



DEMOGRAPHIC RESEARCH

A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 31, ARTICLE 31, PAGES 941–974

PUBLISHED 23 OCTOBER 2014

<http://www.demographic-research.org/Volumes/Vol31/31/>

DOI: 10.4054/DemRes.2014.31.31

Research Article

HealthPaths: Using functional health trajectories to quantify the relative importance of selected health determinants

Michael Wolfson

Geoff Rowe

This publication is part of the Special Collection on “Multistate Event History Analysis,” organized by Guest Editors Frans Willekens and Hein Putter.

© 2014 Michael Wolfson and Geoff Rowe.

This open-access work is published under the terms of the Creative Commons Attribution NonCommercial License 2.0 Germany, which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See <http://creativecommons.org/licenses/by-nc/2.0/de/>

Table of Contents

1	Introduction and approach	942
2	Statistical analysis	950
3	Simulation methods	957
4	Results	960
4.1	Base case scenario	960
4.2	Simulation experiments	963
5	Summary and concluding comments	970
6	Acknowledgements	971
	References	972

HealthPaths: Using functional health trajectories to quantify the relative importance of selected health determinants

Michael Wolfson¹

Geoff Rowe²

Abstract

BACKGROUND

It is widely known that there is a significant dispersion in health status, as well as a strong correlation between health status and socio-economic status. But considerable uncertainty remains as to the quantitative importance of the various explanatory factors typically cited in this context. As health status is intrinsically a reflection of co-evolving dynamic processes, it is important to take a lifetime perspective when seeking to understand its determinants. The “bottom line” measure of overall population health is, though, health-adjusted life expectancy (HALE), which is an aggregation of individuals’ health-adjusted life lengths (HALLs).

OBJECTIVE

In an exploratory mode, we intend to provide a realistic assessment of the relative importance of selected health determinants of HALE.

METHODS

This paper first draws on very detailed estimates of the covariates of vector-valued functional health trajectories, using the National Population Health Survey (Statistics Canada). We then use longitudinal microsimulation to draw out their implications by synthesising first a realistic base case – specifically, a representative longitudinal population sample – and then a series of exploratory counterfactual populations. Comparisons between and among counterfactuals and the base case are then used to estimate the quantitative importance of various factors in accounting for HALE.

RESULTS

Several surprising results emerged. Of the four risk factors explicitly examined, obesity had the smallest impacts on HALE: moving from the fifth to the 95th percentiles of BMI increased HALE 1.5 and 2.5 years for men and women, respectively. Eliminating smoking increased HALE by five and four years, while moving from the lowest to the

¹ University of Ottawa, Canada. E-Mail: Michael.Wolfson@uOttawa.ca.

² University of Ottawa, Canada.

highest levels of education had similar effects of about five years for both men and women. Perhaps surprisingly, moving from the fifth to the 95th percentiles of the psycho-social factor, Antonovsky's sense of coherence, had the greatest impact on estimated HALE, with increases of 5.5 to six years.

CONCLUSIONS

While the analysis is exploratory, the results show the analytical potential of tightly coupling detailed empirical analysis of complex longitudinal data with microsimulation modelling. Substantively, the results of the analysis contribute to the evidence that the widespread public health preoccupation with the "obesity epidemic" is over-stated, while psycho-social and mental health problems continue to be seriously under-emphasised.

1. Introduction and approach

It is a commonplace that there is great variation in health status among individuals in a population. It is also increasingly appreciated that these variations are not random. Researchers have long known that some diseases run in families, and that workers exposed to noxious working conditions are more likely to become ill and die prematurely. More recently, a number of studies have provided clear evidence of strong and pervasive associations between health status and socio-economic status (SES) (e.g., the Black Report, Whitehead et al. 1982), while other studies have shown that SES is a major determinant of health, and have thus introduced the SES gradient in health. (Evans et al. 1994; Wolfson et al. 1993).

But understanding *how* SES affects health and causes the gradient – and how it compares with other factors in terms of its relative importance as a determinant of health – is much more challenging. For example, it is well understood that smoking is bad for health, and data show that smoking is more prevalent among individuals of lower SES. Thus, part of the reason for the observed SES gradient in health is the SES gradient in smoking. Other evidence indicates that psycho-social stressors, such as an effort-reward imbalance and relational injustice in the workplace, are also bad for health (Head et al. 2007).

A fundamental question is, therefore, to what extent each of these empirically observed relationships contribute to, "explain," or even "cause" the overall observed patterns in health status. Moreover, it is clear that many of the factors affecting an individual's health status unfold over the life course. Thus, we would ideally like to have an explanation not only for an individual's health status at a single point in time, but also for the person's health status trajectory over his or her full life cycle. A

summary indicator of a population sample of these life course health trajectories, or “health paths,” is health adjusted life expectancy (HALE; Sullivan 1971; Wolfson 1996; Wolfson and Rowe 2001; McIntosh et al. 2009).

As an initial attempt to address these issues, we present in this paper the results of an exploratory analysis that will allow us both to estimate HALE, and to construct a tool (a simulation model) which can be used to assess the relative importance of various factors in determining HALE. This analysis draws on data from Canada’s National Population Health Survey (NPHS; Statistics Canada 1998) and from a smaller, specially designed version of the POHEM (POpulation HEalth Model; Wolfson 1994), a microsimulation model we are calling HealthPaths.

Our first main objective is to demonstrate a new approach to the causal attribution of overall population health to a set of candidate factors, which are assumed to interact and co-evolve. The other main objective is—provided the factors are indeed shown to be causal – to determine their relative quantitative importance. Ascribing causality to observed associations is difficult. Since this analysis will use longitudinal data, one important criterion for ascribing causality will be met; i.e., the potential causes and effects will always be appropriately ordered temporally.

The general goal is therefore to conduct the analysis in two distinct but tightly coupled phases: statistical estimation followed by microsimulation modelling. Since this is an exploratory analysis, we focus in both phases on the empirical relationships between health status and a very limited set of four basic health determinants: (1) a measure of SES, in this case educational attainment; (2) daily smoking; (3) obesity, as measured by BMI (body mass index); and (4) a psycho-social factor, Antonovsky’s sense of coherence (Antonovsky 1987).³ The NPHS includes a number of psycho-social variables. We have focused on the sense of coherence because it is hypothesised to play a fundamental role in individuals’ long-term health. This variable was originally included in the NPHS in a more experimental mode in order to support exploratory analyses such as this one; and compared to other psycho-social variables like depression, it has not been well-studied.

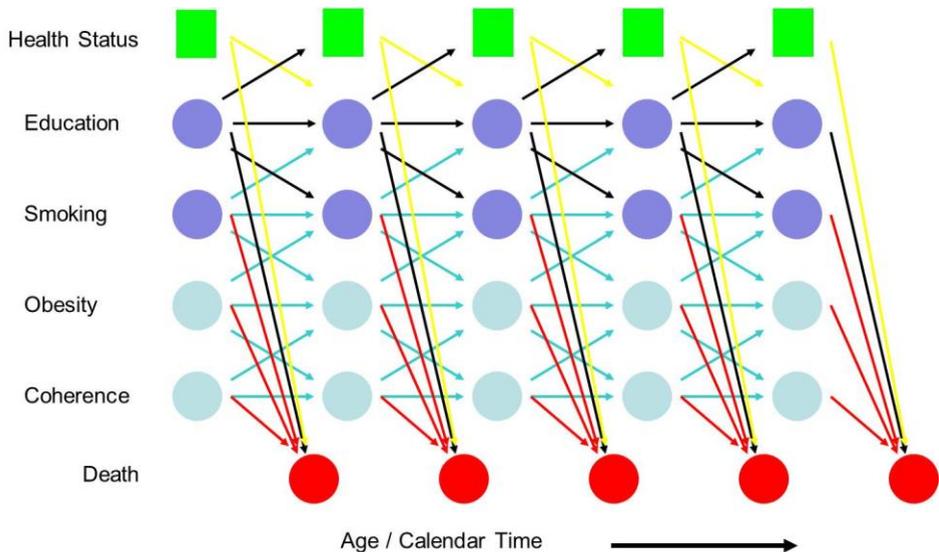
It is of course possible that the specific variables being observed and eventually simulated are themselves proxies for other factors. Furthermore, as only a small subset of all of the factors studied are potential determinants of health, there will be a pervasive risk of omitted variable bias in the statistical analysis to follow.

³ Sapolsky (2004) summarises the Sense of Coherence as “reframing the main psychological modifiers of the stress response in the context of society” in terms of a series of key questions about coping and coping strategies:

“Does a person have the sense of being linked to the mainstream of society? ...
Can a person perceive society’s messages as information, rather than as noise? ...
Has a person been able to develop a set of coping responses? ...
Does a person have the resources to carry out plans?”

The empirical relationships for both phases of the analysis form a network, as illustrated in Figure 1, whereby each “blob” (coloured circle or rectangle) represents a measure of each of the indicated constructs, and each arrow represents a “causal pathway.” The circular blobs for education (measured in terms of attainment), smoking (current, former, never, or occasional on a daily basis), obesity (measured by BMI), and coherence (on a 78-point scale) indicate that the construct is essentially univariate (see below). Death is also shown as a circle, whereby the measure in this case is the exact time of death.

Figure 1: System of recursive equations (n.b. not all possible arrows shown)



Health status, in contrast, is shown as a rectangle. The shape indicates that this construct is actually a composite; i.e., an eight-dimensional vector of health states in each of the domains forming the McMaster health utility index (HUI; Feeny et al. 2001): vision, hearing, speech, mobility, dexterity, pain, cognition, and emotion. For each of these domains, the measure is an ordered categorical variable with five or six levels. Additionally, there is a valuation or scoring function that maps the full vector of eight ordered categorical variables into an index in which one is full health and zero is dead.

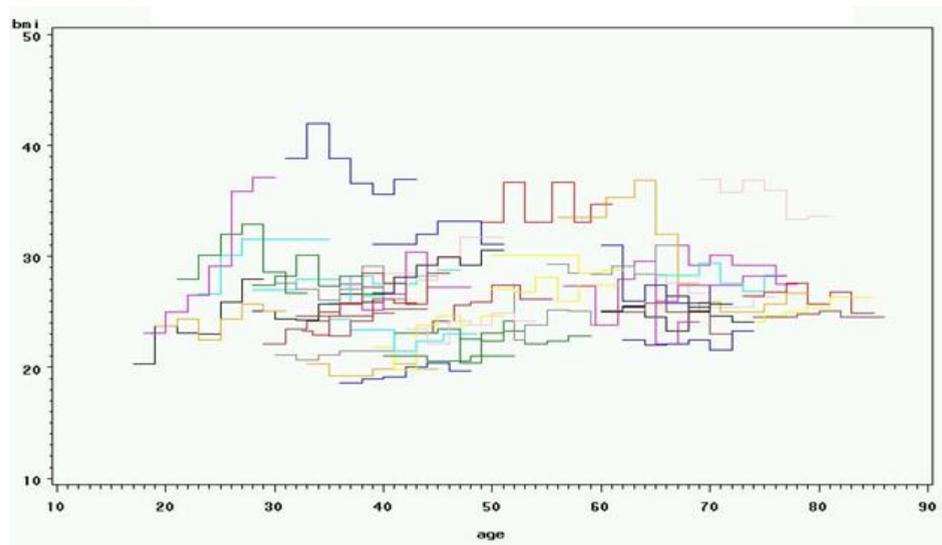
Meanwhile, education and smoking are shown in a different (blue) colour. This is to indicate our simplifying assumption that completed educational attainment does not

change over the life course after college age, and is “pre-ordained” at birth. Similarly, conditional on sex and educational attainment, the ages at smoking initiation and cessation are generated based on an analysis of the NPHS survey, and are also “pre-ordained” at birth. The two remaining state variables, obesity and coherence, are explicitly time-varying.

We should note that to prevent the diagram from becoming too cluttered, we have chosen to show only a few of the underlying arrows. For example, while at each period there are arrows from education to health, smoking, and education in the following time period; and to death during the intervening interval; no arrow is shown from education to obesity, even though there is a relationship.

In Figure 1 we have also simplified matters by implicitly assuming that time is discrete. The underlying data from the NPHS are based on repeated interviews every two years from 1994 to 2006 (the most recent year available at the time of this analysis). For example, Figure 2 shows a random sample of 50 NPHS respondent trajectories for BMI. However, as will be discussed below, while the first statistical estimation phase of our analysis necessarily uses the discrete biennial data from the NPHS, the second simulation modelling phase transforms the statistical estimates so that this part of the analysis is conducted in continuous time.

Figure 2: 50 randomly selected BMI trajectories



While we have refrained from drawing all of the possible arrows indicating causal pathways in Figure 1 in order to avoid clutter, our implicit “theory” is that everything has the potential to affect everything else. We therefore need the capacity to characterise from birth through life to death a complete biography for a sample individual, which would consist of a set of trajectories representing each of the state variables represented by the rows in Figure 1. In turn, the analysis must allow for the possibility that over time, there is reciprocal or mutual causation; i.e., that the variables co-evolve. Thus, the first statistical analysis phase must start with each variable being a dependent variable in one equation, and each variable must be given the chance to be an independent variable in the equations for the other co-evolving state variables.

For example, Figure 1 shows that obesity (BMI) is a dependent variable, the level of which at age a is a function both of its level at age $a-1$ and of the levels of the other variables at age $a-1$. At the same time, BMI is an independent variable for estimates of the transition dynamics of the other co-evolving variables.

While Figure 1 shows only arrows for a one period lag (i.e., a conventional first-order Markov assumption), the analysis allows up to two period lags (i.e., a second-order Markov; see below). The situation with health status is more complicated because we will be estimating individually the dynamics of all eight of the underlying health dimensions.

The result is a coherent and complete network of recursive equations. The network is complete because for every independent variable in this set of equations, there is one equation in which that variable is the dependent variable. All of the variables are available in the NPHS, so every equation in this network of recursive relationships can be estimated from the same data. As a result, the set of equations is coherent in terms of the variables’ definitions, and the measurement and error structures of the variables are identical from one estimated equation to the next.

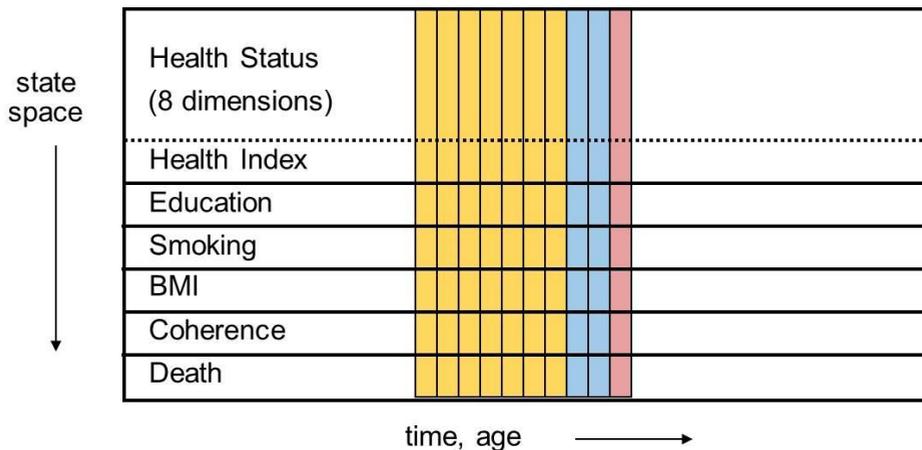
There are of course many other variables which could have been included. For example, income, leisure time physical activity, health care use, and chronic disease are all available on the NPHS; and all of these variables would likely have shown at least some statistically significant relationship to the variables which have been included. However, in this initial exploratory analysis, we are focusing on a small but diverse set of health determinants which have been deliberately chosen to span both proximal (smoking and BMI) and distal (education) factors, and both physical (BMI) and psychosocial (sense of coherence) factors. Still, as Krieger (1994) has pointed out (“where is the spider?”), the overall choice of which determinants are taken into account in constructing the causal web involves making some important judgements.

The network of recursive equations described above should prove extremely useful in assessing the relative importance of various factors as determinants of health. But it is important to note that by itself, the statistical estimation of all the coefficients for this

set of equations is incomplete. Simply looking at the coefficients of the estimated relationships will not tell us very much about the relative quantitative importance of the various factors being considered, given the complex intertwining of the pathways of influence indicated in Figure 1. Nor can the estimated relationships give us more than a suggestion of the effects on an aggregated life course measure such as HALE (health-adjusted life expectancy). It should be noted that we are *not* using the Sullivan method to estimate HALE. Rather, we are building HALE up explicitly from its micro foundations, one individual life course trajectory at a time. Thus, in this HealthPaths analysis HALE is a complex integration over all the individuals in a representative population, and over the lifetimes of each of those individuals.

The set of coherently estimated transition dynamics relationships from the first phase of the analysis, when combined with a Monte Carlo microsimulation model in the second phase of the analysis, meet both of these analytical objectives: i.e., constructing HALE and assessing the quantitative importance of various determining factors. The application of microsimulation in this context – while novel in social and health science – is reasonably straightforward, and is illustrated in Figure 3.

Figure 3: A synthetic/simulated individual biography



Our initial aim is to create one fully synthetic individual biography. A person is “born” (*in silico* in the memory of the computer) by being endowed with the values for all of his or her state variables in the leftmost coloured column of the figure. The set of equations indicated in Figure 1 is then applied recursively to generate all of the

individual's state space values for each time period in succession, thereby filling in the synthetic biography while moving to the right in Figure 3, until the individual dies.

For example, part-way through the synthesis of the individual biography shown in Figure 3, the values for the green and blue columns have been "filled in;" and the model is working to synthesise, in as realistic a manner as possible, the red column immediately to the right. The values for each cell in the red column are determined by the values for the two preceding time periods, shown by the blue columns; plus the coefficients of the estimated network of equations; plus many random number draws to determine whether a given transition occurs (see below).

This second-order Markov assumption—indicated by the colour coding showing the new red column that depends only on the pair of blue columns immediately to the left in Figure 3 – is a compromise. While higher-order terms (i.e., variables' values lagged three or more periods) would likely be statistically significant (e.g., Wolfson et al. 1993), this still represents a notable improvement on analyses which assume the relevant processes are only first-order Markov. Further, the second-order Markov assumption, based on the seven waves of data available from the NPHS from 1994 to 2006, allows for multiple observations of three time period sequences (three-tuples) of events for each individual, thereby allowing for the direct estimation of individual heterogeneity terms.

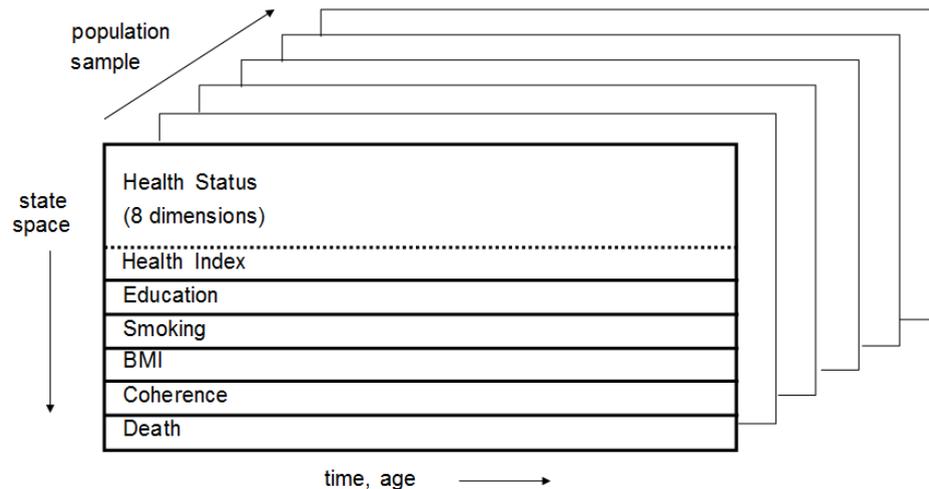
It should be noted, however, that Figures 1 and 3 are misleading for the simulation part of the analysis, because they give the impression that the simulation uses discrete time. In fact, the simulation is considerably more sophisticated in that it uses continuous time. Correspondingly, instead of the model generating changes in status that occur only at fixed annual (for example) time intervals, it generates event histories with discrete events, such as an improvement or a decline in a component domain of health status, which can occur at any time t , where t is a continuous variable. As a result, the statistical analysis, described below, forms a bridge between the discrete time data from the NPHS and the continuous time microsimulation modelling.

Once we have a complete set of empirically based stochastic descriptions of an individual's state transition dynamics which can be used to generate the individual's synthetic but realistic biography, the next step is simply to repeat the process shown in Figure 3 many times (millions in fact) in order to generate a complete representative sample of the Canadian population, as shown in Figure 4.

Individuals in any actual population have been born in a range of years. In other words, the population at any point in time is made up of individuals from a sequence of birth cohorts. However, for the analyses reported below, we focus on the cohort born in a single year, 1960. The population sample synthesised by HealthPaths for the 1960 birth cohort is typically on the order of two million. Once we have a complete population sample for this birth cohort, as in Figure 4, it is straightforward to cross-

tabulate the resulting set of synthetic biographies to generate the desired simulation results.

Figure 4: Population of synthetic biographies



For this analysis, the focus will be on the sum of the values in the “health index” row of Figures 3 and 4. If the index was always one, this sum would simply be each individual’s life length (in years). By using the health index values instead, the sum is the health-adjusted life length (HALL). Correspondingly, the average of all of the individuals’ life lengths (LLs) within a birth cohort is simply that cohort’s life expectancy (LE); while the average of all of the health-adjusted life lengths (HALLs) is health-adjusted life expectancy, or HALE, one of the fundamental indicators of population health.⁴

⁴ Note, however, that what is being estimated here is *cohort* HALE based on full individual- or micro-level life course trajectories. This is different from semi-aggregate *period* HALE, which is typically measured using the Sullivan (1971) method, as in, for example, Wolfson (1996) and McIntosh et al. (2009). See below for further discussion.

2. Statistical analysis

We were able to take a unique approach to the statistical estimation for this analysis due to the existence of bootstrap weights in the main data set used, the National Population Health Survey (NPHS; Yeo et al. 1999). In order to allow users to account for the complex multi-stage clustered design of the NPHS sample, each individual observation has been given a set of 500 bootstrap sample weights. Ordinarily, these are used to estimate more accurate sampling errors for any given statistic derived from the sample. These sampling errors are typically larger than those derived under the assumption that the data had come from a simple random sample (the ratio of the bootstrap sampling variances to the simple random sample sampling variance is termed the “design effect” of the complex sample). For example, the sampling error of a number in one cell of a cross-tabulation, or of a coefficient in a regression, can be simply generated (albeit via intensive computation) by running the tabulation or the regression 500 times: once using each set of bootstrap weights, and then post facto computing the variance of the 500 estimates.

In this analysis, we have used these bootstrap weights in a highly novel fashion, as described in Rowe and Binder (2008). We have used the bootstrap weights not only to determine the variances of the estimated coefficients in each of the regressions in the system of equations outlined above, but also to estimate a form of specification error, and, ultimately, the variances of our simulations. Each regression equation in the coherent network of relationships indicated by Figure 1 was in fact estimated 40 times, once with each of 40 bootstrap weights (sampled from the set of 500 available).

The specification of the equations in the model was developed in several steps. First, for functional health status, measured by the McMaster health utility index (HUI; Feeny et al. 2001), the focus was on changes from one survey wave to the next. The HUI has questions on eight health dimensions: vision, hearing, speech, mobility, dexterity, emotion, and pain. Health status on each of these eight dimensions was assessed on the NPHS using an ordered categorical variable. Given the two-year interval between NPHS interviews, it is possible for multiple (unobserved) changes to have occurred between survey waves. In light of the length of this interval and the categorical nature of the measurement, a conservative use of the data would be to identify two types of event. For example, in the case of the vision dimension:

Improved Vision: $\text{Vision}_{t+2} > \text{Vision}_t$

Declining Vision: $\text{Vision}_{t+2} < \text{Vision}_t$

where Vision_t and Vision_{t+2} are the categories obtained from responses in two successive interviews spaced two years apart. Given this approach, two separate hazard

equations corresponding to these two types of (competing) events were specified for each of the eight functional health status dimensions, resulting in a total of 16 equations.

As we noted earlier, there is considerable evidence that event (transition) probabilities vary not only with the current state, but also with the duration spent in a given state (e.g., Wolfson et al. 1993); i.e., that the underlying processes are not first-order Markov. As a partial response, the longitudinal data extracted from the NPHS were arranged into sequences of three consecutive interviews. This allowed for functional health transition hazards (occurring between times t and $t+2$) to be conditioned on current (t) and lagged ($t-2$) functional health categories, as well as on the other four covariates (BMI, smoking, education, and sense of coherence); a second-order Markov assumption. With seven waves of the panel survey from 1994 to 2006, as many as five such three-tuple sequences were available from each respondent, providing in total more than 50,000 such sequences.⁵

Like Cox regression, the equations estimated assumed proportional hazards, but the estimation strategy was more general than Cox regression, since it involved estimating baseline hazards jointly with covariate relative risks. As outlined by Lawless (1987), this involves the use of a full conditional Poisson likelihood (L) for the dependent event counts y (e.g., representing either improving or declining functional health events), rather than the “partial” likelihood employed in Cox regression. The resulting likelihood equations are shown below, where “ i ” denotes a respondent and “ t ” a time period, 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), and X_{it} represents a vector of time-varying covariates with coefficient vector β . Finally, there is a person-specific, Gaussian random effect z (which accounts for the correlation between observations drawn from the same respondent and otherwise unobserved person-specific factors), with σ representing the estimated variance associated with those person-specific effects.

⁵ Even though the NPHS collected data up to 2010, at the time this analysis commenced, only data up to the 2006 wave were available.

Likelihood equations:

$$y_{it} = PY_{it} \exp (X_{it}\beta + \sigma z_i) + \varepsilon_{it}$$

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

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

$$\hat{\beta} \approx \arg \min_{\beta} [-\sum_{it} \log(L_{it}) + \lambda (\sum_{j=1}^J |\beta_j|)]$$

The estimation of the coefficient vector was not simply a matter of maximising the specified Poisson likelihood. Each of the 16 hazard equations (eight health dimensions times two for improvement or decline in each health dimension) had 1,071 terms (not including the intercept). This large number of terms arose in a natural way not only from the covariates, but also from the inclusion of an extensive range of plausible covariate interaction terms. These 1,071 terms were:

- (a) concurrent and lagged effects for each functional health dimension on each upward or downward transition (across five or six levels for a total of 81 terms);
- (b) additional BMI, smoking, and coherence covariates (13 terms);
- (c) age splines plus age interactions with each of the terms in (a) and (b) above (174 terms); and
- (d) further interactions for all of the terms above with (1) sex, (2) education less than secondary school, and (3) university education (268 terms each).

To help us cope with an estimation with a large number of terms in each regression equation, we added a penalty to the likelihood in the form of a weighted sum of the absolute values of coefficient estimates. Penalties of this kind will generally result in estimated coefficient vectors with a minimal number of non-zero estimates (Tibshirani 1996, 1997). Increasing the penalty weight (λ) either increases the number of elements of β assigned a value of exactly zero, or it shrinks non-zero elements of β towards zero. In effect, the estimation of each regression equation was “open” with regard to the specification. In each instance, the choice of an appropriate value of λ was made by minimising out-of-sample prediction error.

(The dynamics for the other variables in the model – BMI, smoking, coherence, education, and mortality – were simpler, as we will describe.)

As we noted earlier, we were able to take a unique approach to the statistical estimation for this analysis due to the existence of bootstrap weights in the main data set, the National Population Health Survey (NPHS). A novel algorithm was adapted to

deal with as many as 1,000 right-hand-side (RHS) variables in these 16 equations: namely, the Generalized Forward Stagewise Regression, modified from Hastie et al. (2007). The basic steps of this estimation for any one equation were as follows:

1. Start by initialising $\beta_0, \beta_1, \beta_2, \dots, \beta_p = 0$ and utilise centred/standardised covariates X .
2. Find the index “m” corresponding to the largest score $[|\delta \ln(L) / \delta \beta_m|]$ evaluated at the current predictor Y .
3. Update $\beta_m = \beta_m + \varepsilon * \text{sign}(\delta \ln(L) / \delta \beta_m)$, where ε is small (0.005 in this application).
4. Update the intercept β_0 and the predictor Y .
5. Repeat steps 2-4 many times, testing each time for termination by evaluating at each step “s”, the predictive likelihood L_s using an “out-of -sample” subset of the data reserved for testing, with termination as soon as the likelihood starts decreasing ($L_s < L_{s-1}$). The out-of-sample observations are precisely those with zero weights for a given bootstrap weight vector (see below).
6. On termination, the estimates $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are back-transformed to adjust for covariate standardisation.

We used a Poisson likelihood (L) because it is consistent with familiar proportional hazards assumptions. Coefficient estimates evolve with successive steps of this algorithm, and its termination effectively chooses an equation specification (i.e., which β 's end up being non-zero).

Due to the nature of each vector of the bootstrap weights, the weights are non-zero for about 63% of the observations. [If n observations are sampled with replacement and with probability $1/n$, then the probability that any one observation is NOT included in the sample is $(1-1/n)^n$ which for large n approaches $e^{-1} \approx 0.368$]. As a result, at each iteration in which the vector of β 's is updated for each equation, it was a straightforward (although computationally intensive) process to determine the out-of-sample prediction error using the remaining 37% of the observations which had bootstrap sample weights of zero, and which therefore had not influenced the estimation. These out-of-sample prediction errors were incorporated into the overall estimation algorithm by terminating the iterations when the prediction error began to rise, marking the point at which the loss of predictive accuracy due to variance in coefficient estimates outweighs the benefit of further refining the estimates.

For each of the 16 functional health equations being estimated (the probability of an increase and of a decrease in each of eight health status dimensions), and for each of the 40 randomly sampled bootstrap weight vectors, the estimation process was iterated independently in order to determine which subset of RHS variables was most important.

Thus, the RHS variables in each regression did not have to be the same in all 40 of the alternative bootstrap estimation processes, and indeed they often were not. The health status regressions tended to have fewer than 40 RHS variables with non-zero coefficients from among the more than 1,000 candidate variables. The final result was 16 sets of β 's: one set for each equation, times 40 for each of the 40 bootstrap weight vectors used. These 640 equations for health status dynamics correspond to the first row of Figure 1.⁶

The equations for the simulation of both BMI and coherence, the fourth and fifth rows of “blobs” in Figure 1, utilised a common set of covariates and a common form of equation (Y represents BMI or coherence, as appropriate):

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

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

$$\begin{aligned} A \text{ is the median equation:} & \quad A = D \theta_A \\ B \text{ is a dispersion equation:} & \quad B = \max(0.0, D \theta_B) \\ G \text{ is an asymmetry equation:} & \quad G = D \theta_G \end{aligned}$$

and 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 attributes of the distributions (location, dispersion, and asymmetry) depending on covariates 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 generalisations 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 has been fixed at birth for each simulated individual. This was done because the coherence scale was available only from the first two waves of the NPHS, and lagged correlations could not be reliably estimated. The estimation was carried out using SAS Proc NLMIXED to facilitate the specification of u as standard Gaussian.

⁶ Recall that while the estimation is using discrete time survey data (waves every two years), the resulting equations are used in the simulation in a continuous time manner. As a result, more than one transition in the simulation phase can occur within any given two-year interval.

The method for dealing with the 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 a linear least-squares regression of BMI and coherence on the candidate covariates. In a succession of steps, the candidate covariates were eliminated one at a time if the associated p-value was above 0.05. The estimation was carried out using SAS Proc SURVEYREG to ensure that the significance tests took into account 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. In their final form, each of the three equations had the following covariate terms: age, years of daily smoking, time-varying education, HUI, and sex.

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, the simulated individuals are limited to one smoking period over the course of their lives, and the 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 the never, current, and former daily smoker categories, as well as the years of daily smoking covariate.

The daily smoking model utilises NPHS data on the (possibly censored) age at first daily smoking for respondents aged 25 or older, and on the (possibly censored) duration of daily smoking for respondents aged 50 or older. We fitted 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.

Another relatively novel feature in the estimation process was the explicit consideration of individual heterogeneity. By restricting our specifications for the dynamics of the co-evolving variables to second-order Markov (i.e., only one and two period-lagged independent variables covering three time periods overall), we could observe up to five of these “triples” for each individual. An individual fixed effect, which likely proxies for omitted variables, was therefore straightforwardly included in the estimation process after the model selection stage, described above, was completed.

One way to assess the overall contribution of covariates to prediction is to examine the frequencies of non-zero β 's from the estimation process. We may recall that by using 40 bootstrap weight vectors, we have effectively included a degree of model selection. Given the large number of interaction terms in the functional health status (HUI) equations, Table 1 first provides a count of the number of terms in which each covariate is involved, as either a main effect or an interaction for each of the candidate covariates. While there are only 1,072 terms in each equation, the total number of terms shown is 1,600, due to double counting.

Table 1: Number of terms involving each candidate covariate (includes double counting)

Age	BMI	Smoking	Males	Education	Coherence
696	28	48	268	268	24

Table 2 displays the average frequency of the selection of terms involving each candidate covariate (evaluated over 40 bootstrap model selection samples). The results indicate that age was most often chosen (despite the fact that it ranks low relative to the potential number of terms). Coherence appears to be a strong covariate, which is surprising given its relative novelty. In terms of frequency of selection, coherence is on par with more familiar covariates like smoking and BMI. Education has the most marked pattern: i.e., it plays an important role in “getting better” health events, but is never involved in “becoming worse” health events.

Table 2: Average frequency of covariate selection: Functional health status equations (40 bootstrap replicates, blanks indicate cases in which covariates were never selected)

	Improving Health						
	Age	BMI	Smoking	Male	Education <SS	Education ≥ BA	Coherence
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

Table 2: (Continued)

	Declining Health					Coherence
	Age	BMI	Smoking	Male	Education <SS	
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

It should be noted that for the education results, each estimated coefficient is a direct partial effect distinct from the other (correlated) covariates in the equations. Educational attainment, as with sex, will tend to be time-invariant over the range of ages most important to health change; and some of the impact it might have had will inevitably be taken up by lagged functional health status effects.

3. Simulation methods

The Monte Carlo simulation process can now be described more concretely in terms of Figure 3 above. As was already noted, the model actually operates in continuous time, and uses the ModGen simulation environment (Statistics Canada, no date).

Each individual's biography was built up recursively. At age zero, the individual was initialised as a non-smoker, with normal BMI and coherence, and full health. (Note that immigration is ignored).

Lifetime educational attainment was drawn once from the sex-specific distribution observed for the given birth cohort (1960 in this case) based on 2001 census data. We used a simplified process in which, depending on the individual's ultimate educational attainment, there could be transitions at ages 18 (to completed secondary) and 22 (to completed post-secondary or higher), with this variable then remaining constant over the rest of the lifetime.

Smoking trajectories were defined by two variables: age at initiation and age at cessation. These were determined at birth, conditional on sex, birth cohort, and ultimate educational attainment; and were based on NPHS longitudinal data.

Finally, each individual was endowed with his or her vector of “heterogeneity coefficients,” drawn randomly from distributions observed as part of the estimation of transition dynamics, for each of the eight health status dimensions, and for BMI and for coherence.

Using these starting data, the values for the individual’s state variables for the next period were simulated recursively, and one by one. However, this process was complex because it was occurring in continuous time. Some events are special cases of continuous time, such as ageing one year, which is simply modelled as an event which occurs exactly one year after the previous birthday. But other events are characterised by waiting times. The major benefit of this approach is that competing risks can be handled naturally. In particular, the regressions sketched above, and described in more detail in Rowe and Binder (2008) for the health state transitions, generate hazard rates. In turn, these hazard rates are transformed into waiting time distributions. The Monte Carlo simulation process makes uniform random draws from the unit interval, and applies these to the inverted waiting time distribution to generate an event-specific waiting time.⁷

The ModGen microsimulation environment (Statistics Canada, no date) maintains a queue of all the possible future events, ordered from soonest to latest. Generally, the simulation advances the “clock” to the next event in this queue, and makes the event occur. Furthermore, this change in state may affect the independent (right-hand side, RHS) variables that were used to compute other waiting times. ModGen keeps track of all of these inter-dependencies, re-computes all of the affected waiting times, and updates the ordering of anticipated events in the event queue. The ModGen clock then advances to the next event in the queue, implements that event, re-computes any dependent waiting times as needed, reorders the event queue, and continues working through the event queue until the individual dies.

Given this general description of the recursive simulation process, the specifics of our HealthPaths model are generally as follows:

⁷ More specifically, the piecewise exponential distribution is the waiting time counterpart to the piecewise Poisson incidence distribution, in as much as the defining parameter of an exponential distribution is the incidence hazard. (Piecewise distributions are needed, for example, to account for shifting hazards related to integer age, where age is included as a lagged time-varying covariate; i.e., the age at the beginning of each interval during which an event may occur). The complementarity between the exponential waiting time distribution and the Poisson incidence distribution can be characterised as follows: the Poisson represents the chance distribution of event frequencies occurring within a fixed time interval, while the exponential represents the time interval distribution between two chance events. Cox regression is one component of a Poisson process model that is conditional on observed event times. A complete Poisson process model – which is required for simulation – also needs to have a component representing the time pattern of events (duration dependence). (Lawless 1987).

- (a) Increment age by one period by creating a “one year older” event in the continuous time event queue.
- (b) Update as appropriate education per the initial value of lifetime educational attainment imputed at birth and at ages 18 and 22.
- (c) Update as appropriate the smoking status based on the ages at initiation and cessation imputed at birth.
- (d) At the time of each birthday event, impute new BMI and coherence values as functions of all of the included RHS variable values, plus the endowed heterogeneity coefficients which determine how smooth the individual trajectories will be.
- (e) Draw random waiting times to an increase or a decrease in each dimension of the individual’s health status – i.e., 16 waiting times (as appropriate; for example, vision cannot improve if it is already at its maximum) as a function of all the included RHS variable values – plus the endowed heterogeneity coefficient for each dimension of health status.
- (f) Based on the event queue, move the clock forward to the next event with the shortest waiting time, which could be an improvement or a decrement on one of the eight health status dimensions; or to a birthday, smoking, or mortality event.
- (g) Compute the individual-level summary health status index (i.e., the number in the unit interval based on the McMaster HUI scoring function) from the eight HUI dimensions for the interval between birthdays for the given individual in the given year (note that this could involve an average of more than one level of the index if the individual’s health status changed during this time interval), not only because it will be a key output variable, but also because it affects the RHS for BMI and coherence.
- (h) Finally, each time an integer age or HUI changes, re-compute the waiting time probability distribution of death (which depends on the summary health status index – i.e., the individual’s HUI score—not on each of the underlying health status dimensions), draw a random waiting time to death, and add the updated death event waiting time to the event queue.

This recursive loop serves to generate one individual’s complete synthetic biography. The simulation model then repeats this process two million times to generate a representative sample of the population. The summary results are then tabulated, in particular the average health-adjusted life length (HALL = integral of each simulated individual’s HUI scores from birth to death), as well as the nine decile cut-points in the distribution of these HALLs.

But there is one more step. Recall that all of the underlying regressions for the dynamics were estimated from the NPHS 40 times, once for each of a random subset of the 500 bootstrap weights provided with the survey. Thus, the entire simulation process was repeated 40 times, once for each bootstrap replicate of the estimation of the complete network of equations.

As noted above, each estimation process independently determined which candidate independent variables – out of more than 1,000 possibilities – actually have non-zero coefficients; i.e., actually enter into the specification. As a result, the variation across the 40 bootstrap replicates of the simulation is much more than a conventional assessment of the sampling variability of the equation coefficients themselves, based on their statistical significance. It also includes the variability of the regression *specifications* based on out-of-sample predictions. To our knowledge, this is the first time such an extensive account of the error structure of a system of equations has been formally incorporated into a microsimulation analysis.

Finally, it should be noted that the data for education, smoking, and mortality were drawn from a variety of sources other than the NPHS, and represent the 1960 birth cohort. But the data for health status, BMI, and coherence were drawn from the 1994 to 2006 NPHS; and due to a lack of appropriate data, these factors have not been backcast for the years between 1960 and 1993, or projected into the future. Thus, the “1960 birth cohort” being simulated represents a blend of some actual cohort data and “thick” period data centred on the year 2000.⁸

4. Results

4.1 Base case scenario

The main focus of our simulation results is the distribution of individuals’ health-adjusted life lengths (HALLs). The average of the HALLs is simply HALE (health-adjusted life expectancy). Figure 5 shows the distribution of HALLs in the base case simulation. It should be emphasised that these are not predictions of the HALLs of the

⁸ In standard period life tables, including multi-state life tables, “period” refers to the fact that the transition rates all pertain to the same year or the same quinquennium; i.e., the period may be either one year or five years “thick.” In our case, we are using transition dynamics estimated from the NPHS over the 1994 to 2006 period, so we refer to this as a 12-year “thick” period. Further, conventional (multistate) life tables use scalars as transition rates; e.g., an age- and sex-specific mortality rate. In our case, we are in effect using highly multivariate conditional transition probability density functions. However, because we are modelling in continuous time, these transition probabilities are instead represented by conditional waiting time distributions for discrete events

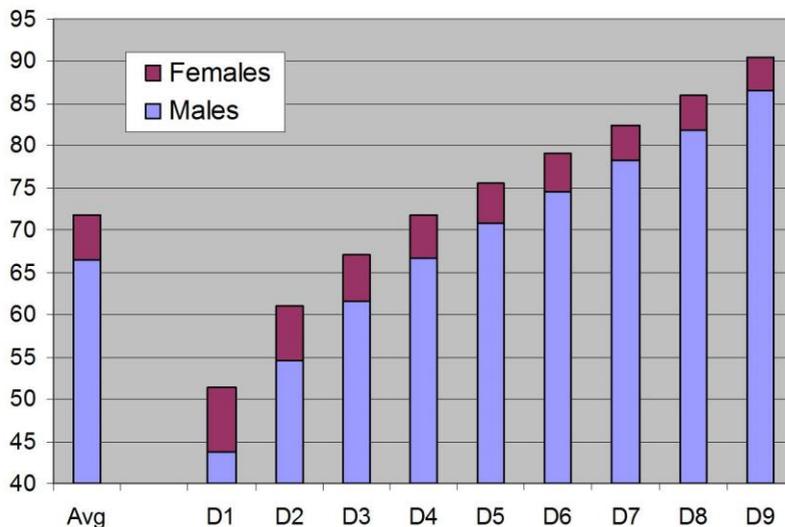
1960 birth cohort. Rather, these estimates are analogous to a “thick” period cohort analysis, and represent an initial exploratory effort.

Overall, the simulation suggests that women in the 1960 birth cohort in Canada can expect to have a HALL of 71.9 years, while men can expect to have a HALL of 66.5 years (leftmost stacked bar).⁹ The nine stacked bars to the right show the decile cut-points of the distribution of these HALLs.

It should be noted that if the pattern in Figure 5 were rotated clockwise 90 degrees, the ends of the bars would in effect trace out the estimated health-adjusted cohort survival curves for men and women born in 1960. Thus, the variation in HALLs does not appear to be unusual. Indeed, when compared to the recent results in McIntosh et al. (2009), they look broadly similar.

The sample sizes of the 40 bootstrap replicates are such that the Monte Carlo standard deviations are small relative to the results of interest, on the order of 0.1 or 0.2 years. This is by design; the smallest samples (i.e., two million cases) were used (to minimise execution time), but subject to achieving a Monte Carlo error not exceeding this magnitude.

Figure 5: Simulated distribution of health-adjusted life lengths, 1960 birth cohort



⁹ Note that the portion of the bars labelled “female” is actually the difference between females and males, both here in Figure 5 and in Figure 8 below.

However, the combined Monte Carlo + sampling + cross equation correlated + out-of-sample prediction errors are an order of magnitude larger, as shown in Table 3. On the one hand, these standard deviations (SDs) appear to be large relative to the magnitudes of interest. On the other, they account for far more sources of variability than is conventional. In particular, these SDs include specification errors estimated by out-of-sample predictions across 40 bootstrap replicates. It is worth noting that the vast majority of microsimulation model-based analyses fail to account for such variation. (An exception is Will et al. 2001).

Table 3: Average health adjusted life lengths (HALLs) and bootstrap standard deviations for the 1960 birth cohort, overall and for decile cut-points (years)

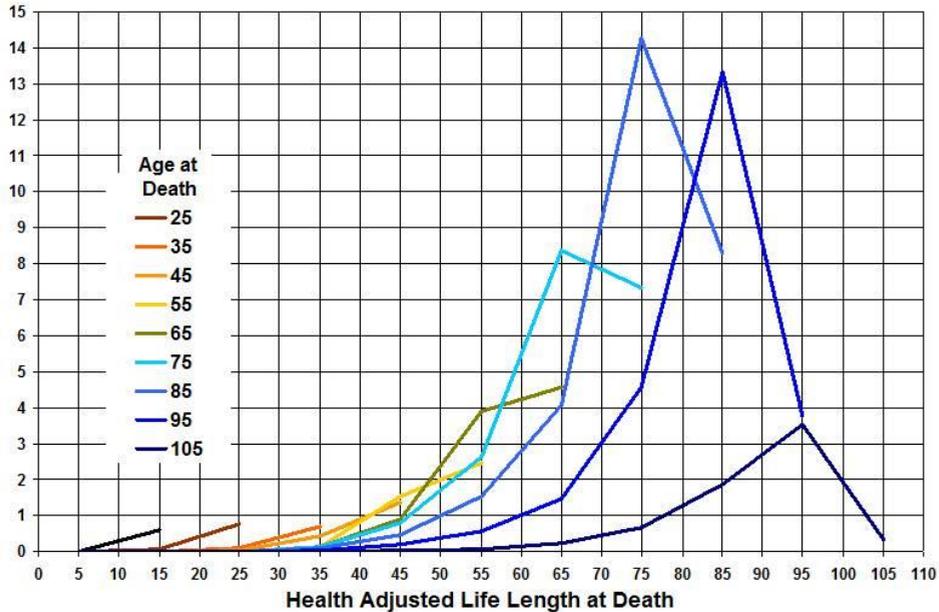
	HALLs		Bootstrap Standard Deviations	
	Females	Males	Females	Males
Average	71.9	66.5	3.2	3.9
D1	51.3	43.7	3.6	3.8
D2	60.9	54.5	4.4	5.2
D3	67.1	61.5	5.0	6.0
D4	71.7	66.7	5.0	6.4
D5	75.6	70.8	4.4	5.9
D6	79.1	74.5	3.8	4.9
D7	82.5	78.2	3.2	3.9
D8	86.0	81.9	2.7	3.1
D9	90.4	86.5	2.2	2.4

The underlying joint distribution of life lengths (LLs) and health-adjusted life lengths (HALLs) is shown in Figure 6. For both LLs and HALLs, the ages were grouped into decades. Each curve in Figure 6 is for one set of decadal LLs, labelled for the midpoint of the decade. For example, the curve labelled 95 is for all of the simulated individuals in the 1960 birth cohort who have been simulated to die at an exact age between 90 and 100. (Recall that the simulation is in continuous time.) The height of each curve at each decade's midpoint is the percentage of the total cohort population (i.e., all LLs) whose HALL was in that decade. So in this case, just over 13% of the cohort are projected to have LLs between 90 and 100, and HALLs between 80 and 90.

For deaths at ages in the 70 to 80, 80 to 90, 90 to 100, and 100 to 110 intervals, the mode of the curve is 10 years less than the age at death. Thus, most of the individuals

dying at these decadal ages had HALLs which were about 10 years shorter than their LLs. But for those who were dying at younger ages, the modal HALL was the same as the LL.¹⁰

Figure 6: Distributions (%) of health-adjusted life lengths by attained life lengths (nearest decades), 1960 birth cohort



4.2 Simulation experiments

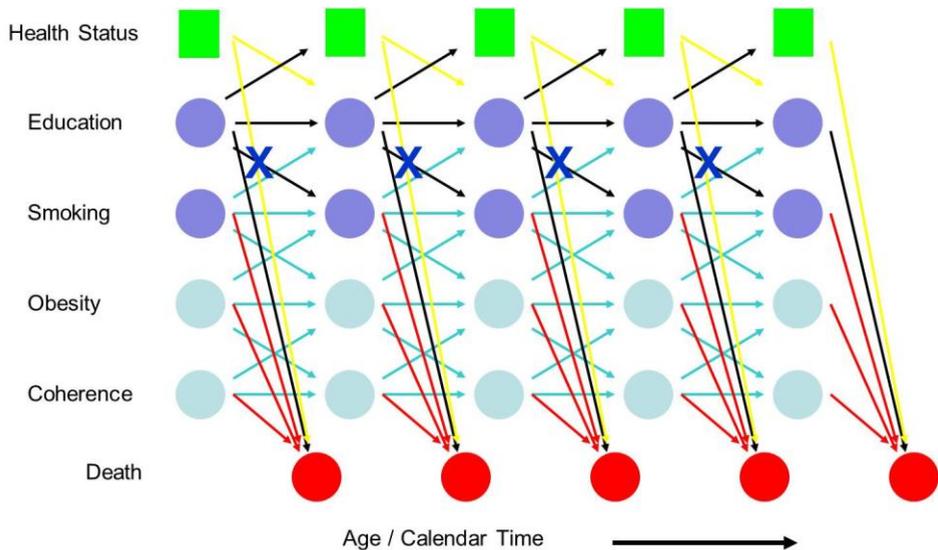
The HealthPaths microsimulation model results just described provide a foundation for a much more extensive analysis. We have now in effect constructed a laboratory within which we can do experiments on the estimated causal web shown in Figure 1 above. The essence of these experiments is to “snip” or modify various connections (shown as arrows in Figure 1) in order to estimate the quantitative importance of various pathways as determinants of the level and distribution of the HALLs just shown in Figures 5 and

¹⁰ Note that the graph is somewhat distorted as an artefact of using decades as the time interval, and then straight lines to connect the data points at mid-points of the decadal intervals.

6. This approach is analogous to the “knockout gene” models used in bio-medical research, except that instead of knocking out a sequence of nucleic acids in an *in vivo* experiment, we are knocking out (zeroing, for example) various coefficients in a series of *in silico* experiments. Alternatively, we can think of these experiments as allowing for the construction of counterfactual worlds; i.e., a world in which we can see what would happen if a given causal pathway differed from the causal pathway observed in our combined statistical estimation and (micro) simulation.

This idea is illustrated in Figure 7, in which a series of Xs indicate that the direct effect of education on smoking is set to zero over the full individual life course. It should again be noted that, as in Figure 1, most of the actual arrows are omitted to prevent the diagram from being too cluttered.

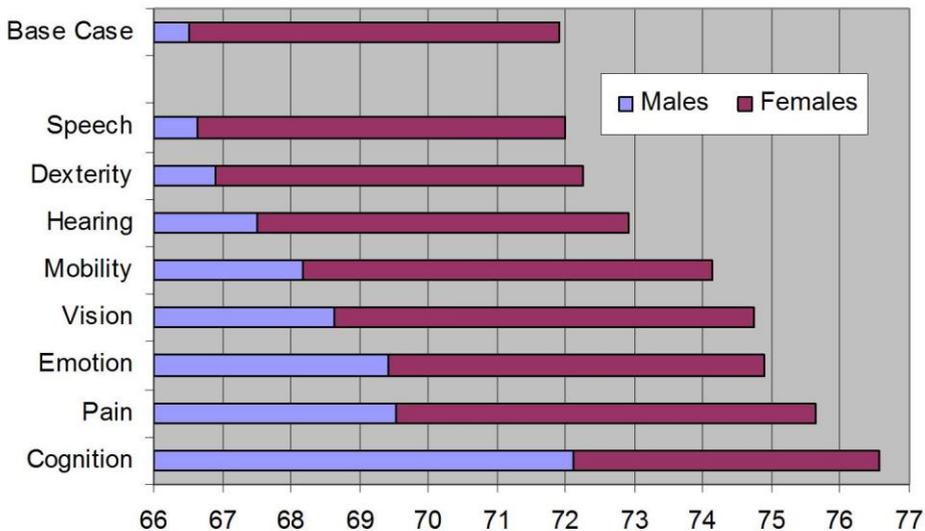
Figure 7: Counterfactual (“knockout”) system of recursive equations (n.b. not all possible arrows shown)



The experiment sketched in Figure 7 is highly simplified. It is clear from this diagram that education affects health status and mortality through many paths, including more circuitously, and in ways that depend on the magnitudes and signs of the myriad coefficients in the estimated network of recursive equations. As a result, a somewhat different set of counterfactual experiments will be presented.

We start with a relatively simple experiment. Figure 8 shows the relative importance of each of the eight dimensions of the McMaster health utility index (HUI). The top pair of stacked bars show the average of the HALLs for men and women, exactly as in Figure 5. Each of the other stacked bars shows the hypothetical average HALLs for the 1960 birth cohort *if* each of the eight domains were, one at a time, forced to their top (i.e., most healthy) levels.

Figure 8: Health status (HUI) attribute-deleted HALE



The results reflect a mixture of the prevalence of less than full health on each dimension observed in the NPHS, and the weight given to each dimension in the (non-linear) scoring function used to compute the summary HUI (Feeny et al. 2001). Figure 8 clearly shows that speech and dexterity problems are of relatively minor importance. Hearing, mobility, and vision are of moderate importance, as the absence of any problems in these dimensions is estimated to increase the HALLs of Canadians born in the 1960s by roughly two years. (e.g., for males, HALE without any vision problems is about 68.5 years; while HALE in the base case is about 66.5 years, for a difference of two years). The most important dimensions appear to be emotion (depression, anxiety), pain, and especially cognition. Absence of any problems in these dimensions would

increase the HALLs of hypothetical 1960 Canadian male and female birth cohorts by roughly three years for emotion and pain, and by 4.5 to 5.5 years for cognition.¹¹

These latter results are roughly twice the impacts of the two leading causes of death in Canada, ischemic heart disease (IHD) and lung cancer. In terms of cause-deleted (n.b. period, not cohort) life expectancy, eliminating IHD deaths would add 1.8 to 2.4 years, while eliminating lung cancer deaths would add 0.7 to 1.0 years; in each case for women and men, respectively (Manuel et al. 2002).

While these counterfactual analyses examining the functional health status components of HALE (based on the HUI) have not fully exploited the experimental machinery of the HealthPaths microsimulation model, they set the stage for later results. The importance of the cognitive dimension is clearly noteworthy.

One of the most pervasive findings is the relationship between socio-economic status and health. Because of the way we have estimated the network of relationships in Figure 1, and due to the construction of the HealthPaths model, constructing an initial hypothetical pair of counterfactual scenarios to assess the quantitative importance of this relationship in a full life-cycle context is straightforward. We simply create a scenario in which everyone always has “low” educational attainment (less than secondary school completed), and another scenario in which everyone has “high” educational attainment (completed post-secondary BA or more).

Figure 9 shows the results along with the base case scenario. The results are in the expected direction, and are relatively uniform across HALL deciles: higher educational attainment increases HALLs across the board; while lower education reduces HALLs. However, the magnitudes of the changes are different: higher education increases HALL by about one to two years compared to the base case education scenario; while lower educations reduces HALL by two to six years.

In Figure 10, we can see the results of hypothetical scenarios in which no one in the 1960 birth cohort ever smokes. In these cases, the patterns of improvements in HALL are differently distributed than for the high and low education scenarios. In the lower deciles, the no smoking scenario increases HALL by as much as five to 10 years;

¹¹ These results differ considerably from the direct estimates of “cause-deleted” HALE in Wolfson (1996). In that analysis, HALE was calculated using the Sullivan method, and vision, hearing, and speech were combined into a single “sensory” dimension. The relative importance of the pain, emotion, mobility, and dexterity dimensions was similar to that shown here. But the aggregated sensory dimension had a much greater impact, while cognition was in the middle. Moreover, the absolute magnitudes of the impacts on HALE were much smaller than those shown here: i.e., about one year in the middle of the range in the 1996 Sullivan method analysis, compared to about three years here. However, this earlier analysis had no cross-equation interactions at all among the eight health status dimensions, nor any correlation of health status with mortality, nor any persistence of each of the health status dimensions at the micro or individual level. The results of that earlier analysis are therefore not really comparable to those found here, especially as the omission of these important factors may have led to a serious underestimation of the impacts. Nevertheless, the differences do demonstrate the importance of further probing the new simulation results presented here.

while in the higher deciles, the improvements are more modest, in the two- to four-year range. The likely explanation for this finding is that because individuals in the higher HALL deciles are more likely to be non-smokers, the elimination of the smoking scenario has less effect on those in the higher deciles.

Figure 9: “What if” education effects by HALE (years) decile

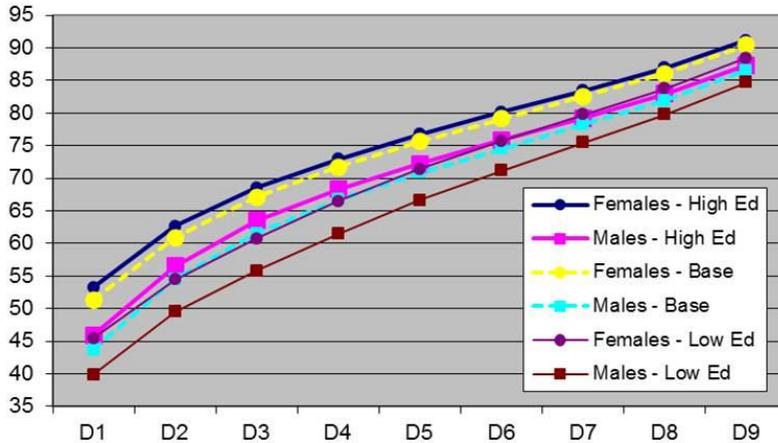
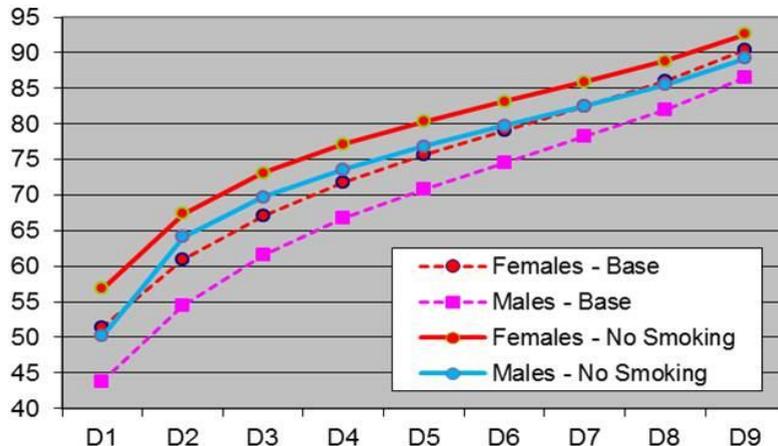
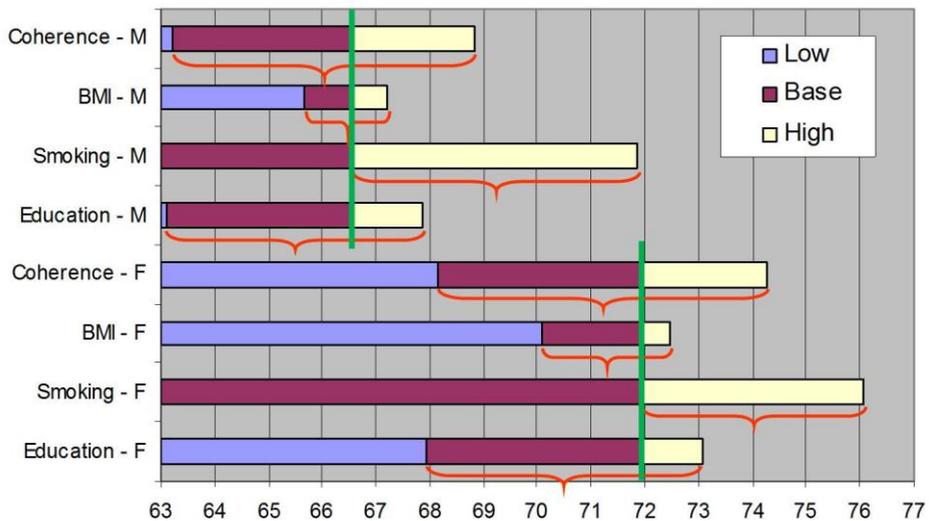


Figure 10: “What if” smoking effects by HALE (years) decile



Finally, Figure 11 and Table 4 show the effects on average HALLs of the two sets of scenarios just shown, education and smoking; plus two further sets of scenarios. These latter two scenarios focus on BMI and on Antonovsky’s sense of coherence. In each of these two cases, the fifth and 95th percentiles of the values in the population were determined. For BMI, these were 19 (low-normal weight) and 33 (obese).¹² Then counter-factual scenarios were simulated where everyone was forced to be at the 5th percentile of BMI all their lives, then at the 95th percentile of BMI, then at the 5th percentile of the Sense of Coherence scale, and then at the 95th percentile – all for males and females separately. In the case of smoking, no “high” scenario is shown because its specification would be highly arbitrary. For both men and women, the vertical green lines indicate baseline HALL.

Figure 11: “What if” comparative knockout effects on HALE of four health determinants



As was already shown in Figure 9 above, the no smoking scenario had a greater impact on HALLs than the high and low education scenarios for some deciles. Figure 11 and Table 4 indicate that on average, the no smoking scenario raises average HALLs

¹² There is general consensus that the relationship between BMI and various measures of health is “J shaped;” i.e., that there are adverse health effects associated with both very high and very low weights (BMIs). The WHO “normal” BMI range, in which individuals are supposed to be at the lowest level of risk in terms of their BMI, is from 18.5 to 25.

by about five years for men and four years for women. The overall difference in average HALLs between the low and high education scenarios is also about five years for men and for women. Thus, the impacts of education are on average at least as large as those for eliminating smoking.

Table 4: HALE data (years) for Figure 11 plus bootstrap standard deviations

	Levels		Bootstrap SD's	
	Females	Males	Females	Males
Baseline	71.9	66.5	3.2	3.9
Low Coherence	68.1	63.2	2.9	3.4
High Coherence	74.3	68.8	3.4	4.0
Low BMI	72.5	67.2	3.2	3.9
High BMI	70.1	65.7	2.9	3.7
No Smoking	76.1	71.8	0.6	0.5
Low Education	67.9	63.1	4.9	4.8
High Education	73.1	67.8	3.3	4.2

Still, because the standard deviations estimated from the 40 bootstrap replicates were generally rather wide, the simulated differences were not always significant. Nevertheless – and perhaps surprisingly given the large amount of attention given to obesity by public policy makers – the quantitative impact of the BMI scenarios was found to be the smallest: i.e., about 1.5 years for men, and about 2.5 years for women.

On the other hand, the impacts of Antonovsky's sense of coherence – or the underlying factors for which it is a marker – were the largest of the four sets of scenarios examined, amounting to over 5.5 years for men and over six years for women. Of course, if the high and low scenarios for coherence had been compared to the difference between a “high” smoking scenario and the elimination of smoking, smoking would likely have had the largest impact. This result for coherence calls for further investigation. One possible reason for this result could be the general importance of cognitive factors in the overall measure of health status shown in Figure 8 above.

5. Summary and concluding comments

The initial analysis described here using the HealthPaths model is exploratory and novel. We have used the longitudinal data from the NPHS in a much more powerful way than usual:

- ◆ estimating not just a disconnected series of empirical relationships, but a coherent network of transition dynamics equations for Canadians' health status;
- ◆ using the bootstrap weights provided with the NPHS for variance estimation (due to its complex sample design) in a novel fashion to account explicitly not only for sampling error, but also for a substantial degree of specification error;
- ◆ using a second (higher)-order Markov assumption for estimating transition dynamics;
- ◆ specifying highly multivariate conditional transition probability functions;
- ◆ applying these transition dynamics in continuous time in order to reflect competing events in a natural manner;
- ◆ fitting based on out of sample prediction errors rather than more conventional goodness-of-fit statistics;
- ◆ using a machine-learning approach, and hence more open equation specifications;
- ◆ taking individual heterogeneity into account; and, finally,
- ◆ constructing a microsimulation model which builds upon and integrates the network of transition dynamics equations in order to conduct a series of experiments.

These experiments have allowed us to quantify, at least in an initial exploratory manner, the relative importance of several key determinants of Canadians' health.

Some of the results were expected, as it is well known that smoking, having a higher BMI, and having a low level of education are all bad for health. But some of the results were more surprising. Based on the scenarios that have been simulated, it appears that the influence of education (specifically, the difference between having completed post-secondary education and not having completed high school) is as great as that of smoking. Both sets of impacts are considerably larger than that of BMI, notwithstanding the very high level of current public and public policy interest in obesity. Finally, our results indicate that the impacts of scenarios relating nominally to Antonovsky's sense of coherence scale – though possibly due to unmeasured factors for which it is a marker – were larger than any of the others. This result is likely attributable to a combination of the weight of the cognitive dimension in the McMaster

health utilities index scoring function, combined with the empirical prevalence and extent of cognitive decline observed in the NPHS. Our finding in this analysis that cognitive factors have much larger effects on health status than BMI is in contrast to the weighting of these factors in the general public discourse.

6. Acknowledgements

This analysis has benefitted greatly from discussions at the NIDI conference in April 2011, as well as at the IEA in Aug 2011. We are further indebted to Frans Willekens and two anonymous referees for the valuable comments. This research was funded by a grant from the Canadian Foundation for Innovation. We, of course, remain responsible for any errors or infelicities.

References

- Antonovsky, A. (1987). *Unraveling the Mystery of Health*. San Francisco: Jossey-Bass.
- Evans, R., Barer, M., and Marmor, T. (eds.) (1994). *Why Are Some People Healthy, and Others Not? The Determinants of Health of Populations*. New York: Aldine de Gruyter.
- Feeny, D., Furlong, W., Torrance, G., Goldsmith, C., Zhu, Z., DePauw, S., Denton, M., and Boyle, M. (2001). Multi-Attribute and Single Attribute Utility Functions for the Health Utility Index Mark 3 System. *Medical Care* 40(2): 113–128. doi:10.1097/00005650-200202000-00006.
- Hastie, T., Taylor, J., Tibshirani, R., and Walther, G. (2007). Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics* 1(1): 1–29. doi:10.1214/07-EJS004.
- He, Y. and Raghunathan, T.E. (2006). Tukey's gh Distribution for Multiple Imputation. *The American Statistician* 60(3): 125–129. doi:10.1198/000313006X126819.
- Head, J., Klivmaki, M., Seigrist, J., Ferrie, J., Vahtera, J., Shipley, M., and Marmot, M. (2007). Effort-reward imbalance and relational injustice at work predict sickness-absence: the Whitehall II Study. *Journal of Psychosomatic Research* 63(4): 433–440. doi:10.1016/j.jpsychores.2007.06.021.
- Hoaglin, D.C. (1985). Summarizing Shape Numerically: The g-and-h Distributions. In: Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (eds.). *Exploring Data Tables, Trends, and Shapes*. New York: Wiley.
- Krieger, N. (1994). Epidemiology and the Web of Causation: Has Anyone Seen the Spider? *Social Science & Medicine* 39(7): 887–903. doi:10.1016/0277-9536(94)90202-X.
- Lawless, J.F. (1987). Regression Methods for Poisson Process Data. *Journal of the American Statistical Association* 82: 808–815. doi:10.1080/01621459.1987.10478502.
- Manuel, D., Schulz, S.E., and Kopec, J. (2002). Measuring the Health Burden of Chronic Disease and Injury Using Health-Adjusted Life Expectancy and the Health Utilities Index. *Journal of Epidemiology and Community Health* 2002(56): 843–850. doi:10.1136/jech.56.11.843.

- McIntosh, C., Fines, P., Wilkins, R., and Wolfson, M. (2009). Income disparities in health-adjusted life expectancy for Canadian adults, 1991 to 2001. In: *Health Reports, November 2009, Statistics Canada*. <http://www.statcan.gc.ca/pub/82-003-x/2009004/article/11019-eng.pdf>.
- Rowe, G. and Binder, D. (2008). *Can Bootstrap Replicates be used for Cross-validation? 2008 Proceedings of the Survey Research Methods Section*. Paper presented at the American Statistical Association Annual Meeting.
- Sapolsky, R. (2004). *Why Zebras Don't Get Ulcers: A Guide to Stress, Stress-Related Diseases and Coping*. New York: Holt.
- Statistics, Canada (ModGen, no date). <http://www.statcan.gc.ca/microsimulation/modgen/modgen-eng.htm>.
- Statistics, Canada (1998). National Population Health Survey Overview, 1996/7. <http://www.statcan.gc.ca/pub/82-567-x/82-567-x1997001-eng.pdf>.
- Sullivan, D.F. (1971). A Single Index of Morbidity and Mortality. *HSMHA Health Reports* 1971 86(4): 347–355. doi:10.2307/4594169.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58(1): 267-288.
- Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* 16(4): 385-395. doi:10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.
- Whitehead, M., Townsend, P., and Davidson, N. (1982). *Inequalities in Health: The Black Report and the Health Divide*. Penguin Social Sciences <http://www.sochealth.co.uk/Black/black.htm>.
- Will, P., Nobrega, K.M., Berthelot, J.-M., Flanagan, W., Wolfson, M.C., Logan, D.M., and Evans, W.K. (2001). First, Do No Harm – First do no harm: extending the debate on the provision of preventive tamoxifen. *British Journal of Cancer* 85(9): 1280–1288. doi:10.1054/bjoc.2001.2125.
- Wolfson, M. (1994). POHEM – A Framework for Understanding and Modeling the Health of Human Populations. *World Health Statistics Quarterly* 47(3/4).
- Wolfson, M. (1996). Health-Adjusted Life Expectancy. *Health Reports Statistics Canada* 8(1).
- Wolfson, M. and Rowe, G. (2001). On Measuring Health Inequalities. *Bulletin of the World Health Organization* 2001(79): 553–556.

- Wolfson, M., Rowe, G., Gentleman, J., and Tomiak, M. (1993). Career Earnings and Death: A Longitudinal Analysis of Older Canadian Men. *Journal of Gerontology: Social Sciences* 48(4): S167–S179. doi:10.1093/geronj/48.4.S167.
- Yeo, D., Mantel, H., and Liu, T.-P. (1999). Bootstrap Variance Estimation for the National Population Health Survey. In: *American Statistical Association: Proceedings of the Survey Research Methods Section*. 49–57 https://www.amstat.org/sections/srms/Proceedings/papers/1999_136.pdf.