tukey's *gh* Distribution for Multiple Imputation. *The American Statistician* 60(3), pp. 125-129.
- Hoaglin, David C. (1985) "Summarizing Shape Numerically: The g-and-h Distributions" in *Exploring Data Tables, Trends, and Shapes*, David C. Hoaglin, Frederick Mosteller and John W. Tukey eds. Wiley, New York.
- Krieger, Nancy (1994). Epidemiology and the Web of Causation: Has Anyone Seen the Spider? *Soc. Sci. Med.* 39(7). pp. 887-903,
- Lawless, J.F. (1987). "Regression Methods for Poisson Process Data". *Journal of the American Statistical Association*, 82, pp. 808-815.
- Orpana, Heather M., Ross, Nancy, Feeny, David, McFarland, Bentson, Bernier, Julie, and Kaplan, Mark (2009). The natural history of health-related quality of life: A 10-year cohort study. *Statistics Canada, Catalogue no. 82-003-XPE • Health Reports*, 20(1), pp. 29-35.
- Pinheiro, J.C. and Bates, D.M. (1995). "Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model," *Journal of Computational and Graphical Statistics*, 4, pp. 12-35.

Rowe, Geoff and Binder, David (2008). "Can Bootstrap Replicates be used for Cross-validation?". 2008 *Proceedings of the Survey Research Methods Section*. American Statistical Association Annual Meeting.

Statistics Canada (1999), "Information about the National Population Health Survey", Catalogue No. 82F0068XIE, www.statcan.ca/english/concepts/nphs/index.htm

Stow, Craig A., Reckhow, Kenneth H., and Qian, Son S. (2006). A Bayesian Approach to Retransformation Bias in Transformed Regression. *Ecology*, 87(6), pp. 1472–1477.

Tibshirani, Robert (1996). Regression shrinkage and selection via the lasso.

J.Royal.Statist.Soc. B, 58, pp. 267-288.

(1997). The lasso method for variable selection in the Cox model.

Statistics in Medicine, 16, pp. 385-395.

Wolfson, Michael and Rowe, Geoff (2004), "Disability and Informal Support: Prospects for Canada." in Cohen S B and J M Lepkowski (Eds.), *Eighth Conference on Health Survey Research Methods, PHS 04-1013*, National Center for Health Statistics, Hyattsville, MD.

Yeo, Douglas, Mantel, Harold and Liu, Tzen-Ping (1999). "Bootstrap Variance Estimation for the National Population Health Survey". *1999 Proceedings of the Survey Research Methods Section*. American Statistical Association Annual Meeting.

Zellner, Arnold (1962). An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association*, 57, pp. 348–368.

Zellner, Arnold and Palm, Franz (1974). Time Series Analysis and Simultaneous Equation Econometric Models. *Journal of Econometrics*, 2, pp. 17-54.