Chaire de recherche Industrielle Alliance
sur les enjeux économiques
des changements démographiques

CEDIA.ca

COMPAS

A HEALTH MICROSIMULATION MODEL FOR QUEBEC

DOCUMENT TECHNIQUE
TECHNICAL DOCUMENT

David Boisclair, Aurélie Côté-Sergent, Jean-Yves Duclos,
Alexandre Lekina, Steeve Marchand and Pierre-Carl Michaud

December 2014

UNIVERSITÉ LAVAL
Faculté des sciences sociales

ESG UQÀM

La Chaire de recherche Industrielle Alliance sur les enjeux économiques des changements démographiques est une chaire bi-institutionnelle qui s'appuie sur un partenariat avec les organisations suivantes :

- Centre interuniversitaire de recherche en analyse des organisations (CIRANO)
- Industrielle Alliance, Assurance et services financiers inc.
- Régie des rentes du Québec

COMPAS: A health microsimulation model for Quebec

David Boisclair*
Aurélie Côté-Sergent†
Jean-Yves Duclos‡
Alexandre Lekina§
Steeve Marchand¶
Pierre-Carl Michaud‖

December 2014

Abstract

In this document we present the design and structure of COMPAS, a dynamic microsimulation model of Quebecers' state of health and health care use. This model has two major dimensions: a dynamic component, which allows for the evolution of simulated individuals' health status throughout their life cycle, and a cross-sectional component, which links use of medical resources with the health status of these simulated individuals. The model uses the 2010 population as the starting point, and several sources of data to model dimensions of state of health and mortality, as well as various transitions between the states. It also lets users analyze just one cohort at a time, or the entirety of the population as it evolves over the course of time with its incoming individuals.

Keywords: microsimulation, health, life expectancy, Québec.
Codes JEL: C63, I14, J11.

*ESG UQAM
†ESG UQAM
‡Université Laval, CIRPÉE, CIRANO
§ESG UQAM
¶Université Laval
‖ESG UQAM, CIRPÉE, CIRANO, RAND. Correspondance: Département des sciences économiques, Université du Québec à Montréal, C.P. 8888, Succ. Centre-ville, Montréal (QC) H3C 3P8, Canada.

Contents

List of figures

List of Tables

Introduction

Context

The share of national income dedicated to health care has not ceased to increase over time. In Canada, this share has increased from 7.1% in 1981 to 11.7% in 2011, and from 8.6% to 12.6% in Quebec (Canadian Institute for Health Information, 2012). The same observation can be made for the past few decades — but with a greater effect — with regard to the share of health spending in the budget of the state. This is particularly true in Canada and in Quebec, where three-quarters of health expenditures are public. In Quebec, the share of government revenues dedicated to health has risen from nearly 32% in 2000 to more than 45% in 2014.

In most industrialized countries, the increase in spending has been accompanied by an increase in life expectancy. Thus, a Canadian born in 1960 had a life expectancy at birth of 71 years, compared to 75.3 years if born in 1980 and 80.8 years if born in 2008 (World Bank, 2014). A major share of the gains have been achieved for those over the age of 50 years: among men, between 1991 and 2008, life expectancy at 50 years rose from 27.6 to 31 years, and women also saw appreciable gains, from 32.8 to 34.7 years (Bohnert, 2013). In Quebec, the situation is similar: life expectancy at birth went from 78 years in 1997 to more than 81 years in 2012, while life expectancy at 65 years rose by about 2.5 years over the same period (Régie des rentes du Québec, 2013; Statistics Canada, 2014a,b).

Some important trends are underway. To start with, long-term demographic trends such as an aging population and longer life expectancy are exerting pressure on public finances and the health system as a whole. Also, medical technological progress and the capacity to provide home care to elderly persons are key factors in the future evolution of resources devoted to health.

Moreover, we observe a progression of chronic illness associated with obesity in Quebec, a progression which has increased pressures on the health system and the budget of the state. The prevalence of hypertension thus rose by 5.1% annually between 2000 and 2007(Institut national de santé publique du Québec, 2011). Among the population aged 65 years and over, the prevalence of heart disease increased at an annual rate of 11.0% between 1994 and 2003 (Institut national de santé publique du Québec, 2006). It is imperative to be able to adequately model and analyse these trends to obtain reliable projections of the future situation.

Certain studies, such as Clavet et al. (2013), use the microsimulation model developed by the SIMUL team (see below) and project that health spending will amount to at least 56.5% of public revenues in 2030 and could even exceed 68% according to a plausible scenario. Using a methodology which makes use of aggregate data, Godbout et al. (2014) obtain results of the same magnitude (58.6%), as have many previous similar works going back to the early 2000s. However, evaluation of policies which aim to attenuate these pressures while continuing to improve the population's health requires models capable of accounting for the aforementioned factors in a detailed manner.

Long-term microsimulation

Since the observed trends are multidimensional, modeling at the aggregate, population level does not permit to treat these questions while accounting simultaneously for a large number of factors.

Among other aspects, it cannot account for increasing heterogeneity in life trajectories of elderly persons. This element was brought up during the work of a panel aiming to bring together researchers and organizations endeavouring to study the future evolution of health expenditures in the United States (National Research Council, 2012). The *Future Elderly Model* (FEM), developed by a team of researchers at the RAND Corporation, is one of the only long-term models in the United States and in the world to incorporate complex interactions between the diverse trends noted above. Initially built using an official database on 100 000 Medicare beneficiaries, thus Americans aged 65 years and over, the model now covers the entire population aged 50 years and over of the United States and seven European countries. It allows users to project health status, use of medical resources, as well as labour income, labour supply and individual retirement decisions; the impact of future changes in health status, longevity and medical technologies can also be forecast. The model has been used in vriouas studies analyzing different questions (recently, see for example Goldman et al., 2013; Michaud et al., 2012).

This type of model always begins with an initial assumption that there is no future change in government policy.[1] These models are nevertheless very useful, both to consider medium- and long-term scenarios, and to evaluate the potential benefits of policies which aim to improve the population's health using " alternative " scenarios constructed by users. These benefits go well beyond reducing health expenditures since they also cover the economic benefits flowing from improved health and from the effects on revenues and expenditures of different public programs. It is in this perspective that we have gone about developing a Quebec health microsimulation model.

Microsimulation linked to aging and health in Canada

A certain number of microsimulation models have been developed and applied to aging or health in Canada over the years. Some have been general in nature, such as Statistics Canada's Lifepaths, while others were always oriented towards health, such as POHEM, also developed by Statistics Canada. To our knowledge, however, none of the existing models has been developed specifically for Quebec, and none of them model simultaneous transitions between different states of health defined by the presence of various illnesses and risk factors. These existing models are briefly described below, based on the inventory performed by Décarie et al. (2012) as well as the summary included in Clavet et al. (2012).

The oldest microsimulation model found in Canada is DYNACAN, set up in the 1970s in what is today Employment and Social Development Canada. This dynamic model, based on the 1971 Census and on the American CORSIM model, is oriented towards public pension programs (Old Age Security, Canada Pension Plan) and on the effect of policies relating to these. DYNACAN was abandoned at the end of the 2000s, but still served as partial inspiration for the socioeconomic model developed by the SIMUL team (see below), also based on a representative sample of households.

The model currently used by the Canadian government — and used by a number of researchers — is the Lifepaths model, developed by Statistics Canada over the past 25 years. As opposed to others

[1]The idea underlying the base (or " reference") scenario is not that policies will actually not change, but rather that the long-term projections should be carried out on the basis of what is known of government policy at the time of performing the simulations.

mentioned here, this model is based on the concept of a dominant person in a household, and simulates this person's situation as a whole. The model covers dimensions such as retirement and institutionalization, but has very little on health status and nothing on health care use (only one measure based on self-reported health status is included, in addition to one measure of disability). A large number of articles and analyses have been produced using Lifepaths, but this last has basically no provincial specifications (Clavet et al., 2012).

To our knowledge, the only true health microsimulation model existing in Canada is named the POpulation HEalth Model (POHEM). Since the 1990s, Statistics Canada has built many modules of this model in order to model different illnesses and risk factors. To date, however, the illnesses and risk factors are modelled separately and only permit an analysis " by illness". POHEM uses many sources of data, including several major surveys, the census, illness and hospitalization records, etc.

Combining POHEM with a variety of data types and sources is also the spirit in which our team has sought to develop a more general dynamic microsimulation model linked to health for Quebec. The model aims to respond to several concerns with respect to existing models, and aims, in some ways, to further improve the excellent analysis and research infrastructure developed with these models.

COMPAS

Our team has already developed, in collaboration with the Ministry of Finance of Quebec, a microsimulation model which can be used to analyze the impacts of demographic changes and economic policies on public finances and Quebecers' standard of living (Clavet et al., 2012). However, this model does not consider health care, which accounts for nearly 50% of the Quebec budget.

We have therefore been developing a dynamic health microsimulation model using existing sources of administrative and survey data. Since the SIMUL team has already developed infrastructure for demographic and economic microsimulation, we have created a complementary model linked to health that will extend the scope of application of SIMUL. This model is named COMPAS. COMPAS was created in 2013 and its development is pursued on a continuous basis.

The health model has two important components: a) a dynamic component which allows us to simulate individuals progressing through their life cycle, and thus also to incorporate the evolution of their health status, captured by variouse indicators such as the presence of illnesses, the presence of activity limitations, etc.; and b) a cross-sectional component aiming to quantify the cost of resource use associated with the health status of persons alive in a given year. This second component will benefit from the rich infrastructure of existing data, including data from the RAMQ as well as MED-ECHO.

Each year, the living population is characterized by a set of demographic, economic and health variables. The health status variables make it possible to attribute use of health care and related expenses in a given year. This population then passes through a transition phase where health status may change. For example, an individual aged 80 years with hypertension would have a set of probabilities to suffer a cardiovascular incident. These probabilities depend on the current health status. During this transition, each individual faces a risk of dying or of developing some incapacity(ies) which, for instance, might lead to their withdrawal from the labour market or entry into a long-term care facility.

It is this wealth in states and transitions which makes it possible to perform long-term forecasts using microdata and to analyze different scenarios that can affect a fraction or the entirety of the population. For example, what would happen if we could delay entry into a long-term care facility? Or if we could reduce mortality risk for individuals having suffered a stroke? Or reduce the prevalence of diabetes? Since the health model will be linked to the socioeconomic model already developed by SIMUL, it will be possible to analyze the impact of these scenarios on government income and other expenditures.

Already, COMPAS produces baseline results which are presented elsewhere (Boisclair et al., 2014). For example, COMPAS predicts that the prevalence of hypertension in 2050 will be in the 21% range for individuals aged 70-74 years. The model also captures aging of the Quebec population. Between 2010 and 2050, the share of elderly individuals over 75 years of age (in the population aged 30 years and over) will increase from 10% to just under 30%. We also note an increase in life expectancy between 2010 and 2050. Individuals aged 30 years in 2050 will live an average of 3.3 years longer than individuals aged 30 years in 2010. As for 60-year old individuals, the difference will be 2.8 years.

Objectives and progress

The research objectives associated with the development of the model are as follows:

1. Develop a longitudinal database making it possible to estimate transitions between states of health as well as mortality at the population level. This database is built using the Quebec sample from the National Population Health Survey (NPHS).
2. Estimate dynamic relationships between states of health using the database built in point 1 and a suitable econometric modelling. These relations are multidimensional and enable us to preserve the heterogeneity of health trajectories in the population.
3. Develop a database of costs and use of resources according to the state of health identified in point 1. This database should incorporate remuneration for health care services provided, spending on medications and budgetary amounts allocated for each diagnosis related group (DRG) for hospital stays. Il should enable us to create a measure of total annual expenditures for each individual in the target population.
4. Estimate relationships between costs and use of health care using the dataset created in point 3 together with suitable econometric models.
5. Construct the simulation model based on the demographic and economic statistical infrastructure of the SIMUL model and integrating the relationships estimated in points 2 and 4.
6. Calibrate and validate the simulation model on a forward-looking basis, based on projections of the ISQ and RRQ for mortality and those of the ministries for forecasted health spending in Quebec.
7. Evaluate the sensitivity of future forecasts to economic and social policies, trends in health, improvements in longevity, and economic conditions.

At the time of drafting this report, we have completed objectives 1 and 2 as well as parts of objectives 3 and 4 related to health care use. Also, we have finished building the initial model (preliminary achievement of objective 5), and we can present preliminary results. Moreover, we have completed most of the work relating to objective 6. For objective 3, the part involving administrative data requires a lot of work to obtain access to the data. To date, we have access to the data of the RAMQ on consultations with general practitioners and medical specialists; we must file a request with the RAMQ in order to also have access to data on pharmaceutical expenditures. This work began in the autumn of 2014. For short-term hospitalizations, we have access to MED-ECHO data, which will enable us to estimate the cost of hospitalizations. There is still work to be done on the cost of institutionalization and that of home services; the search for databases to support this work began in the summer of 2014.

1 Model structure

1.1 Dynamics of COMPAS

The microsimulation model includes several components. The structure of the model is almost entirely based on the population dynamics. The model tracks agents, characterized by socioeconomic and health attributes, which transition from one period to another. In each period, new agents enter the simulation at the default starting age, which we set at 30 years, or at any age if they immigrate.[2] Their life in the model ends at their time of death or when they reach the maximum age permitted in the model, which we have set at 110 years. In each simulation cycle [3], we can observe not only their health status, but also their use of health care. A group of agents born in the same year — or the same two-year period when the simulation cycles are two years —is denoted as a cohort. The model enables us to perform simulations for just one cohort or for a population, comprised of many cohorts. A population is comprised of the number of economic agents who are alive at the end of a given year. In each simulation cycle, the population moves forward in time. To start with, it loses some agents due to mortality. It also gains some because there is renewal, i.e., a new cohort of economic agents over the age of 30 years (or $30-31$ years when using 2-year simulation cycles) enters the model. Finally, it gains and loses members due to migration. Each simulated agent has a given demographic weight. Thus, the statistics of interest can be calculated at the population level.

The model begins in 2010, which is the most recent year for which reliable data is available for many of the sources of information that we use. For that year, we create an initial population using an existing survey. This initialization phase gives a statistically representative sample of Quebec in 2010. Then begins the process of demographic evolution of the population up to the target year: 2050. Depending on the desired output variables, the model can keep estimating its variables up to the year 2130 if any remaining agents are still alive. This could be useful, for example, to analyze the evolution of life expectancy of each cohort entering the model. In effect, cohorts entering the model between 2010 and 2050 will see a major proportion of agents who remain living in 2050 (the oldest will only be 70 years old). Figure 1.1 summarizes how the model works in population mode and the dynamics of COMPAS.

[2] The age of entry is set at 30 years in order to avoid having to model individual education trajectories and choices, and because illnesses and incapacities considered in the model are uncommon before this age.

[3] In COMPAS, a simulation cycle corresponds with a period of two years. This choice is not restrictive and can be changed according to the objectives of the analysis.

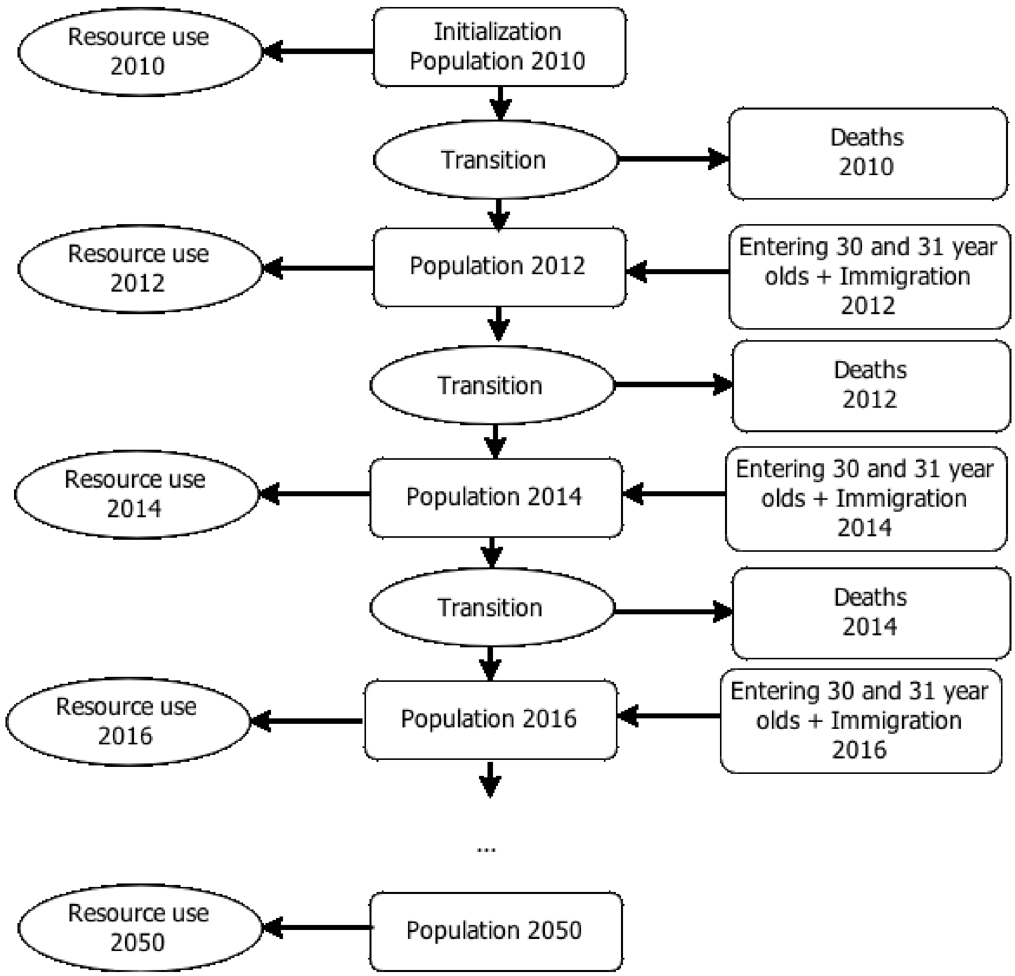

Figure 1.1: Dynamics of the COMPAS model with 2-year simulation cycles

In the case of the cohort mode, the renewal component (**newcohort.f95**; see below) is called up just once, while the immigration ((**immigration.f95**) and initialization (**initialize.f95**) components are inactive. This mode only generates a single cohort of individuals aged 30 years, or $30-31$ years with the 2-year simulation cycle (see chapter 3).

Since the simulations include a stochastic element, the model enables the user to make several replications in order to ensure that the result obtained is not purely a matter of chance. By default, 100 replications are used in our work.

1.2 Structure of the FORTRAN code

This section documents the structure of the model and provides a guide for an advanced user. Therefore, it is not needed to understand the subsequent chapters, and it can be omitted without loss for the reader who does not wish to be informed of the model's programming characteristics.

COMPAS is mainly built using FORTRAN, a simple language with great flexibility and adaptability for different modeling needs. The design of the model is modular, which enables us to modify the parameters and components separately, and without necessarily — or immediately — affecting the entire model. This modular structure of the FORTRAN code is reproduced in figure 1.2.

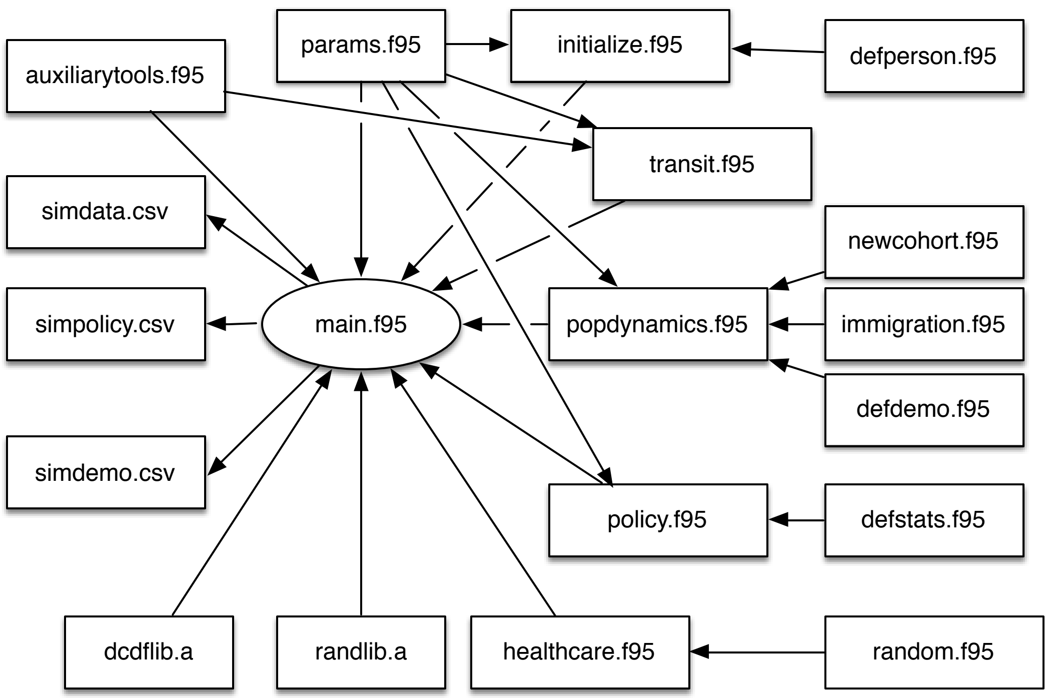

Figure 1.2: Structure of the FORTRAN code

The main program of the model is thus **main.f95**. It calls up a set of FORTRAN modules corresponding to the different components of the model as presented in figure 1.1, above. The following subsections present the different modules.

1.2.1 Parameters

The *params* module contains the different parameters of the model. These are used in nearly all modules. The start year (*startyear*) and end year (*stopyear*) of the simulation, as well as the year when the last cohort enters (*stopyear_ enter*), are parameters which can be changed by the user. As mentioned, the default values of these parameters are respectively 2010, 2130 and 2050.

The *nyears* parameter does not correspond with the number of years, but rather with the number of cycles in the simulation. This can vary for different reasons, such as the end year of the simulation: if *stopyear_ enter* prevails, then the simulation ends in the last year that a new cohort enters into the model (2050 by default); if textitstopyear prevails, then the simulation continues until all the economic agents are deceased or up to *stopyear* (year 2130 by default). Importantly, *nyears* also depends on the number of years between each demographic transition (*gapyears*; see below for the transitions), set to 2 by default. That is to say that, for the period from 2010 to 2050, by default the model would pass through 20 cycles of the simulation; here, we thus have $nyears = (stopyear - startyear)/gapyears$.

The parameters *startage* and *stopage* determine the minimum and maximum age allowed in the simulation (by default 30 and 110 years), while *nages* corresponds with the number of different ages or to the difference between *stopage* and *startage*.

The initial sample size (*nstartpop*) and that of new cohorts (*nstartcohort*) are then determined. This is not linked to initial population size. Since each agent will have a population weight according to his/her characteristics, this weight will be adjusted as a function of initial sample size. The larger the values of *nstartpop* and *nstartcohort*, the more precise and stable the simulation results are. The simulation will, however, take longer.

The next four parameters determine choices in the model. The *cohort* parameter allows to set the simulation mode: cohort mode or population mode. One can also choose to save simulated individual data (*savedata*), the ''dashboard" (*savepolicy*) and the demographic data (*demo*).

One can also choose to perform many replications of the simulations (*doreps*). The *nreps* parameter indicates the number of replications.

The last three parameters deal with demographic assumptions. The *migration* parameter determines whether there is migration while *immigrate* determines the net immigration rate as a share of the population. Finally, the *qxgain* parameter determines the annual improvement in mortality rates, in %.

1.2.2 Agents

The microsimulation model works from a population of agents. These agents are defined in the file **defperson.f95**. The " person" type is a structure (in C++) which defines the attributes of a person; this stores all the characteristics of the person. For example, one can declare a population of agents comprised of n persons by creating " type (person) pop(n)". Then, one can obtain the age of person i using the " pop(i)%age" command. The combination of *id* and *byear* is a unique indicator for each person. The " wgt" attribute is the population weight of the agent.

1.2.3 Initialization

The *initialize* module is that which creates the population in the start year of the model using CCHS 2010. The 3 chapter explains in more detail the manner in which the initialisation is done. The model thereby draws from this survey a sample with replacement of size *nstartpop*. It then calibrates the age distribution drawn from the survey on the one produced by the ISQ and adjusts the weight of each agent accordingly.

The program that initialises the population is found in **src/initialize.f95**.

1.2.4 Transitions

The *transit* module is called by **main.f95** to give the starting population for year $t+1$ using the population in year t, before the demographic transitions take place (i.e., before those having died are removed from the model and new entrants and immigrants are added). This module contains the definitions of the transition probabilities for each attribute of the individual. It is in this module that we update the agent's characteristics. If the agent presents the value alive=.false., he/she is then eliminated in the module *popdynamics*. Chapter 4 presents the different modules used in order to calculate the transition probabilities.

The program of the module is given in **src/transit.f95**.

1.2.5 Population dynamics

The *popdynamics* module is called by **main.f95** to perform the demographic transition and to calculate the demographic statistics. In this module, those who have died are first eliminated. Then, the *newcohort* module is used to generate a new cohort. Finally, the *immigration* module is called to adjust the weights of immigrants such that they will account for net immigration to Quebec. The demographic statistics are calculated *before* each demographic transition; these statistics are defined in the **defdemo.f95** module.

The program which accomplishes the demographic transition is found in **src/popdynamics.f95**.

1.2.6 New cohorts

The *newcohorts* module generates the new cohorts in two steps. First, it generates a sample of size *nnewcohort* using the joint distribution of initial conditions in the sample of individuals 30 and 31 years of age in the CCHS. Then, the weights are adjusted in order to account for the size of entering cohorts as projected by the ISQ in its demographic forecasts. This module is called by *popdynamics*.

The programme which generates the new cohorts is found in **src/newcohorts.f95**.

1.2.7 Immigration

The *immigration* module adjusts the weights of the current population to account for the entry of new immigrants. Thus, no new observations are added to the population to account for that entry; rather, new immigrants are accounted for by reweighting the existing observations. The *immigration* module is called by *popdynamics*.

1.2.8 Dashboard

The dashboard is controlled by the *policy* module. This module uses a population of agents as input. It calculates different statistics about their health and their health care use.

The program of the module is given in **src/policy.f95**.

The program that specifies which statistics are available is found in **src/defstats.f95**.

1.2.9 Execution of the program

The FORTRAN program uses two support libraries, **dcdflib.a** and **ranlib.a**, which combine probability functions and random draws. These are found in the **libs/** directory and **create.sh** bash files can create these libraries for the desired platform (PC, Mac, Linux 32 and 64 bit).

In Linux and Mac, the file which creates the executable of the model is found in **build/compile_linux.sh**. It contains a collection of bash commands which will compile the FORTRAN file codes. The FORTRAN compiler used is gfortran. If the program executes correctly, all that remains is to run the model with the ./runmodel command.

The model creates three output files: **simdata.csv**, **simpolicy.csv** and **simdemo.csv**. These files can be read using a software such as Stata to produce figures, etc.

1.3 Organization of files and *subversion*

The " santé " directory is in *version control*, which means that the versions of the model are saved in an Assembla directory (http://www.assembla.com). This allows us to save old versions without having to save copies of the model. An introduction to *subversion* is available from http://dev.nozav.org/intro_svn.html.

2 Data

2.1 Overview

This chapter presents the preparation of the data used to build the microsimulation model. We use various databases to construct the model. By far the most important survey is the National Population Health Survey (NPHS), which is the only longitudinal survey on health in Canada. We use it to build the transitions module and the health care use module. The 2010 Canadian Community Health Survey (CCHS) is used to construct the initial population of the model. Is it representative of the Canadian population in 2010. Finally, we use the data of the Quebec Statistical Institute (ISQ) on population by age to replicate the age distribution of the model's initial population. We give more details on complementary sources of data in the subsequent chapters.

2.2 NPHS

The NPHS is a longitudinal survey carried out on a sample of the Canadian population. The questionnaire was administered on the same sample of respondents every two years beginning in 1994-1995. There are thus nine cycles in this survey, the last of which being 2010-2011. While there are three components of the NPHS (the Households component, the Health Institutions component, and the North component), only the Households component is used here. The North component has not been developed by the NPHS since 2000-2001, while the Health Institutions component had to be ended in 2002-2003 because too high a share of the sample population had deceased (Réseau de recherche en santé et en sécurité du travail du Québec, 2014).

A particularity with the Households component is that all respondents interviewed necessarily live in a private household in the first cycle of the survey. The survey thus excludes residents in Indian reserves, on Crown land and in remote regions of Quebec and Ontario, as well as full-time members of the Canadian Forces and persons living in institutions (Statistics Canada, 2012c). It should be noted that if a respondent from the Households component transitions to an institution between two cycles of the survey, they are still followed in the Households component.

In 1994, there were 17,276 respondents in Canada, including 2,740 aged 65 years or over. In comparison, there were 3,000 respondents in Quebec, including 369 aged 65 years or over. Table 2.1 presents the number of observations in the NPHS by age group for Quebec and the rest of Canada. It can be concluded that the number of observations in Quebec is small, especially over 65 years. Given that certain health problems are relatively rare, it seems prudent to use the entire Canadian sample and to tolerate deviations for Quebec as needed.

	Quebec	Rest of Canada
20 to 24 years	172	1047
25 to 29 years	254	1068
30 to 34 years	290	1361
35 to 39 years	286	1264
40 to 44 years	264	1049
45 to 49 years	218	1012
50 to 54 years	189	784
55 to 59 years	151	731
60 to 64 years	148	670
65 to 69 years	138	685
70 to 74 years	117	670
75 to 79 years	68	492
80 to 84 years	25	356
85 to 89 years	22	136

Table 2.1: Number of observations for Quebec and all of Canada in the NPHS (1994-2011)

2.2.1 Weighting

In order to proceed with estimations using NPHS data, one must use sampling weights. This allows the selected sample to be representative of the entire Canadian population. Each person in the sample represents a certain number of individuals in the population. An initial weight is calculated and then adjusted to account for certain survey particularities. This first weight is the inverse of the selection probability. Thus, the greater the chance of an individual to be selected into the sample, the lower their weight. This weight is calculated in the first cycle of the survey (1994-1995) and does not vary in time. Since there is no resampling in the NPHS, this weight enables us to represent the population of interest as it was in 1994, when initial observations were present (Statistics Canada, 2012c).

Statistics Canada (2012c) then computes other weights which allow a subset of the sample to be representative of the Canadian population. The second weight only applies to individuals having responded to the nine cycles of the survey (or who deceased or were in an institution in some cycles). Indeed, some individuals did not respond to all survey cycles.

A longitudinal survey on health requires good follow-up of respondents who could not be interviewed. In particular, it is important to be able to clearly identify those respondents who are deceased. It is also important to verify the frequency at which people leave NPHS and see if their characteristics are different from those who remain. We therefore deal with this last point before discussing mortality.

2.2.2 Attrition and selection

Table 2.2 shows cumulative response rates, attrition rates (the share of those having responded in the preceding cycle who did not respond in the current cycle) and the share of those entering (respondents who left between two cycles but who subsequently returned). For those who entered, the software used by Statistics Canada to record responses ensures that there are no inconsistencies in the responses of individuals between the cycle preceding their absence from the survey and the cycle following it (Statistics Canada, 2012c).

	Response rate	Attrition rate	Share of entrants
1994-1995	100.00%	.	.
1996-1997	92.86%	7.14%	0.00%
1998-1999	87.55%	6.75%	1.44%
2000-2001	83.61%	7.24%	3.29%
2002-2003	79.62%	7.57%	3.59%
2004-2005	74.66%	8.53%	3.57%
2006-2007	73.13%	6.50%	4.97%
2008-2009	65.62%	10.05%	2.54%
2010-2011	64.79%	6.50%	5.66%

Table 2.2: Response and attrition rates by cycle in the NPHS

Attrition rates are generally low and are comparable with those observed in other frequently used longitudinal surveys, such as the Health and Retirement Study (HRS) and the Panel Study of Income Dynamics (PSID) in the U.S. The attrition rate in the HRS was 35.7% between 1992 and 2002 (Kapteyn et al., 2006). In 2002, 64.3% of the initial sample had responded to all cycles. As for the PSID, observed attrition was between 2.5 and 3% per year between 1968 and 1989 (Fitzgerald et al., 1998). In comparison, between 1994 and 2004, attrition in the NPHS is only 22.4%. Finally, in the NPHS, a large share of respondents who leave the panel eventually return.

In table 2.3, we analyze whether the health status of individuals remaining in the panel continuously from 1994-1995 to 2010-2011 is significantly different from non-deceased individuals who did not respond to one or more cycles. The table presents Student's t-statistics for difference in averages calculated relative to 1994 data. For example, a value of more than 1.96 or less than -1.96 for a variable means that we reject at the 5% confidence level the null hypothesis that the average of this variable for individuals having responded throughout the entire surveying period is identical to that of individuals who will eventually stop responding. Otherwise stated, a value of more than 1.96 or less than -1.96 indicates a difference between the two groups, at a 5% threshold. Significant differences could cause problems, because in that case non-responses would reduce the sample's representativeness.

The tests prove significant at the 5% threshold for some variables. For some age groups, the presence of diabetes (65 years and over) and heart diseases (40 to 64 years) is higher in 1994 for those who will later stop responding to the survey. This is also true for more elderly smokers, for former smokers under the age of 65, and for individuals suffering from several incapacities.

	under 40 years	40 to 64 years	65 years and over
Diabetes	-1.0222	-0.3774	-2.1553
Hypertension	-0.2972	-1.9783	-0.7707
Cancer	0.8821	-1.5148	-1.5080
Heart diseases	0.3777	-2.2666	-1.6501
Stroke	1.4179	-0.3828	-0.3224
Lung diseases	0.4293	0.0259	-0.9210
Dementias	.	0.3665	-0.4008
Current smokers	-1.0222	-0.3774	-2.1553
Former smokers	-5.2900	-4.1568	-1.5910
No incapacity	0.5909	-1.0680	4.9315
1 ADL or cognitive impairment	0.8370	0.9817	-2.0228
2+ ADL	-2.5082	0.4724	-4.4845

Table 2.3: Statistical selection tests of, NPHS (1994-2011): Difference in the shares between individuals responding to all cycles and those for whom there are missing cycles. T-statistics are presented. A negative value indicates a higher prevalence among individuals who eventually leave.

2.2.3 Mortality

The NPHS offers two ways to detect mortality. The first is for an interviewer to obtain this information when they attempt to contact the respondent. The second is by validation with the national registry of deaths. For the validation to be possible, the respondent had to consent to a matching with administrative data. About 90% of respondents consented to this matching.

Table 2.4 presents the annual mortality rates of the NPHS by cycle for the population aged 30 years and over. In earlier cycles, the rates increase over time. A reason leading to the rates initially being low is that the initial population of the NPHS is only comprised of individuals who are not in an institution. Mortality rates then decline in later survey cycles, notably due to problems with validation with the national registry of deaths.

Figure 2.1 presents the mortality rates by age from the NPHS from 1994 to 2006, as well as the 95% confidence interval around these rates. We compare these rates with those from the periodic mortality rates from Human Mortality Database for a similar period, i.e. from 1995 to 2004. All Canadian data comes from Statistics Canada. The correspondence is good up to about 75 years of age. However, there is divergence after 75 years, that is from the moment where the number of individuals in institutions increases. Since individuals responding to the NPHS are not in an institution in the first survey cycle,

	Mortality rates
1994-1995	0.98%
1996-1997	1.13%
1998-1999	1.13%
2000-2001	1.15%
2002-2003	1.28%
2004-2005	1.25%
2006-2007	1.05%
2008-2009	0.85%
Total	1.10%

Table 2.4: Annual mortality rate by NPHS cycle, populated aged 30 and over

they are probably in better health than the Canadian population as a whole. This could explain why the mortality rate in the NPHS is lower than that calculated using the periodic mortality tables beginning at 75 years of age. We therefore recommend eventual calibration of the NPHS using the survival tables.

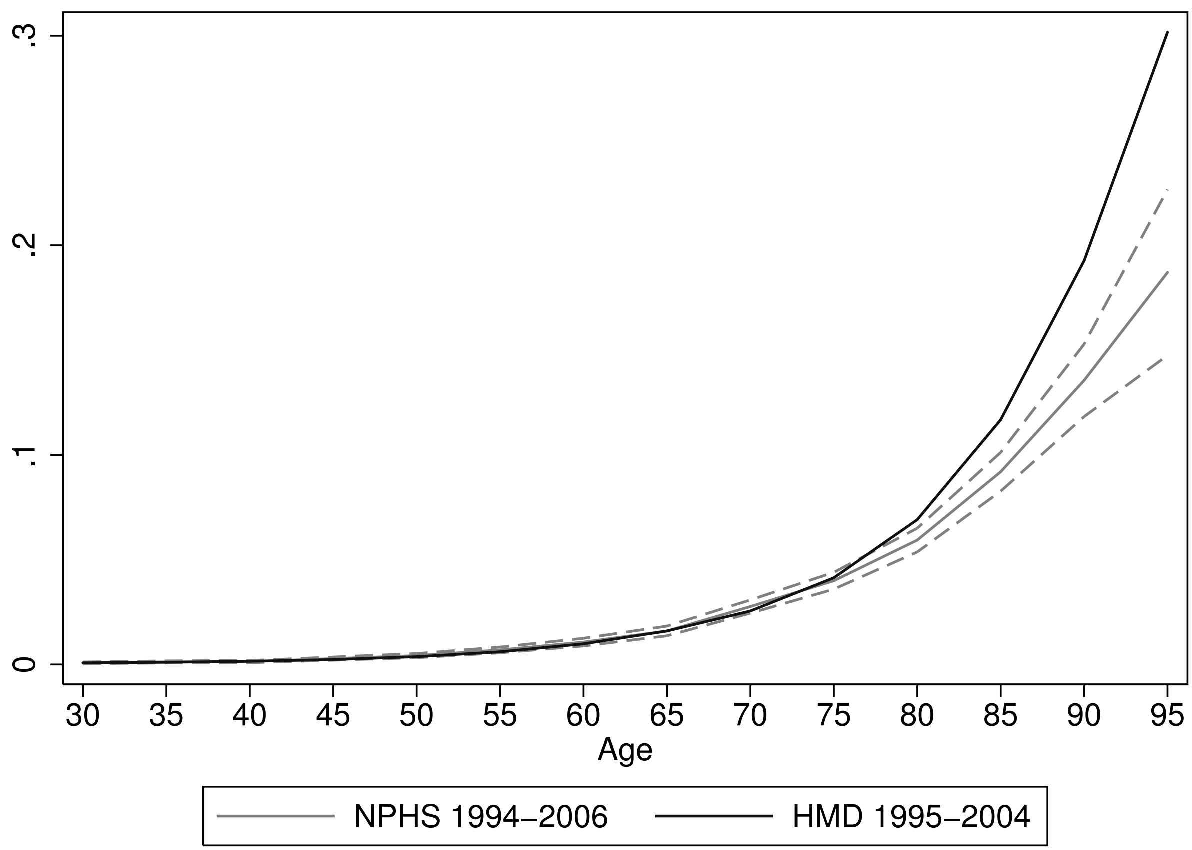

Figure 2.1: Comparison of mortality rates: NPHS vs. Human Mortality Database (HMD). 95% confidence interval is dotted line.

2.3 Constructing the variables

The NPHS questionnaire is rich and the questions are for the most part identical from one cycle to the next. We divide this section into three. First, we present the construction of variables relating to self-reported health problems. These variables deal with measures of incidence (having started to suffer from an illness between two points in time) and prevalence (suffering from an illness at a given moment). Then, we describe the measures of disability. We also present the measures of risk factors. Finally, we describe the health care use variables.

The dimensions of health are presented in table 2.5. A person's health status includes many dimensions in the model.

Diabetes
Hypertension
Cancer
Stroke
Heart diseases
Lung diseases
Alzheimer's and other dementias
Obesity
Tobacco
Number of incapacities
Institutionalization

Table 2.5: Dimensions of state of health accounted for in COMPAS

In terms of health care use variables, for now we use the dimensions shown in table 2.6. In future versions of the model, we will refine these variables and consider other dimensions such as costs, etc.

Number of visits to a generalist
Number of visits to a specialist
Number of nights of short-term hospitalization
Use of medicines (yes / no)
Use of home care services (yes / no)

Table 2.6: Dimensions of health care use accounted for in COMPAS

We use the weights to produce the following statistics. In this version, we have not stratified by province or sex. However, in order to verify the differences which may exist between Quebec and Canada, we present for all results Student's comparison tests, which indicate whether the proportion for the rest of Canada is significantly different from that for Quebec.

2.3.1 Self-reported health conditions

The model includes seven self-reported health measures: presence of diabetes, hypertension, cancer, heart diseases, stroke, lung diseases and presence of dementias (Alzheimer's and others). These illnesses are the most common in the population. The questions asked in the NPHS take the following form: " *We are interested in health problems diagnosed by a health professional. Do you suffer from diabetes?,*

... *hypertension?, etc."* (Statistics Canada, 2012d). Once aggregated, the answers of all respondents to these questions lead directly to measures of prevalence. To develop the measures of incidence, we use positive responses in the current cycle in cases where the response was negative in the preceding cycle.

Table 2.7 gives the prevalence rates by 5-year age group for each illness. Unsurprisingly, we find that prevalence increases with age, except for certain illnesses (diabetes, cancer) for which prevalence decreases at more advanced ages. The most common illness among those aged 40-44 years is hypertension, with a prevalence rate of 6.8%. Among those aged 65-69 years, the most common illness is also hypertension, at 39.7%.

We observe differences between Quebec and Canada in terms of the prevalence of illnesses. As indicated by the t statistics presented in the second part of table 2.7, among individuals under 40 years old, the prevalence of diabetes, cancers and dementias is higher in Quebec than in the rest of Canada, while the opposite is true for hypertension.[4] From 65 years on, the prevalence of diabetes is always higher in Quebec than in the rest of Canada. However, a smaller share of Quebecers suffer from cancer, heart diseases or strokes than Canadians who live outside of Quebec.

Age	Diabetes	Hypertension	Cancer	Heart diseases	Stroke	Lung diseases	Dementias
20 to 24 years	0.59%	0.94%	0.50%	1.28%	0.21%	5.10%	0.02%
25 to 29 years	0.86%	2.30%	0.65%	1.14%	0.24%	4.84%	0.17%
30 to 34 years	0.93%	3.37%	0.85%	1.32%	0.24%	4.79%	0.26%
35 to 39 years	1.23%	5.11%	1.03%	1.93%	0.26%	4.99%	0.25%
40 to 44 years	2.18%	6.78%	1.54%	2.59%	0.35%	5.21%	0.20%
45 to 49 years	3.23%	10.22%	1.98%	3.53%	0.60%	5.12%	0.24%
50 to 54 years	4.61%	16.91%	3.03%	5.18%	0.93%	5.33%	0.26%
55 to 59 years	6.97%	25.33%	4.24%	7.50%	1.51%	6.09%	0.33%
60 to 64 years	9.40%	34.01%	6.10%	11.65%	2.65%	7.56%	0.72%
65 to 69 years	11.77%	39.68%	7.46%	16.37%	3.36%	8.36%	1.07%
70 to 74 years	14.48%	44.95%	8.71%	21.08%	5.59%	10.73%	1.84%
75 to 79 years	16.45%	49.15%	9.80%	26.28%	6.93%	11.98%	3.70%
80 to 84 years	17.63%	52.55%	11.05%	32.66%	9.25%	13.72%	6.01%
85 to 89 years	16.27%	53.99%	15.23%	35.96%	11.92%	14.35%	11.28%
90 to 94 years	15.06%	52.15%	14.22%	37.76%	13.33%	13.73%	14.68%
95 to 99 years	14.79%	55.45%	13.18%	37.13%	20.47%	20.20%	18.81%
Total	5.17%	17.12%	3.41%	7.39%	1.75%	6.41%	0.90%
t-test of difference between rates in Quebec and Canada							
under 40 years	-4.3785	4.758	-4.2359	-1.8561	0.7613	1.3227	-3.4427
40 to 64 years	2.4845	4.591	2.6713	0.347	-3.3479	1.35	-2.9588
65 years and over	-3.1319	-0.1904	7.401	2.5295	3.997	-4.6766	1.2087

Table 2.7: Self-reported prevalence of illnesses in NPHS (1994-2011)

Interpreting the prevalence by age is difficult for two important reasons: a) cohort effects (intergenerational differences for a given age) in the incidence of illnesses; and b) higher mortality among those

[4] In all the tables in this chapter, a negative t statistic indicates a *higher* proportion in Quebec than in Canada, while a positive t statistic indicates a *lower* proportion in Quebec. The significance of these differences depends on the size of the statistic; e.g., a t-statistic of greater magnitude than an absolute value of 1.96 indicates a significant difference at the 5% level. The direction of the difference depends on the sign of the statistic, as explained above.

with an illness. Thus, we could very well observe that incidence increases with age while prevalence decreases.

In table 2.8, we present the incidence rates by age group and for each illness. Among those aged 40-44 years, incidence is highest for hypertension, with 1.6%. Indeed, hypertension exhibits the highest incidence up until80 years of age. We also note that incidence generally increases with age, although it declines for some illnesses at advanced ages. A number of reasons could explain this, including selection and the biological processes leading to these illnesses.

Age	Diabetes	Hypertension	Cancer	Heart diseases	Stroke	Lung diseases	Dementias
20 to 24 years	0.15%	0.45%	0.11%	0.13%	.	0.63%	0.06%
25 to 29 years	0.11%	0.71%	0.15%	0.15%	0.03%	0.52%	0.05%
30 to 34 years	0.17%	1.00%	0.09%	0.41%	0.03%	0.63%	0.03%
35 to 39 years	0.33%	1.19%	0.27%	0.43%	0.04%	0.59%	0.02%
40 to 44 years	0.67%	1.61%	0.37%	0.54%	0.13%	0.52%	0.07%
45 to 49 years	0.69%	2.80%	0.44%	0.99%	0.18%	0.54%	0.03%
50 to 54 years	1.12%	4.24%	0.84%	1.18%	0.15%	0.66%	0.03%
55 to 59 years	1.55%	5.68%	1.09%	1.74%	0.42%	1.07%	0.18%
60 to 64 years	1.96%	5.52%	1.21%	2.45%	0.64%	0.96%	0.22%
65 to 69 years	1.68%	6.44%	1.49%	2.76%	0.71%	1.37%	0.38%
70 to 74 years	2.00%	5.46%	2.36%	4.51%	1.22%	1.43%	1.10%
75 to 79 years	2.55%	5.74%	1.83%	4.24%	1.78%	1.75%	1.69%
80 to 84 years	1.24%	4.68%	2.74%	4.99%	2.22%	1.83%	3.80%
85 to 89 years	2.27%	4.31%	1.67%	5.18%	3.00%	2.13%	6.08%
90 to 94 years	1.10%	4.53%	1.94%	6.70%	4.99%	4.31%	7.09%
95 to 99 years	.	4.90%	.	5.41%	4.68%	2.89%	8.16%
Total	0.89%	2.82%	0.73%	1.34%	0.40%	0.83%	0.39%
t-test of difference between rates in Quebec and Canada							
less than 40 years	-1.3992	0.3709	-2.3159	-0.2394	0.1002	1.0728	-1.2011
40 to 64 years	1.1932	-0.2533	0.9679	-0.9549	-1.7444	-1.9855	0.1098
65 years and over	-1.0975	-1.2604	2.7038	0.5042	3.1001	-1.6169	0.8102

Table 2.8: Self-reported incidence of illnesses in the NPHS (1994-2011)

2.3.2 Disability and institutionalization

We had wanted to construct a measure of disability that included cognitive and physical incapacities as well as institutionalization. Thus, we first used the variables from the questionnaire to build a measure of cognitive impairment, a measure of physical incapacity and a measure of institutionalization. We present each of these measures.

Cognitive impairment

We use a self-reported variable indicating whether the individual has frequently had memory problems that negatively affect their activities.

Number of activities of daily living limitations (ADL)

We create a variable to denote the number of activities of daily living limitations reported by the individual.

Institutionalization

The NPHS indicates clearly whether the individual is in an institution. The institutions, or establishments, are defined as follows in the NPHS:

1. Institutions for elderly persons, including care facilities that welcome elderly persons and long-term care hospitals;

2. Institutions for mental health, including facilities that welcome children with emotional problems and persons presenting mental troubles or developmental delay and psychiatric hospitals;

3. Other institutions for rehabilitation, including rehabilitation centres, paediatric hospitals and other specialized hospitals, hospitals with long-term care units, as well as care facilities welcoming physically handicapped persons.

Incapacities variable constructed for the model

The variable created for the model combines these three measures. We define 4 states: 1) No incapacity; 2) Only one ADL or cognitive incapacity; 3) 2 ADLs or more, with or without cognitive incapacity; 4) In institution.

Table 2.9 presents the distribution of this variable by age group. Among those aged 40-44 years, few respondents suffer from incapacities, or less than 5%. This rate increases to 9.7% for those aged 65-69 years. About 58% of those aged 90-94 years suffer from an incapacity. Finally, in Quebec, the proportion of individuals with no incapacity is higher than in the rest of Canada, and this holds across all age groups. This conclusion also applies to the share of Quebecers aged 40-64 years that is in an institution.

In table 2.10, we show two transition matrices between the states that we have defined. One is calculated for respondents aged under 65 years and the other for those aged 65 years and over. These transition rates are calculated for those respondents having survived from one cycle to the next. Among persons under the age of 65, there are few transitions from one period to the next for those without a disability. Indeed, 98% of persons with no incapacity in the current period do not have one in the next period. Meanwhile, only 42% of persons with one incapacity remain in this state in the next cycle. After 65 years, the transitions to different states of incapacity become more numerous. As shown in the table, only 38% of persons with an incapacity in the current period still have one in the following period. The others report either two or more incapacities (21%) or none (37%). Thus, it is possible for an individual reporting one incapacity in the current period to not have it in the following period. Since physical incapacities stem from limitations with regard to activities such as preparing meals, it is possible that taking medication, for example, would allow an individual to not have problems anymore with a certain activity. We also observe that there are few differences between Quebec and the rest of Canada in terms of transitions from one period to another, and this holds across all age groups.

Age	No incapacity	1 ADL or cognitive impairment	2+ ADL	Institution
20 to 24 years	96.34%	2.68%	0.97%	.
25 to 29 years	95.85%	3.09%	1.03%	.
30 to 34 years	96.43%	2.51%	1.06%	.
35 to 39 years	95.92%	2.56%	1.52%	.
40 to 44 years	95.28%	2.78%	1.92%	.
45 to 49 years	94.33%	2.90%	2.66%	0.10%
50 to 54 years	93.51%	3.52%	2.90%	0.07%
55 to 59 years	92.47%	3.80%	3.64%	0.08%
60 to 64 years	91.23%	4.76%	3.93%	0.09%
65 to 69 years	90.30%	4.73%	4.82%	0.15%
70 to 74 years	85.56%	7.43%	6.44%	0.58%
75 to 79 years	78.27%	10.12%	9.96%	1.65%
80 to 84 years	68.50%	12.18%	15.32%	4.00%
85 to 89 years	53.69%	13.85%	23.93%	8.53%
90 to 94 years	41.81%	12.37%	30.37%	15.45%
95 to 99 years	21.84%	9.14%	41.01%	28.01%
Total	91.80%	4.10%	3.62%	0.48%
t-test for difference between rates in Quebec and Canada				
less than 40 years	-4.5711	4.5356	1.3265	1.1331
40 to 64 years	-6.6771	4.5927	5.1438	-2.1051
65 years and over	-1.6175	0.9178	0.0735	2.2091

Table 2.9: State of disability according to NPHS (1994-2011)

2.3.3 Risk factors

We have considered two risk factors for the time being, namely obesity and tobacco use. In terms of tobacco use, we have defined three possible states: 1) never smoked; 2) current smoker; and 3) former smoker. Table 2.11 shows the distribution of respondents by age. In table 2.12 we show, by age group, the initiation and quit rates, or transitions between states of tobacco use. We see that the rate of tobacco use is about 25% between 20 and 24 years, then reaches a peak of over 30% between 30 and 34 years, then declines with age. The share of former smokers rises up to the age of 80 years and declines thereafter, which probably reflects cohort effects in tobacco habits. Initiation rates are higher among youth, while quit rates are higher at the two ends of the age distribution — a dynamic which also no doubt masks certain cohort effects.

There are notable (and statistically significant) differences in the shares of current and former smokers in Quebec and Canada. In all age groups, the share of current and former smokers is higher in Quebec than in Canada.

As for obesity, we have constructed a variable based on the body mass index (BMI) with three categories (under 30, 30-35 and 35+). In the vocabulary of the World Health Organization, an individual with a BMI of 30 or more is considered obese. In our model, class I obesity groups together individuals with a BMI of 30 to 35 (class 1 obesity of the WHO), while classes II-III group together all respondents with a BMI of 35 and over (classes II and II of WHO) (World Health Organization, 2006). Table 2.13 gives the distribution of respondents by age group. We see that class I obesity is at its highest in the 60s, while classes II-III is highest in the 50s.

Under 65 years				
	Disability in the following period			
Disability in the current period	No disability	1 ADL or cognitive impairment	2+ ADL	Institution
No incapacity	97.55%	1.69%	0.75%	0.01%
1 ADL or cognitive impairment	49.93%	42.19%	7.78%	0.10%
2+ ADL	27.56%	12.53%	59.56%	0.35%
Institution	.	.	.	.
	t-test for the difference between rates in Quebec and Canada			
No incapacity	-1.293	1.6794	0.0974	-0.2918
1 ADL or cognitive impairment	-2.0808	2.2628	0.4195	0.8894
2+ ADL	-4.3156	1.0613	3.1151	0.0207
Institution	0.6176	0.6176	0.6176	-1.1072

65 years and over				
	Disability in the following period			
Disability in the current period	No disability	1 ADL or cognitive trouble	2+ ADL	Institution
No incapacity	88.61%	5.88%	4.82%	0.70%
1 ADL or cognitive impairment	37.38%	38.48%	21.00%	3.14%
2+ ADL	16.86%	10.88%	62.92%	9.33%
Institution	6.47%	.	.	92.25%
	t-test for difference between rates in Quebec and Canada			
No incapacity	0.6783	0.5198	1.6862	0.2358
1 ADL or cognitive impairment	0.0340	0.4836	0.5498	0.1221
2+ ADL	-1.1897	-0.2968	0.0194	1.6029
Institution	1.8798	.	2.1114	2.5498

Table 2.10: Transitions between states of disability according to the NPHS (1994-2011)

Table 2.14 presents the individual transitions between levels of obesity. The rows indicate the level of obesity in the current cycle, while columns 2, 3 and 4 complete the crosstab with respect to the following cycle. On average, a little more than 3% of non-obese individuals become obese in each 2-year cycle; 17.7% of individuals with class I obesity become non-obese in each cycle, while 8.1% go on to class II-III obesity; and 17.2% of individuals with classes II-III obesity become obese class I while 2.9% become non-obese.

2.3.4 Health care use

The NPHS is rich in data on health care use (Statistics Canada, 2012d). Five variables were created for modelling purposes. The first two variables are the number of visits or telephone consultations over the 12 previous months, on the one hand with a family doctor, paediatrician or general practitioner; and on the other hand, with another doctor or specialist (such as surgeon, allergist, orthopaedist, gynaecologist or psychiatrist). These two variables were censored at the 99^{th} percentile because the last percentile had very many observations, which might unduly impact results at the average.

Since doctor consultations do not include stays in a care facility, a third variable was created to account for these: the number of nights in a hospital, in nursing care or in a convalescent home. As was the case for medical consultations variables and for the same reason, this variable was censored at the 99^{th}

Age	Current smokers	Former smokers
20 to 24 years	24.26%	14.04%
25 to 29 years	28.48%	19.95%
30 to 34 years	30.63%	21.35%
35 to 39 years	28.95%	25.95%
40 to 44 years	28.54%	28.87%
45 to 49 years	26.89%	32.11%
50 to 54 years	25.04%	35.29%
55 to 59 years	22.42%	38.66%
60 to 64 years	19.37%	41.23%
65 to 69 years	16.26%	44.69%
70 to 74 years	13.73%	45.90%
75 to 79 years	11.03%	47.30%
80 to 84 years	7.10%	45.83%
85 to 89 years	4.60%	40.45%
90 to 94 years	3.02%	36.14%
95 to 99 years	.	40.61%
Total	.	31.35%
t-test for difference between rates in Quebec and Canada		
under 40 years	-5.4252	-1.0811
40 to 64 years	-7.7491	-4.5301
65 years and over	-8.8095	-6.2277

Table 2.11: Tobacco use in the NPHS (1994-2011)

Age	Initiation	Quit	Reuptake
20 to 24 years	4.67%	20.37%	18.94%
25 to 29 years	3.09%	15.63%	20.69%
30 to 34 years	2.72%	13.19%	10.96%
35 to 39 years	2.54%	12.15%	9.08%
40 to 44 years	1.95%	12.40%	8.35%
45 to 49 years	2.23%	10.90%	6.33%
50 to 54 years	2.40%	12.92%	4.79%
55 to 59 years	2.18%	14.10%	3.84%
60 to 64 years	1.65%	14.58%	3.37%
65 to 69 years	2.56%	18.13%	3.54%
70 to 74 years	2.25%	17.94%	2.94%
75 to 79 years	2.19%	20.85%	1.56%
80 to 84 years	1.05%	16.66%	1.65%
85 to 89 years	1.10%	24.81%	1.56%
Total	2.60%	14.26%	6.71%
t-test for difference between rates in Quebec and Canada			
< 40	0.6266	1.1721	1.5307
40-64	-1.402	2.109	1.9207
>65	-0.6881	0.4028	-1.3191

Table 2.12: Initiation and quit rates according to the NPHS (1994-2011)

percentile.

The fourth variable is an indicator of use of medications, consisting of a binary variable indicating the consumption of at least one medication in the previous 12 months. We use this variable because we

Age	BMI<30	30 ≤ IMC <35	BMI ≥ 35
20 to 24 years	91.39%	5.65%	1.98%
25 to 29 years	87.24%	8.97%	2.53%
30 to 34 years	85.36%	10.29%	3.10%
35 to 39 years	83.46%	11.41%	3.87%
40 to 44 years	81.93%	12.46%	4.35%
45 to 49 years	81.08%	13.13%	4.64%
50 to 54 years	78.90%	14.41%	5.52%
55 to 59 years	78.88%	14.68%	5.25%
60 to 64 years	78.05%	15.95%	5.10%
65 to 69 years	78.98%	15.67%	4.35%
70 to 74 years	81.69%	14.25%	3.33%
75 to 79 years	84.20%	12.75%	2.38%
80 to 84 years	89.16%	9.08%	1.31%
85 to 89 years	92.35%	6.34%	1.20%
90 to 94 years	92.48%	6.13%	1.17%
Total	83.16%	11.93%	3.82%
t-test for difference between rates in Quebec and Canada			
less than 40 years	-9.8752	7.4653	6.1378
40 to 64 years	-12.5711	8.7684	8.3062
65 years and over	-0.2379	-0.2379	0.9002

Table 2.13: Body mass index (BMI) according to the NPHS (1994-2011)

	Obesity in following cycle		
Obesity in current cycle	BMI<30	30 ≤ BMI <35	BMI ≥ 35
BMI<30	96.47%	3.18%	0.16%
30 ≤ IMC <35	17.65%	74.15%	8.09%
BMI ≥ 35	2.90%	17.21%	79.76%
Total	85.37%	10.90%	3.56%
t-test for difference between rates in Quebec and Canada			
BMI<30	-5.1179	5.1754	1.7965
BMI 30-35	-0.9912	-1.6353	0.804
BMI>35	-0.4775	-0.3086	1.7589

Table 2.14: Transitions between states of obesity according to the NPHS (1994-2011)

do not have data on the amount of medications consumed.

The last variable concerns home care services. Since the main objective here is to evaluate the use of public resources, the variable that we create is an indicator of home care services used in the previous 12 months, but for which the cost is at least partially borne by the government. The NPHS gives examples of home care services, which can be as varied as nursing care, help to take a bath or delivery of meals.

Table 2.15 gives the distribution by age. The number of consultations with a general practitioner increases fairly constantly with age, especially beyond 45 years (one may guess that, before this age, pregnancy increases the number of visits). Over the course of the reference 12-month period, individuals consult a general practitioner three times on average. Consultations with specialists in the previous 12 months, however, drop to under one. The number of nights of hospitalization increases faster after 75 years of age. Individuals aged 85 to 89 years spend an average of more than 3 nights in a short-term facility over the previous 12 months. Between 80% and 93% of the entire population responded that they had taken at least one medication in the reference period. Finally, after 85 years of age, more than 25% of the population receive home care services.

Age	Nb. consultations generalist	Nb consultations specialist	Nb hosp. nights
20-24	2.64	0.55	0.36
25-30	2.75	0.69	0.38
30-34	2.77	0.72	0.46
35-39	2.68	0.68	0.40
40-44	2.54	0.65	0.36
45-49	2.64	0.66	0.41
50-54	2.98	0.76	0.64
55-59	3.15	0.77	0.68
60-64	3.32	0.80	0.73
65-69	3.76	0.85	1.10
70-74	4.19	0.88	1.43
75-79	4.46	0.86	2.11
80-84	5.06	0.84	2.42
85-89	5.05	0.69	3.57
90-94	5.29	0.52	3.53
95-99	4.35	0.42	3.98
Total	3.07	0.72	0.71
Age	**1+ medication**	**Home care services (yes)**	
20-24	80.91%	0.50%	
25-30	80.58%	1.19%	
30-34	79.77%	1.19%	
35-39	79.96%	1.30%	
40-44	81.60%	0.91%	
45-49	81.70%	1.16%	
50-54	82.93%	1.40%	
55-59	85.10%	1.72%	
60-64	87.27%	2.60%	
65-69	89.20%	4.01%	
70-74	90.37%	7.04%	
75-79	91.32%	11.35%	
80-84	93.34%	17.91%	
85-89	90.88%	25.45%	
90-94	90.95%	38.41%	
95-99	92.17%	33.73%	
Total	83.51%	2.93%	

Table 2.15: Health care use by age according to NPHS (1994-2011)

Table 2.16 indicates that the average number of consultations with a generalist is about twice as high among individuals suffering from one of the health conditions considered. The difference is statistically significant in every case. Individuals who suffer from diabetes, cancer, heart diseases, stroke or dementias consult a specialist more than once a year on average. This difference is statistically significant. Those who suffer from any form of dementia, a heart disease, or who have had a stroke spend more than five times as many nights in hospital as those who do not. This difference is also statistically significant. Among those with diabetes, hypertension, a heart disease or a stroke, the prevalence of having consumed at least once medication is highest, at 96%. It seems that consumption of at least one medication does not increase by much with age. However, due to the definition of this variable, a respondent who is in overall good health but who consumes a single medication in the entire year cannot be distinguished from a respondent suffering from many illnesses and consuming several medications. Thus, 80% of individuals aged 20 to 24 years had consumed at least one medication during the reference period. The more than 10-percentage point increase between the ages of 20 and 99 years of the proportion of individuals who consumed at least one medication is, once the definition of the variable is accounted for, fairly important after all. About 21% of people having had a stroke or who suffer from a form of dementia receive home care services, about 10 times higher than those who never suffered from these — again here, there is a statistical difference.

2.3.5 Other socio-economic variables

We have constructed socio-economic variables that will be useful in the analysis.

- Immigrant: dummy variable equal to 1 if the individual is an immigrant and 0 otherwise.
- Woman: dummy variable equal to 1 if the individual is a woman and 0 otherwise.
- Education: three dummy variables for the highest level reached among secondary completion, a college diploma and a university diploma.

We have chosen to only model these variables because it is reasonable to assume that they are fairly stable after the age of 30. If we were to include other variables, such as marital status or income, we would have to model individual behaviours over time. At this stage of the model's conceptualization, such modelling would be too complex.

Table 2.17 presents descriptive statistics drawn from the NPHS regarding these variables. We note that the share of immigrants increases slightly with age, as does the proportion of women — in particular over the age of 60 in this second case. Due to important cohort effects, the proportion of university graduates decreases with age, but after 30 years, an age where individuals are likely to have completed most of their studies.

2.4 CCHS

We also use " public" data, or public use data, from the CCHS (Statistics Canada, 2010).This survey is used to create the initial population in 2010, aged 30 to 110 years. We only use the Quebec observations of the CCHS, because they are sufficiently numerous (about 11 000) to create the initial population. We cannot use the NPHS because its population is only representative of the Canadian population in 1994.

CCHS 2010 shares many variables with the NPHS, which we have been able to use directly. We thus have access to all variables on the presence of illnesses considered in the model, with the exception of Alzheimer's and other dementias. Table 2.18 presents the prevalence of illnesses in CCHS 2010. Similar to what is observed in the NPHS, the prevalence of most illnesses increases with age. The prevalence of lung diseases increases up to 75 years and then decreases slightly thereafter. The most common illness is hypertension, with a prevalence of 48.6% among persons aged 65 to 69 years.

Diseases	Nb. consultations generalist	Nb. consultations specialist	Nb. hosp. nights
Diabetes			
No	2.78	0.63	0.57
Yes	5.40	1.06	2.48
Hypertension			
No	2.59	0.60	0.47
Yes	4.66	0.98	1.74
Cancer			
No	2.84	0.62	0.59
Yes	4.77	1.81	2.59
Heart diseases			
No	2.73	0.61	0.50
Yes	5.22	1.27	2.88
Stroke			
No	2.85	0.65	0.59
Yes	6.04	1.12	4.75
Lung diseases			
No	2.80	0.63	0.60
Yes	4.46	1.00	1.42
Dementias			
No	2.88	0.65	0.63
Yes	5.61	1.13	4.90
Illnesses	1+ medications	Home care services (yes)	
Diabetes			
No	82.40%	2.04%	
Yes	97.86%	9.95%	
Hypertension			
No	80.53%	1.61%	
Yes	96.62%	6.86%	
Cancer			
No	82.79%	2.14%	
Yes	93.64%	10.56%	
Heart diseases			
No	82.10%	1.76%	
Yes	96.87%	11.36%	
Stroke			
No	82.92%	2.12%	
Yes	97.91%	21.01%	
Lung diseases			
No	82.56%	2.11%	
Yes	92.14%	6.77%	
Dementias			
No	83.10%	2.30%	
Yes	93.03%	21.63%	

Table 2.16: Use of health care services by presence of illnesses, NPHS (1994-2011): All these differences are statistically significant at the 5% threshold.

Age	Immigrants	Women	Secondary dipl.	College dipl.	University diploma
20-24	9.84%	48.36%	54.23%	17.01%	8.63%
25-29	12.65%	49.33%	42.39%	22.97%	21.48%
30-34	16.93%	50.29%	43.39%	24.75%	20.96%
35-39	19.35%	50.52%	44.63%	24.21%	19.49%
40-44	20.99%	50.37%	43.44%	23.92%	19.51%
45-49	22.39%	49.33%	42.06%	23.05%	20.31%
50-54	23.76%	48.59%	40.64%	20.31%	19.97%
55-59	25.11%	49.22%	38.86%	17.59%	18.89%
60-64	24.85%	51.06%	36.57%	15.81%	15.69%
65-69	25.15%	53.92%	34.05%	14.24%	11.95%
70-74	28.16%	56.13%	31.59%	13.43%	10.11%
75-79	27.62%	59.99%	30.71%	10.93%	7.90%
80-84	28.30%	62.58%	29.51%	8.88%	8.28%
85-89	25.64%	66.32%	29.40%	8.23%	8.46%
90-94	28.09%	69.60%	31.11%	8.68%	7.52%
95-99	30.27%	65.47%	32.38%	3.43%	6.77%
Total	20.72%	51.17%	41.20%	19.83%	16.98%

Table 2.17: Socio-economic characteristics drawn from the NPHS (1994-2011)

Age	Diabetes	Hypertension	Cancer	Heart diseases	Stroke	Lung diseases
20 to 24 years	1.01%	3.02%	0.41%	0.78%	0.00%	0.00%
25 to 29 years	0.64%	3.59%	1.22%	0.57%	0.12%	0.00%
30 to 34 years	0.63%	5.24%	1.57%	1.51%	0.02%	0.00%
35 to 39 years	1.52%	7.03%	1.89%	1.11%	0.39%	1.53%
40 to 44 years	3.16%	12.63%	2.69%	0.93%	0.15%	1.00%
45 to 49 years	2.94%	12.11%	4.46%	1.90%	0.27%	3.71%
50 to 54 years	4.70%	23.39%	6.06%	3.29%	1.02%	2.79%
55 to 59 years	8.76%	36.54%	7.53%	9.66%	0.68%	4.52%
60 to 64 years	9.74%	38.95%	12.64%	6.69%	1.47%	5.41%
65 to 69 years	16.05%	48.61%	15.83%	18.12%	2.75%	6.33%
70 to 74 years	16.01%	49.51%	18.97%	20.34%	2.66%	9.79%
75 to 79 years	17.33%	63.16%	20.59%	23.61%	5.20%	8.74%
80 years and over	17.26%	56.81%	22.09%	25.11%	5.98%	7.30%
Total	6.10%	23.02%	7.00%	6.34%	1.10%	3.21%

Table 2.18: Self-reported prevalence of illnesses in CCHS 2010

3 Initialization

3.1 Overview

Initialization consists of creating a database for the initial simulation year (2010 by default). We cannot use NPHS as the initial database because this survey was designed to be representative of the Canadian population at the time of sampling (1994). Therefore, it is not representative of the population's state of health in 2010. Moreover, the sample of Quebec respondents is too small for us to create the initial population in a reliable manner.

We thus use the 2010 CCHS as the database to construct the initial population. This survey has the advantage of presenting an up-to-date picture of the state of health of the Canadian population. Also, it contains a large number of observations (11 271 for Quebec alone), which allows us to use only Quebec observations from the survey. The disadvantage of the CCHS compared to the NPHS is that, as opposed to the latter, the former is not longitudinal in nature. This characteristic, however, is not needed to create an initial database.

Using this second survey introduces a difficulty, however: the initial database must contain all variables needed for the transition and health care use models. For the majority of variables that we use, this is not a problem since the NPHS and CCHS variables are identical. However, in certain cases, it is necessary to impute values for certain variables that are not directly available in the CCHS. Moreover, we use the public use version of the CCHS, which means that the data is limited for certain variables. This chapter describes the assumptions which allow us to create the initial database — and thus population — using the 2010 CCHS.

3.2 Imputation

3.2.1 Imputation of age

The public use CCHS categorizes age of individuals by 3- to 4-year intervals.[5]

To overcome this problem, we attribute a precise age to each individual within his/her age group, in a random and uniform manner, then calibrate the survey weights in order to mirror the age distribution of the Quebec population by sex provided by the ISQ (see Institut de la statistique du Québec 2009).

3.2.2 Imputation of education level

The education levels available in the public use CCHS are " less than secondary diploma", " secondary diploma" and " post-secondary studies". In order to have similar levels to those of the NPHS, we have to impute university diplomas among individuals in the last category. While it would have been possible to use the NPHS to impute the education level, the small number of respondents from Quebec in the survey dissuaded us from doing so.

We instead use public use data from the 2010 Labour Force Survey (LFS) of Statistics Canada, which provides a detailed picture of education levels among the population in 2010. We only retain Quebec observations that have at least a college diploma, and then we estimate a logit model to obtain the effects of individual characteristics on the probability of having a university degree, conditional on having at least a college diploma. The structure of the model is as follows:

$$y_i^* = \boldsymbol{\beta}\mathbf{x}_i + \epsilon_i \tag{1}$$

$$y_i = \begin{cases} 1 & \text{si} \quad y_i^* \geq 0 \\ 0 & \text{si} \quad y_i^* < 0 \end{cases}, \tag{2}$$

[5]We use the public use CCHS data to construct the initial database to ensure that we can initialize the model at will outside the walls of a Statistics Canada Research Data Centre (RDC). Use of restricted access data would require the user to run the model within an RDC.

where y_i is equal to 1 if individual i has a university degree and 0 otherwise; y_i^* is the latent variable; $\mathbf{x}_i$ is the vector of explanatory variables of individual i; and ϵ_i is a random term that follows a logistic distribution.

Table 3.1 presents the variables used and the estimation results. Since the model used is not linear, it is not the parameters themselves which are interesting to analyze the impact of the explanatory variables, but rather the marginal effects, or the change in the conditional expected value of y_i when the value of a variable x_i changes by one unit. In the COMPAS framework, we use the average marginal effect (AME). The AME is an average across all individuals of the effect of a change of one unit of a variable x_i on the expected value of y_i.[6] We include a spline for each age group to capture the non-linearity of the age variable, a constant effect of female gender, as well as interaction effects between sex and these age effects. These interaction effects allow the effect of being a woman on the probability of having a university diploma to vary by age. This is important given the significant changes observed between the generations with regard to the gap in education level between men and women.

	Marginal effect
Age (if 25 years and under)	0.0745 ***
Age (if 26 to 30 years)	0.0113 ***
Age (if 31 to 35 years)	-0.0090 *
Age (if 36 to 40 years)	0.0077
Age (if 41 to 45 years)	-0.0181 ***
Age (if 46 to 50 years)	-0.0012
Age (if 51 to 55 years)	0.0148 ***
Age (if 56 to 60 years)	-0.0012
Age (if 61 to 65 years)	0.0026
Age (if 66 years and over)	-0.0090 *
Woman	0.3304 **
Woman*[Age (if 25 years and under)]	-0.0089
Woman*[Age (if 26 to 30 years)]	-0.0085 **
Woman*[Age (if 31 to 35 years)]	0.0145 **
Woman*[Age (if 36 to 40 years)]	-0.0180 ***
Woman*[Age (if 41 to 45 years)]	0.0008
Woman*[Age (if 46 to 50 years)]	0.0203 ***
Woman*[Age (if 51 to 55 years)]	-0.0246 ***
Woman*[Age (if 56 to 60 years)]	0.0197 ***
Woman*[Age (if 61 to 65 years)]	-0.0075
Woman*[Age (if 66 years and over)]	-0.0040
Number of observations	110,626

Table 3.1: Average marginal effects of variables on the probability of having a university diploma for individuals having completed post-secondary studies – 2010 public use LFS

Figure 3.1 presents more clearly the estimated probabilities by age and sex. The age effects for the

[6]The models used in this chapter as well as in chapter 4 all have a non-linear form. Thus, we always present the MEM rather than the estimated parameters.

youngest are positive and significant; this simply captures the fact that older students are more advanced in their studies. The age effect then becomes negative, reflecting the increase in education level among the youngest cohort. Finally, the age effects are ambiguous at higher ages, because the most educated persons tend to live longer. The effect of being a woman is positive and significant for the youngest, because younger women tend to attend university more than men do. The effect of being a woman decreases with age, however, and becomes negative after 50 years.

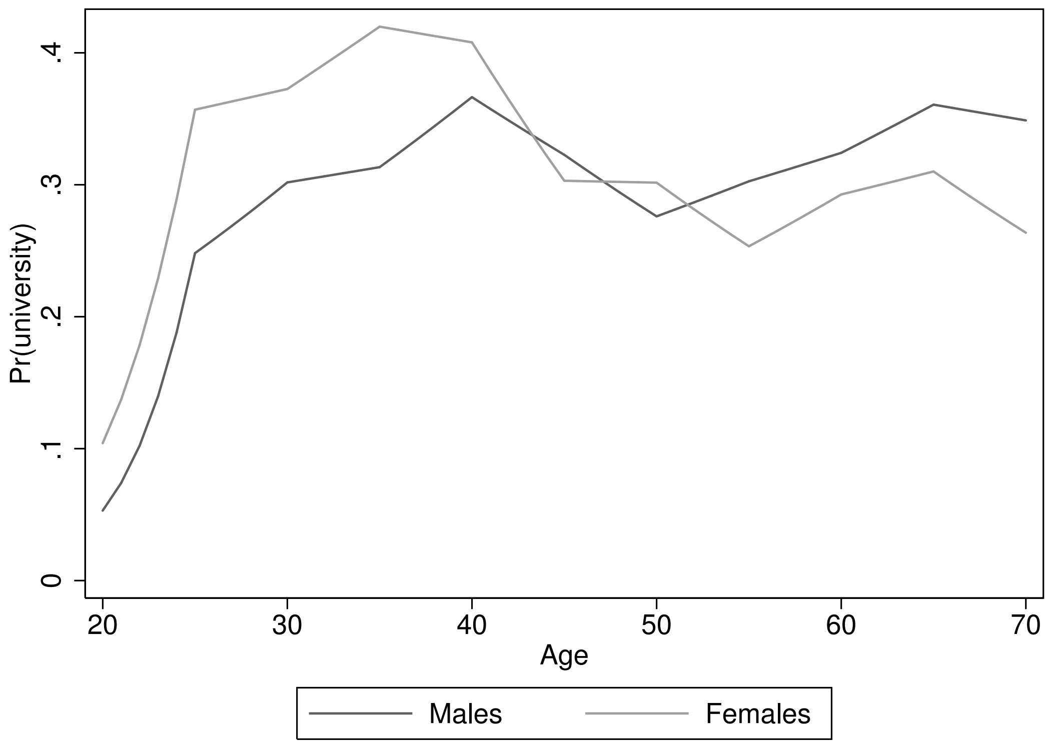

Figure 3.1: Probability of having a university degree for individuals having completed post-secondary studies — 2010 public use LFS

The model coefficients are then used to impute education levels in the CCHS. For each observation i with post-secondary studies, we estimate the probability of having a university diploma using the following formula:

$$\hat{P}(univ_i = 1|\mathbf{x_i}) = \frac{\exp\left(\mathbf{x_i'}\hat{\boldsymbol{\beta}}\right)}{1 + \exp\left(\mathbf{x_i'}\hat{\boldsymbol{\beta}}\right)} \tag{3}$$

We then draw a random variable from a uniform distribution (0,1) for each observation and impute a university degree if this is smaller than the probability calculated above.

3.2.3 Imputation of disability

The CCHS allows us to count the number of activities of daily living limitations among the following:

- need help to prepare meals;
- need help shopping;
- need help for personal care;
- need help to get around the house.

Compared to the ones we use from the NPHS to compute the transitions (see following chapter), the only missing limitation is the need for help to accomplish household chores. We ignore this problem for the time being and calculate the number of other limitations to obtain the same disability variables as in the NPHS (only one limitation or cognitive impairment; two limitations or more; in institution).

The CCHS does not allow us to observe whether the person is in an institution. We must therefore impute the institutionalization status. The imputation is also made using a logit model, this time estimated on 2010-2011 NPHS data. We include a spline at 75 years in order to allow the age effect to change after 75 years; a sex effect; an immigrant status effect; and two binary variables for the first two types of disability (one limitation or cognitive problem; two limitations or more). Table 3.2 shows the results.

	Marginal effect
Age (if under 75 years)	0.0003 **
Age (if 75 years or more)	0.0008 ***
Immigrant	-0.0033 *
Woman	-0.0014
Disability: 1 ADL or cognitive impairment	0.0097 ***
Disability: 2 ADL or +	0.0150 ***
Number of observations	14,598

Table 3.2: Average marginal effects of variables on the probability of being in an institution — NPHS 2010-2011

We see that the age effects are significant. The effect for women is not significant, but that for immigrants is slightly negative. The dummy variables for the other states of disability have large coefficients, because individuals who suffer from physical or cognitive limitations are more likely to be in an institution.

We subsequently use the estimated coefficients to calculate the probability of being in an institution for each individual observed in the CCHS and to proceed with the imputation.

3.2.4 Imputation of Alzheimer's and other dementias

The prevalence of Alzheimer's and other dementias is the only chronic health problem that we do not observe in the public use CCHS. We therefore impute the variable using a logit estimated with 2010-2011 NPHS data. The model is similar to that used for institutionalization, except that it includes a binary variable for institutionalization status as an explanatory variable. The estimation results are presented in table 3.3.

We find that the probability of suffering from dementias increases strongly with age and stops increasing significantly after 75 years. Sex and immigration status do not have a significant effect. As we might expect, the disability statuses are for their part strongly associated with the presence of dementias.

	Marginal effect
Age (if under 75 years)	0.0007 ***
Age (if 75 years or more)	0.0002
Immigrant	0.0010
Woman	-0.0038 *
Disability: 1 ADL or cognitive impairment	0.0122 ***
Disability: 2+ ADL	0.0250 ***
Disability: in institution	0.0225 ***
Number of observations	14,598

Table 3.3: Average marginal effects of variables on the probability of having Alzheimer's or another dementia — 2010 NPHS

The model coefficients are then used to impute the presence of dementias in the CCHS. Thus, for each observation i, we estimate the probability of suffering from dementias using the following formula:

$$\hat{P}(d\acute{e}mences_i = 1|\mathbf{x_i}) = \frac{\exp\left(\mathbf{x_i'}\hat{\boldsymbol{\beta}}\right)}{1+\exp\left(\mathbf{x_i'}\hat{\boldsymbol{\beta}}\right)} \tag{4}$$

where $\mathbf{x}_i$ is the vector of explanatory variables of individual i and β is the vector of coefficients. In order to determine whether an observation i will be imputed as having dementia, we draw a random variable from a (0,1) uniform distribution for each observation. If the latter is less than the calculated probability, observation i is imputed as having dementia.

3.2.5 Characteristics of the initial population

The initial population of the model, following the imputations performed on 2010 CCHS data, contains, with one exception, all variables needed for the transition and health care use models. Table 3.4 presents certain statistics regarding the variables of this population.

	Average	Standard deviation	Minimum	Maximum
Age	53.80	15.14	30	110
Immigrants	14.94%	35.65%	0	1
Women	51.30%	49.98%	0	1
No diploma	19.38%	39.53%	0	1
Secondary diploma	17.76%	38.22%	0	1
College diploma	42.43%	49.42%	0	1
University diploma	20.44%	40.32%	0	1
Presence of diabetes	7.03%	25.57%	0	1
Presence of hypertension	27.23%	44.51%	0	1
Presence of heart diseases	7.54%	26.41%	0	1
Presence of stroke	1.04%	10.15%	0	1
Presence of cancer	8.26%	27.53%	0	1
Presence of lung diseases	3.89%	19.33%	0	1
Presence of dementias	1.32%	11.40%	0	1
Never smoked	30.23%	45.93%	0	1
Current smokers	23.59%	42.46%	0	1
Former smokers	46.18%	42.46%	0	1
Healthy weight	82.76%	37.78%	0	1
Class I obesity	12.54%	33.12%	0	1
Classes II and III obesity	4.70%	21.17%	0	1
Disability: no incapacity	89.76%	30.32%	0	1
Disability: 1 ADL or cognitive impairment	4.63%	21.02%	0	1
Disability: 2+ ADL	4.87%	21.52%	0	1
Disability: in institution	0.74%	8.58%	0	1

Table 3.4: Description of variables in the initial population

4 Transitions

4.1 Overview

The transition models make it possible to calculate the probability that an individual changes state of health or behaviour as a function of their individual characteristics. Such models require a priori estimation of the parameters of discrete choice econometric models, with the transitions that we want to model serving as dependent variables. In COMPAS, these parameters are estimated using data from the NPHS because the longitudinal nature of this survey makes it possible to observe individuals' transitions between the different states. Then, the estimated parameters are used to calculate the transitions probabilities of individuals in the simulations as a function of their individual characteristics, which then allows us to simulate the future evolution of their health conditions.

In section 4.2, we present the econometric models used in COMPAS. These are used to estimate the effects of individual characteristics on transitions occurring between cycles of the NPHS. By default, these cycles are 2-year cycles; as such, the individual probabilities of changing state are calculated for a 2-year interval.

4.2 Econometric models

4.2.1 Models for illnesses

A transition model is used to estimate the probabilities of incidence of each disease considered. The probabilities are a function of age, different risk factors, socio-economic characteristics, and in certain cases, the prevalence of other diseases. Restrictions imposed include that some illnesses do not impact the incidence of other illnesses, so as to not create false links between these diseases. The restrictions were evaluated by a panel of experts in the context of the Future Elderly Model (FEM) and are based on medical research on the relationship between the different diseases (Goldman et al., 2005).

Table 4.1 shows the different effects permitted in the model. For example, a "×" at the intersection of the row with diabetes and the column with hypertension means that we allow the presence of diabetes to have an effect on the probability of starting to suffer from hypertension. Moreover, the diseases are considered as being " absorbing " states. This means that once an individual suffers from an illness, he/she has it until the end of his/her life (see explanations on prevalence provided in chapter 2).

	Diabetes	Hypertension	Cancer	Lung diseases	Heart diseases	Stroke	Dementias
Diabetes	n.a.	x			x	x	
Hypertension		n.a.			x	x	
Cancer			n.a.			x	
Lung diseases				n.a.			
Heart diseases					n.a.	x	
Stroke						n.a.	
Dementias							n.a.

Table 4.1: Permitted effects of diseases on each other

The specification is as follows:

$$inc^*_{i,j,t+1} = \boldsymbol{\lambda}_j \mathbf{y}_{i,t} + \boldsymbol{\beta}_j \mathbf{x}_{i,t} + \epsilon_{i,j,t}, \tag{5}$$

with

$$inc_{i,j,t+1} = \begin{cases} 1 & \text{if} \quad inc^*_{i,j,t+1} > 0 \\ 0 & \text{if} \quad inc^*_{i,j,t+1} \leq 0 \end{cases} \tag{6}$$

where:

- $inc^*_{i,j,t+1}$ is a latent variable of the incidence of illness j for individual i in period $t+1$;
- $inc_{i,j,t+1}$ is a binary variable indicating the incidence of illness j for individual i in period $t+1$;
- $\mathbf{y}_{i,t}$ is a vector of binary variables indicating the presence or absence of each illness in individual i in period t;
- $\boldsymbol{\lambda}_j$ is a vector that includes the effects of the different diseases on the probability of incidence of illness j, accounting for the links permitted as described in table 4.1;
- $\mathbf{x}_{i,t}$ includes all explanatory variables accounted for;

- $\boldsymbol{\beta}_j$ includes the effects of these variables on illness j;
- $\epsilon_{i,j,t}$ is a random term specific to individual i and illness j in period t.

We estimate the parameters only on the population that does not yet suffer from illness j in period t. We assume that the distribution function of $\epsilon_{i,j,t}$ follows an extreme value distribution, which enables us to use a complementary log-log model. This model is different from the discrete choice logit and probit models, which are more commonly used, in that it relaxes the assumption of symmetry of the $\epsilon_{i,j,t}$ around 0, which makes it a better specification when the incidence of the dependent variable occurs only rarely. The parameters so estimated can then be used to project the probabilities of individual transitions in the simulation by the following formula:

$$\hat{P}(inc_{i,j,t+1} = 1|\mathbf{y_i}, \mathbf{x_i}) = 1 - \exp(-\exp(\boldsymbol{\lambda}_j \mathbf{y}_{i,t} + \boldsymbol{\beta}_j \mathbf{x}_{i,t})). \tag{7}$$

	Diabetes	Hypertension	Cancer	Heart diseases
Age (if 50 years or under)	0.0008 ***	0.0024 ***	0.0007 ***	0.0010 ***
Age (if over 50 years)	0.0002 ***	0.0003 ***	0.0003 ***	0.0006 ***
Smoker	-0.0002	0.0046 **	0.0004	0.0046 ***
Former smoker	0.0010	0.0037 **	0.0005	0.0019 *
Class I obese	0.0101 ***	0.0158 ***	-0.0006	0.0025 *
Classes II-III obese	0.0169 ***	0.0208 ***	0.0041 ***	0.0060 ***
Woman	-0.0022 ***	-0.0022	0.0002	-0.0033 ***
Immigrant	0.0002	-0.0002	-0.0019 **	-0.0021 *
Secondary diploma	0.0003	-0.0029 *	0.0009	-0.0017
College diploma	-0.0003	-0.0012	0.0020 **	-0.0008
University diploma	-0.0037 ***	-0.0036	-0.0001	-0.0031 *
Resides in Quebec	0.0010	0.0015	-0.0005	0.0006
Presence of diabetes	—	0.0074 ***	—	0.0050 ***
Presence of hypertension	—	—	—	0.0075 ***
Presence of heart diseases	—	—	—	—
Presence of cancer	—	—	—	—
Number of observations	105,502	98,313	106,099	103,918
Average incidence	0.89%	2.82%	0.73%	1.34%

	Stroke	Lung diseases	Dementias
Age (if 50 years or less)	0.0004 ***	0.0001	0.0001
Age (if over 50 years)	0.0003 ***	0.0003 ***	0.0005 ***
Smoker	0.0026 ***	0.0069 ***	0.0008
Former smoker	0.0008	0.0031 ***	0.0011 *
Class I obese	-0.0003	0.0003	-0.0019 **
Classes II-III obese	-0.0029 **	0.0037 **	-0.0014
Woman	-0.0004	0.0024 ***	-0.0002
Immigrant	0.0009 **	-0.0029 ***	0.0010 ***
Secondary diploma	-0.0005	-0.0026 ***	-0.0008
College diploma	-0.0016 *	-0.0032 ***	-0.0014 *
University diploma	-0.0037 ***	-0.0040 ***	-0.0027 ***
Resides in Quebec	-0.0005	0.0008	0.0003
Presence of diabetes	0.0030 ***	—	—
Presence of hypertension	0.0004	-	—
Presence of heart diseases	0.0017 ***	—	—
Presence of cancer	-0.0009	—	—
Number of observations	107,222	104,807	107,709
Average incidence	0.40%	0.83%	0.39%

Table 4.2: Average marginal effects of variables on the probabilities of incidence of illnesses

Table 4.2 presents the average marginal effects of estimations for the different illnesses. The first two effects are age effects, created as follows:

- up to 50 years: $min(age, 50)$;
- over 50 years: $max([age - 50], 0)$.

We impose a spline at 50 years, which means that the effect of age in the model is permitted to change after 50 years. We see in the table that, up to age 50, the probability of incidence of most illnesses increases significantly with age, with the exceptions of lung diseases and dementias. After 50 years, age continues to increase the probability of incidence of illnesses, but less rapidly than before — except for lung diseases and dementias, two categories of illness with a steeper slope in this higher age range.

Being a smoker positively and significantly affects the probability of incidence of certain illnesses, such as hypertension, heart diseases, stroke and lung diseases. Being a former smoker also positively impacts the probabilities of incidence, but less so than for smokers. Obesity has a generally positive effect on the probabilities of incidence, with the exception of Alzheimer's and other dementias.

Women seem less affected by diabetes and heart diseases and more affected by lung diseases. As for immigrants, they have less of a chance of being affected by cancer, heart diseases and lung diseases, but are more likely to suffer from stroke or dementias. A higher level of education is generally associated with lower probabilities of incidence of illnesses, with the exception of cancer. The estimated effects of education may capture two effects which could counter each other: first, a generally better state of health among educated individuals, and second, a higher probability among more educated individuals to consult with a health professional, which increases their incidence of illnesses as observed in the survey data.

The dummy variable for Quebec residents is non-significant in all cases, despite the fact that we find significant differences between the rates of incidence in Quebec and the rest of Canada for certain illnesses in our descriptive statistics (see table 2.8 of chapter 2). This suggests that the observed differences in the incidence of illnesses between Quebec and Canada are mostly captured by the other variables included in the models, and it is thus appropriate to use NPHS data for the whole of Canada instead of only those from Quebec.

The effects of the prevalence of other illnesses on the incidence of illnesses are generally positive and significant when we allow for non-null effects. The exceptions are that the prevalence of cancer and hypertension do not seem to significantly affect the probability of incidence of a stroke.

4.2.2 Model for mortality

Mortality is modeled similarly to the incidence of illnesses, except that we also include disability statuses as explanatory variables and that we do not impose any restrictions on the effects of the different illnesses' prevalence.

Table 4.3 presents the marginal effects of the variables on the probability of dying in the two following years. Unsurprisingly, the probability of dying increases significantly with age, both before an after

50 years of age. It may seem surprising, however, that the increase is less rapid after 50 years, but this can be explained by the fact that death at older ages is better predicted by other variables in the model that are correlated with age. Indeed, the presence of illnesses positively affects the probability of death, and this relationship is statistically significant in most cases (diabetes, cancer, heart diseases, Alzheimer's and other dementias).

	Marginal effect
Age (if 50 years or less)	0.0013 ***
Age (if over 50 years)	0.0009 ***
Presence of diabetes	0.0039 ***
Presence of hypertension	0.0000
Presence of cancer	0.0078 ***
Presence of heart diseases	0.0039 ***
Presence of stroke	0.0015
Presence of lung diseases	0.0009
Presence of dementias	0.0027 *
Smoker	0.0101 ***
Former smoker	0.0047 ***
Class I obese	-0.0035 ***
Classes II-III obese	-0.0005
Woman	-0.0069 ***
Immigrant	-0.0005
Secondary diploma	-0.0034 ***
College diploma	-0.0048 ***
University diploma	-0.0061 ***
Resides in Quebec	-0.0005
Disability: 1 ADL or cognitive impairment	0.0089 ***
Disability: 2+ ADL	0.0152 ***
Disability: in institution	0.0178 ***
Number of observations	124,958
Average mortality	1.1%

Table 4.3: Average marginal effects on probabilities of death

Being a smoker or former smoker also significantly affects the probability of death. The link between obesity and death is ambiguous: this can be explained by the weight loss often occurring in the last stages of life.

Being a woman decreases the probability of death, which is consistent with the higher life expectancy of this group. A higher education level also leads to a significant decrease in the probability of death. The effect of the dummy variable for Quebec is not significant, which corresponds with the fact that Quebecers' life expectancy is comparable with that of Canadians when accounting for other differences.

It is worth noting that, as discussed in chapter 2, mortality rates observed in the NPHS are slightly different from official mortality rates, especially at higher ages. As a result, the number of deaths per year in our model could be biased. In a later version of COMPAS, we will calibrate the estimated probabilities in order to match the mortality rates from Human Mortality Database, which correspond with the mortality rates observed by Statistics Canada.

4.2.3 Models for tobacco use

We estimate three transition models for tobacco use:

1. initiate smoking (estimated for individuals who have never smoked),
2. cease smoking (estimated for smokers),
3. restart smoking (estimated for former smokers).

We again use the complementary log-log model for the three transitions. The presence of illnesses is, however, excluded from the explanatory variables so as to not introduce a simultaneity mechanism in the model. Indeed, tobacco use increases the probability that individuals develop certain illnesses (lung diseases, cancers) and then die as a consequence. Table 4.4 presents the average marginal effects.

	Initiation	Quit	Reuptake
Age (if 50 years or less)	-0.0005 ***	-0.0029 ***	-0.0033 ***
Age (if over 50 years)	-0.0001	0.0044 ***	-0.0024 ***
Class I obese	-0.0042	0.0146	0.0066
Classes II-III obese	-0.0014	0.0307 **	0.0067
Woman	-0.0198 ***	-0.0133 **	-0.0010
Immigrant	0.0013	0.0289 ***	-0.0144 **
Secondary diploma	-0.0045 *	0.0451 ***	-0.0109 *
College diploma	-0.0129 ***	0.0756 ***	-0.0181 ***
University diploma	-0.0213 ***	0.1217 ***	-0.0342 ***
Resides in Quebec	0.0035	0.0021	-0.0053
Number of observations	52,555	32,168	40,235
Average incidence	2.60%	14.26%	6.71%

Table 4.4: Average marginal effects on the probabilities of transition between states of tobacco use

Up to 50 years, age has a negative effect on all transition probabilities. After 50 years, age increases the probability of quitting smoking and decreases that of restarting. As for obesity, we only find a significant impact for classes II-III obesity on smoking cessation. Being a woman reduces the probability of all three transitions, and being an immigrant increases the chances of quitting and decreases those of restarting. Finally, a higher level of education is associated with a higher chance of quitting smoking and a lower probability of initiating or restarting.

4.2.4 Model for disability

We model the probabilities of transition between all states of disability in the model ((“ no incapacity", “ only one activities of daily living limitation or cognitive impairment", “ two or more limitations, with or without cognitive impairment", and “ in institution"). The estimation is done using a multinomial logit model, which allows us to estimate the probabilities that an individual finds themselves in each of the possible states of disability in the following cycle.

We assume that individual characteristics affect the probabilities of being in each state of disability through four latent variables (one for each state):

$$inv^*_{i,j,t+1} = \boldsymbol{\beta}_j \mathbf{x}_{i,t} + \epsilon_{i,j,t},\ \forall j = 1, 2, 3, 4 \tag{8}$$

with

$$Pr(inv_{i,t+1} = j) = Pr(inv^*_{i,j,t+1} > inv^*_{i,k,t+1}),\ \forall k \neq j \tag{9}$$

where:

- $inv^*_{i,j,t+1}$ is the latent variable of the state of disability j for individual i in period $t+1$;
- $inv_{i,t+1}$ is the categorical variable for disability of individual i in period $t+1$;
- $\mathbf{x}_{i,t}$ is the vector of individual characteristics of individual i at period t;
- $\boldsymbol{\beta}_j$ is the vector of parameters of the effects of the $\mathbf{x}_{i,t}$ variables on the latent variable of the state of disability j;
- $\epsilon_{i,j,t}$ is a random term specific to individual i and disability j in period t.

Vector $\mathbf{x}_{i,t}$ includes the presence of all illnesses and risk factors, socioeconomic characteristics, as well as a dummy variable for each state of disability in the preceding period. We assume that $\epsilon_{i,j,t}$ follows a standard type 1 extreme value distribution, which means that the difference between the error terms follows a logistic distribution. Since we use the difference between the values of the latent variables to calculate the probabilities, we have to normalize to 0 the vector of parameters for one of the states. We can nevertheless obtain the marginal effects for all states.

Table 4.5 presents the estimated average marginal effects for each state of disability. Age increases the probability of becoming disabled, especially after the age of 50. The presence of illnesses generally has positive and significant effects on the probabilities of suffering from the different levels of disability. We observe, however, that the effect of certain illnesses on being in an institution is not significant. This is the case for hypertension, cancer, heart diseases and lung diseases.

Being a smoker increases the risks of suffering from all levels of disability, while being a former smoker only increases the risk of suffering from the first level of disability. The effects of obesity are not significant.

Being a woman significantly affects the probabilities of suffering from a disability, while immigrants seem to find themselves in an institution less often. A higher level of education is associated with lower risks of suffering from a disability. The effects of disability dummy variables are obviously strong because individuals tend to remain in their current state of disability in the following period. Thus, the probability of having an incapacity or cognitive problem in the following period is 9.75% higher among individuals with an incapacity or cognitive impairment in the current period. Individuals in an institution in the current period have a 12.70% higher chance of having at least two incapacities in the next period. Finally, Quebecers are somewhat less likely to suffer from the first level of disability than the rest of Canadians.

	No incapacity	1 ADL or cogn. impair.	2+ ADL	In institution
Age (if 50 years of less)	-0.0003 *	-0.0001	0.0003 ***	0.0000
Age (if over 50 years)	-0.0022 ***	0.0008 ***	0.0010 ***	0.0004 ***
Diabetes	-0.0155 ***	0.0050 ***	0.0096 ***	0.0009 ***
Hypertension	-0.0092 ***	0.0070 ***	0.0028 **	-0.0005
Cancer	-0.0102 ***	0.0073 **	0.0032 *	-0.0003
Heart diseases	-0.0190 ***	0.0116 ***	0.0078 ***	-0.0005
Stroke	-0.0283 ***	0.0120 ***	0.0149 ***	0.0013 ***
Lung diseases	-0.0213 ***	0.0116 ***	0.0104 ***	-0.0008
Dementias	-0.0381 ***	0.0160 ***	0.0170 ***	0.0051 ***
Smoker	-0.0177 ***	0.0117 ***	0.0046 ***	0.0014 ***
Former smoker	-0.0047 **	0.0044 **	0.0000	0.0004
Class I obesity	-0.0034	0.0024	0.0011	-0.0001
Classes II-III obesity	-0.0071 *	0.0048	0.0043	-0.0020
Woman	-0.0192 ***	0.0134 ***	0.0052 ***	0.0005 **
Immigrant	-0.0013	-0.0018	0.0044 ***	-0.0013 **
Secondary diploma	0.0096 ***	-0.0046 ***	-0.0042 ***	-0.0008 **
College diploma	0.0118 ***	-0.0048 **	-0.0061 ***	-0.0008 **
University diploma	0.0241 ***	-0.0098 ***	-0.0119 ***	-0.0023 ***
Disability: 1 ADL or cogn. impair.	-0.1412 ***	0.0975 ***	0.0417 ***	0.0019 ***
Disability: 2+ ADL	-0.1698 ***	0.0663 ***	0.0987 ***	0.0047 ***
Disability: in institution	-0.2250 ***	0.0688 **	0.1270 ***	0.0289 ***
Resides in Quebec	0.0057 **	-0.0052 ***	-0.0005	0.0000
Number of observations	108,355			
Average incidence	91.20%	4.26%	4.06%	0.34%

Table 4.5: Average marginal effects on probabilities of transition towards the states of disability

4.2.5 Model for obesity

We model the transitions between the different possible states of obesity (BMI<30, 30≤BMI<35, and BMI≥35) in the same manner as disability. However, the model does not include the presence of illnesses and states of disability as explanatory factors, since obesity is considered a risk factor that affects these variables and not a consequence of these. Table 4.6 presents the average marginal effects.

We find that an additional year before the age of 50 increases the probability of beginning to suffer from class I obesity by 0.06%, but does not affect the probability of starting to suffer from classes II-III obesity. After the age of 50, age decreases the probability of being obese, both for class I and for classes II-III obesity. Being a smoker significantly increases the probability of becoming classes II-III obese, while being a former smoker has a positive and significant effect on the incidence of the two types of obesity. This last relation is often observed because of weight gain occurring after cessation of tobacco use.

The effects of the obesity dummy variables are of course very significant, since individuals tend to remain in their present state in the following period. Women have a slightly lower tendency than men to become obese, much like proportionally fewer immigrants become obese than non-immigrants. A higher level of education generally has a negative effect on the probability of becoming obese. Finally, residing in Quebec is associated with a slightly lower probability of becoming obese.

	BMI < 30	30 ≤ BMI < 35	BMI ≥ 35
Age (if 50 years or under)	-0.0006 ***	0.0006 ***	0.0000
Age (if over 50 years)	0.0009 ***	-0.0005 ***	-0.0005 ***
Smoker	-0.0001	-0.0005	0.0005 ***
Former smoker	-0.0078 ***	0.0047 *	0.0032 ***
Class I obesity	-0.2324 ***	0.2012 ***	0.0311 ***
Classes II-III obesity	-0.2715 ***	0.1779 ***	0.0936
Woman	0.0045 **	-0.0108 ***	0.0063
Immigrant	0.0058 **	-0.0061 **	0.0003
Secondary diploma	0.0040 *	-0.0031	-0.0009 **
College diploma	0.0068 **	-0.0060 *	-0.0008 ***
University diploma	0.0134 ***	-0.0097 ***	-0.0037
Resides in Quebec	0.0086 ***	-0.0053 **	-0.0033
Number of observations		107,566	
Average incidence	83.3%	11.93%	3.82%

Table 4.6: Average marginal effects on probabilities of transition towards the states of obesity

4.3 Conclusions and future improvements to the model

The main conclusions of this chapter concern the risk factors and differences between Quebec and Canada.

We note that the risk factors (tobacco use and obesity) have an important impact on transitions towards different illnesses. For example, smokers and former smokers have a greater chance of developing hypertension, cancer, a heart disease, a stroke or dementia than persons who have never smoked. Similarly, the risk of developing diabetes, hypertension, a heart disease or a lung disease is higher among persons who suffer from obesity than those who do not.

Moreover, tobacco use and obesity are associated with a greater probability of developing at least one incapacity. Persons with a BMI over 30 have a lower chance of starting to smoke than those with a BMI under 30, while the risk of dying between two cycles of the NPHS is higher for smokers and former smokers than for those who have never smoked.

The models of transition towards illnesses and towards death indicate that there are no significant differences between Quebec and Canada in terms of transition probabilities. Analogously, the dummy variable for Quebec is not statistically significant when we look at the transitions between different states of tobacco use. However, there are some differences between Quebec and Canada in the models of transition towards disability and obesity. Individuals who live in Quebec have a lower probability of developing an incapacity and becoming class I obese than individuals in the rest of Canada. Residing in Quebec decreases by 0.52% the annual probability of developing an incapacity or a cognitive impairment. The probabilities of transition towards classes II-III obesity, towards institutionalization, and towards two incapacities or more do not differ between Quebec and Canada.

The results of the transition model towards states of obesity indicate that some improvements remain to be made with respect to persistence of obesity in our model. Indeed, the probability of suffering from class I obesity in the following period only increases by 20% for an individual who suffers from class I obesity in the current period. The probability of suffering from classes II-III obesity in the following period does not depend statistically on classes II-III obesity in the current period. It therefore seems that the simulated individuals change BMI categories relatively easily. But certain studies (for example Daouli et al. (2014)) indicate instead that it is difficult for an individual who suffers from obesity to change BMI category. Estimation of the dynamics of obesity in the model should thus be revised in order to more adequately represent the presence of obesity over the life cycle of an individual. Due to the lack of persistence, it is possible that the prevalence of obesity in COMPAS is underestimated.

The modeling of the entering cohorts also requires further work. At this stage of developing the model, one cohort can be distinguished from another according to the characteristics at 30 years old. For example, the prevalence of obesity at 30 years of age differs between cohorts. However, the probabilities of transitions towards different states are the same for each cohort. Including cohort effects in the incidence of different illnesses and risk factors would allow us to account for changes between different cohorts in terms of the probability of changing health states. For example, this would imply that the probability of becoming classes II-III obese could vary between cohorts. NPHS data would allow us to calculate average incidence by cohort between 1994 and 2010. Thus, if the incidence of an illness proves lower in younger cohorts, we could reduce the probability of the incidence of illnesses in cohorts entering the model. Such changes could result from medical advances or behavioural changes between cohorts.

5 Renewal

5.1 Overview

In COMPAS, entry of a new cohort of individuals aged 30 and 31 years is done in each simulation cycle, that is every 2 years by default. Each entering individual has many characteristics. In proportion, these should at all times reflect the joint distribution of the initial conditions (at $t = 0$) of the target population aged 30 and 31 years, as well as potential trend changes that we will have applied. In this chapter, we start by describing our methodology in section 5.2, then in section 5.3 we present some results of our renewal model.

5.2 Modeling

In each simulation cycle, we introduce individuals of a fixed age into COMPAS. Denote by $y = (y_1, \ldots, y_M)$ their M individual characteristics. Table 5.1 gives the list of characteristics of an individual entering the model. Each of these characteristics corresponds with only one type of variable: binary, ordinal, or integer.

Each year, the average characteristics of thecohort entering the model change. These characteristics are correlated with each other. Let $F_t(y)$ be the joint probaility distribution of these characteristics in year t and let $F_m(y_m)$ be the marginal distribution of the variable y_m. We should underline that several of these characteristics have a degenerate distribution, i.e., they only take one value with a probability of 1, for example 0. [7]

[7] X is a degenerate random variable if and only if X is almost certainly equal to a constant.

Variables	Types
Age	integer
Birth year	integer
Sex	binary
Immigrant	binary
Education level	ordinal
Diabetes	binary
Hypertension	binary
Heart diseases	binary
Stroke	binary
Cancer	binary
Lung diseases	binary
Dementias	binary
Tobacco use	ordinal
Obesity	ordinal
Disability	ordinal
Consultations with a generalist	integer
Consultations with a specialist	integer
Hospitalization nights	integer
Consumption of at least one medication	binary
Home care services	binary

Table 5.1: Types of variables associated with entering individuals

We wish to model the evolution of entering cohorts over time. To do this, using information from the public use microdata files discussed in chapter 2, we look at the evolution of $\{F_{m,t}(y_m),\ m = 1, \ldots, M\}$. We arrive at this by modeling the evolution of the expected value of y_m. This modeling depends on the type of the variable representing an individual characteristic. In the present section, we only discuss certain binary or ordinal type variables. Thus, we do not address either the "consultations with a doctor" variables or the variable that captures the number of nights in hospital. Also, there is no reason to model age or birth year since all individuals or 30 or 31 years old when entering the model. Thus, we only account for correlations between binary and ordinal variables of table 5.1, and these are the only ones maintained among entering cohorts.

Let $\{\tau_{m,t},\ m = 1, \ldots, M\}$ be a sequence such that:

$$E_{m,t}[y_m] = \tau_{m,t} E_{m,0}[y_m], \tag{10}$$

where $E_{m,0}[y_m]$ is the average of y_m in the year of initialization of the model ($t = 0$), and $E_{m,t}[y_m]$ is the expected value of y_m at t. In COMPAS, the year of initialization is the year 2010. Now let us look first at binary variables. These variables can be written as

$$y_m = \mathbb{1}_{(\beta_m + \epsilon_m > 0)}, \tag{11}$$

where $\mathbb{1}$ is an indicator variable, β_m is a threshold and ϵ_m is a normally distributed error term. Thus, we have

$$E[y_m] = \Phi(\beta_m), \tag{12}$$

where $\Phi(.)$ is the cumulative distribution function of a normal distribution.

As for the ordinal variables with K categories, we have $y_m = k$ if

$$\beta_{m,k} > \epsilon_m > \beta_{m,k-1}, \tag{13}$$

where $\beta_0 = -\infty$ and $\beta_K = \infty$. Thus, if $y_{m,k} = \mathbb{1}_{(y_m=k)}$, it follows that:

$$E[y_{m,k}] = \Phi(\beta_{m,k}) - \Phi(\beta_{m,k-1}). \tag{14}$$

We assume in this case that the vector of unobserved residuals is $\epsilon = (\epsilon_1, \ldots, \epsilon_M) \sim \mathcal{N}(0, \Omega)$.

For binary characteristics, for each year we adjust the parameters such that

$$\begin{cases} \tau_{m,t}E[y_m] & = & \Phi(\beta^*_m) \\ \beta^*_m & = & \Phi^{-1}(\tau_{m,t}E[y_m]) \end{cases} \tag{15}$$

where β^*_m is a threshold and $\tau_{m,t}$ is a factor of change such that $\tau_{m,0} = 1$. The factors of change $\tau_{m,t}$ are based on observed historical trends in the Canadian population. As explained elsewhere (Boisclair et al., 2014), we use the Labour Force Survey (LFS) and the CCHS to calculate these historical trends, for instance with respect to the level of education. These are then used to calculate the factors of change $\tau_{m,t}$.

Similarly, for ordinal variable, we have for $k = 1$

$$\begin{cases} \tau_{m,1,t}E[y_{m,1}] & = & \Phi(\beta^*_{m,1}) \\ \beta^*_{m,1} & = & \Phi^{-1}(\tau_{m,1,t}E[y_{m,1}]) \end{cases} \tag{16}$$

where $\beta^*_{m,1}$ is a threshold and $\tau_{m,1,t}$ is a factor of change such that $\tau_{m,1,0} = 1$. Recursively, we can then adjust the other parameters using

$$\tau_{m,k,t}E[y_{m,k}] \quad = \quad \Phi(\beta^*_{m,k}) - \Phi(\beta^*_{m,k-1}). \tag{17}$$

Thus, we have

$$\beta^*_{m,k} \quad = \quad \Phi^{-1}(\tau_{m,k,t}E[y_{m,k}] + \Phi(\beta^*_{m,k-1})). \tag{18}$$

With the new parameters, we can then draw a new cohort using a random draw from the multivariate normal distribution. Indeed, if $\tilde{\eta}$ is a drawing vector independent from the standard Gaussian random variable, then a draw of $\tilde{\epsilon}$ is given by $L_\Omega\tilde{\eta}$ where L_Ω is the triangular matrix resulting from the Cholesky factoring of matrix Ω.

In summary, if we know the factors of change $\{\tau_{m,t},\ m = 1, \ldots, M\}$ estimated using official data, then a binary type characteristic for an individual entering in period t is given by:

$$\tilde{y}_m = \mathbb{1}_{\left(\beta^*_m + \sum_{j=1}^{m} L_\Omega(m,j)\eta > 0\right)}. \tag{19}$$

Similarly, if we know $\{\tau_{m,k,t},\ m = 1,\ldots,M\}$, then the ordinal type characteristic with $k = 1,\ldots,K$ categories is given by:

$$\tilde{y}_m = k, \quad \text{si } \beta^*_{m,k} < \sum_{j=1}^{m} L_\Omega(m,j)\eta < \beta^*_{m,k+1}. \tag{20}$$

The size and individual characteristics of new cohorts may then be adjusted using projections performed using official data — here, data from Statistics Canada (see section 5.3).

5.3 Implementation

In COMPAS, we mainly use the public use microdata published by Statistics Canada. While the NPHS is mostly used to calculate transitions of individuals already in the model, we instead use the CCHS and the LFS to construct, based on historical trends, factors of change to be applied to future trends. Only the four exogenous characteristics that are immigration, education level, tobacco use and obesity are the object of such adjustments. For more on this, we refer to Public Health Agency of Canada and Canadian Institute for Health Information (2011).

5.3.1 Historical trends

In order to identify the historical trends concerning obesity, education and tobacco use, we use the public use microdata files of the CCHS and LFS from 2000 and 2012. The sample considered consists of the whole Canadian population aged 25 to 34 years. This means we implicitly assume that the evolution of these three characteristics is the same in Quebec as in the rest of Canada, although their levels may differ.We retain individuals aged 25 to 34 years in order to have sufficient observations while remaining as close as possible to the age of entering cohorts, which are 30 and 31 years old. Only average trends are considered, i.e., we assume that the trends are the same for men and women.

According to Statistics Canada (2000a, 2012a, see table 5.2), between 2000 and 2012 the obesity rate increased for the population aged 25 to 34 years. The annual average growth rate (AAGR) was 1.09%. This growth is mostly due to an important increase in classes II-III obesity — of just over 2% per year — given that class I obesity only increased by 0.12% per year.

Statistics Canada (2000b, 2012b, see table 5.2) indicates an important decrease in the share of Canadian aged 25 to 34 years who do not have any diploma. The annual decrease was 3.41%. Conversely, it is not surprising to find that the share of individuals with a university diploma rose, by 2.10% per year.

While the trend is upward for obesity, the inverse is observed for tobacco use (Statistics Canada, 2000a, 2012a, see table 5.2). The share of smokers declined by about 2.6% per year. While the proportion of former smokers is relatively stable, the share of Canadians who never smoked increased during the period under study, at an AAGR of 0.93%.

Table 5.2 presents a summary of the AAGRs calculated using the survey data.

Data sources	Characteristics	AAGR
	OClass I obesity	0.12%
ESCC	Classes II-III obesity	2.11%
	Total obesity	1.09%
	No diploma	-3.41%
EPA	Secondary diploma	-1.46%
	College diploma	0.38%
	University diploma	2.10%
	Current smoker	-2.57%
ESCC	Former smoker	0.10%
	Never smoked	0.93%

Table 5.2: Annual average growth rate (AAGR) of the shares of certain individual characteristics, Canadian population aged 25 to 34 years, 2000 to 2012

5.3.2 Projections

The historic AAGRs obtained for obesity, education and tobacco use are not used as such in each simulation cycle. The different the growth rates are adjusted in order to, on the one hand, reflect some uncertainty about the future and, on the other, account for certain particularities about the Quebec population.

Between 1990 and 2004, the AAGR of the proportion of 25 to 34 year-olds who were obese was 8.34% in Quebec (Audet, 2007). This rate is higher than the AAGR found between 2000 and 2012 when considering only those Quebecers aged 25 to 34 years (Statistics Canada, 2000a, 2012a). It seems, then, that there has been some slowdown in the obesity growth rate in recent years. In our opinion, this suggests that future cohorts will most likely have only slightly higher obesity rates compared to their predecessors (Institut national de santé publique du Québec, 2012).

A similar effect is observed for education. The Quebec observations of the LFS between 1990 à 2000 show higher absolute values of AAGRs in all categories between 2000 and 2012. Thus, it seems that there is some slowdown in the variation in the share of individuals with a given level of education.

Thus, for the obesity and education variables, the 2000-2012 AAGR presented in table 5.2 is only assumed to hold between 2010 and 2020. It is then reduced by half in 2020, 2030 and then again in 2040. The AAGR for 2040 is then held constant between 2040 and 2050.

The trends relating to tobacco use in Quebec are relatively stable as of 2009. The prevalence of tobacco use is stable in the general population, at around $17-20\%$ (Health Canada, 2012). This seems to indicate that the decline in prevalence recorded bewteen 2000 and 2012 is mainly the result of an important decline in the number of smokers in the early 2000s. There is therefore no reason to believe that the share of smokers will decline in each future cohort between 2010 and 2050. Due to this, each new cohort entering in the model has, at 30 and 31 years old, approximately the same share of smokers as the preceding cohort. Otherwise stated, the trends are stable for tobacco use in the model : the

AAGR is zero. This is precisely what is shown in table 5.4 where the proportions are shown for entering cohorts.

As opposed to education, obesity and tobacco use, the treatment of immigration in the model is not based on historical trends that are adjusted and then projected into the future. We instead use the forecasts of the Régie des rentes du Québec (2013). The RRQ forecasts that the net migratory balance, which includes both migration between provinces and international immigration/emigration, will decline slightly, from 38,600 in 2012 to 34,800 in 2015. It then remains at its 2015 level until 2062. While the addition of migratory waves ends in 2050 in COMPAS, these assumptions are used "as is" in the microsimulation.

In table 5.3, we have coded the characteristics included in table 5.2. In table 5.4, we present the distribution of certain individual characteristics in the entering cohorts. These are adjusted using equations 15 and 16.

6 Demographics

The present chapter describes the assumptions with regard to mortality and immigration as well as the demographic models which serve as the basis for COMPAS. It describes the modelling of mortality improvements.

6.1 Mortality

6.1.1 Definitions and statistics

The (gross) mortality rate is the ratio of the number of deaths in a year to the average total population in the year. Otherwise stated, it is the probability of dying in the course of a year. Among other things, this varies depending on the age structure of the population, sex and year. According to ISQ statistics (Payeur, 2013), the mortality rates remained stable at around 7.4 per thousand between 2001 and 2012, and sat at 7.5 per thousand in 2012.

While this probability is of interest, to analyze the evolution of mortality with precision, other statistical indicators such as life expectancy at birth are preferred (Payeur, 2013). The United Nations defines life expectancy at birth as the number of years that a newborn should live if general mortality conditions prevailing at the moment of birth were to remain constant throughout his/her life (United Nations, 2008): *it characterizes mortality independently from the age structure of the population.* In 2010, life expectancy at birth in Quebec was 79.4 years for men and 83.5 years for women according to statistics of the Régie des rentes du Québec (2013). According to provisional ISQ data, in 2012 life expectancy at birth sat at 79.8 years for men and 83.8 years for women (Payeur, 2013).

In the historical data, we observe a clear improvement in mortality between 1962 and 2010 (Statistics Canada, 2013, tab. 102-0512), (Régie des rentes du Québec, 2013, tables 14 and 15). In COMPAS, given that young cohorts aged 30 and 31 years enter the modeling process up until 2050, the very efficiency of our model requires forecasts of mortality rates over a long period. Moreover, given that an individual

Variables	Codes	Characteristics
	0	Non immigrant
Immigration	1	Immigrant
	1	No diploma
	2	Secondary diploma
Education level	3	College diploma
	4	University diploma
	1	Never smoked
Tobacco use	2	Current smoker
	3	Former smoker
	1	Absence of obesity
Obesity	2	Classe I obesity
	3	Classes II-III obesity

Table 5.3: Coding of certain variables characterizing an individiual

Cycle t	Year	Immigration		Level of education				Tobacco use			Obesity		
		0	1	1	2	3	4	1	2	3	1	2	3
0	2010	78.53	21.47	4.07	12.39	55.46	28.09	39.62	28.09	32.29	84.30	13.06	2.64
1	2012	78.58	21.42	4.47	12.03	53.94	29.56	41.06	27.86	31.08	84.93	12.43	2.64
5	2020	81.17	18.83	3.31	11.05	54.74	30.90	37.97	29.83	32.20	83.05	12.75	4.20
10	2030	81.48	18.52	2.55	9.57	56.35	31.53	40.38	26.79	32.83	81.13	13.51	5.37
15	2040	82.96	17.04	2.15	8.99	61.94	26.92	39.58	28.62	31.80	80.14	13.28	6.57
20	2050	80.90	19.10	2.37	9.79	58.99	28.85	40.61	27.33	32.07	80.46	14.18	5.37

Table 5.4: Distribution of certain individual characteristics in entering cohorts (in %), in simulation cycles $t \in \{0, 1, 5, 10, 15, 20\}$

may reach 110 years old in COMPAS, the forecasts should ideally include age groups for the most elderly of the elderly, or those aged 90 and over.

We deem it important to underscore that this modeling procedure gives rise to an endogeneity problem. Improved overall mortality depends mainly on two factors: *(i)* trends linked to diseases and risk factors and *(ii)* technological progress, particularly in medicine. The contribution of diseases and risk factors is largely modeled in COMPAS, and it is likely that its overall combined effect is nearly nil as a result of the various opposing effects. Thus, in the long term, the significant improvement in mortality mostly comes from technological progress.

In the following section, we present some methods that enable us to make mortality rate forecasts.

6.1.2 Estimation methods

In order to quantify, for instance, the impact of increasing life spans on the cost of pensions, prospective mortality tables have been built around the world (Macdonald, 1997; McDonald et al., 1998, for a description see). Numerous models exist, but the development of existing forecasts is largely based on the *Lee-Carter* model introduced in Lee and Carter (1992), as well as on certain variants, such as the Poisson log-bilinear model (Brouhns et al., 2002; Renshaw et al., 1996).

The Lee-Carter model consists of first decomposing mortality into two components, age and calendar time, and then extrapolating into the future observed past trends. In its log-bilinear form, this model is written (Renshaw and Haberman, 2003c) as:

$$\ln \mu_x(t) = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t}, \tag{21}$$

where:

- $\mu_x(.)$ is the instantaneous mortality rate, assumed constant "by parts" at date t for age x;
- α_x represents the component specific to age x: it is the average behaviour of the logarithms of $\mu_x(.)$ over time;
- κ_t describes the general evolution of mortality rate $\mu_x(.)$ over time ;
- β_x indicates the sensitivity of instantaneous mortality with respect to the general evolution of mortality: this is the gap between the $\ln(\mu_x(.))$ and the average behaviour α_x;
- $\epsilon_{x,t}$ is an error term which, by the null hypothesis, reflects the specifics of age x or date t which are not captured by the model. The errors $\epsilon_{x,t}$ are assumed i.i.d. $\sim \mathcal{N}(0, \sigma^2)$.

In reality, the logarithm of the observed instantaneous mortality rate is generally more volatile at very old ages due to the small number of observations: i.e., the absolute number of deaths among the most elderly persons is small simply because there are fewer of them. As a result, the assumption of homoscedasticity remains difficult to validate. In order to fix this problem and improve the quality of the estimation, Brouhns et al. (2002) propose to model the number of deaths observed at age x in year t, which we denote as $D_{x,t}$, using a Poisson distribution of the parameter $E_{x,t}\mu_x(t)$:

$$D_{x,t} \sim \mathcal{P}oisson\left(E_{x,t}\mu_x(t)\right), \tag{22}$$

where $E_{x,t}$ is the average number of persons of age x in year t (this is the exposure to risk) while $\mu_x(.)$ is the instantaneous rate defined by:

$$\mu_x(t) = \exp\left(\alpha_x + \beta_x \kappa_t\right). \tag{23}$$

Compared to the Lee-Carter model, the Brouhns et al. (2002) model not only enables us to lift the assumption of homoscedasticity, but also accounts for the integer nature of the number of deaths $D_{x,t}$ and better captures the greater variance of the instantaneous mortality rate at higher ages that is due to the small number of observations. Ramifications of the models defined by equations (21) to (23) were proposed in Renshaw and Haberman (2003a) and Renshaw and Haberman (2003b,c), respectively.

Below, we present the integration of a prospective mortality table into the COMPAS model. We mainly use the mortality rates forecasted by the Régie des rentes du Québec (2013, tab. 14), an excerpt of which appears in tables 6.1 and 6.2. This choice, which is not irreversible, is nevertheless consistent with the selected method: the RRQ uses an extension of the Lee-Carter model, a model that has become standard despite some critiques (see for example Gutterman and Vanderhoof, 1998).

6.1.3 Integration into COMPAS

Let $s \in S = \{h, f\}$ with S being a set representing human genus where the letters h and f respectively represent men ("hommes" in French) and women ("femmes"). Let $\{ga_i,\ i = 1,\ldots,n\}$ be the age groups observed in the population and $\{pp_j,\ j = 1,\ldots,m\}$ be the projection periods. The corresponding samples ranked in increasing order are denoted as $ga_{1,n} < \ldots < ga_{n,n}$ et $pp_{1,m} < \ldots < pp_{m,m}$. The prospective mortality table by sex, age group (increasing) and period (also increasing) is denoted by $M_{s,n,m}$. Let g be the number of years between two demographic transitions as described in chapter 4. Let us denote by $nages$ and $nyears$ the number of age ranges and simulation cycles in the COMPAS model:

$$nyears = 1 + (stopyear - startyear)/g, \tag{24}$$
$$nages = stopage - startage + 1, \tag{25}$$

where $startyear$, $stopyear$, $startage$ and $stopage$ are respectively the year of initialization of the model, the last simulation year, the minimum age an individual can have, and the maximum age at death permitted for an individual in the model. It is important to specify that the notions of ' age group in the population" and ' age range " in the model do not necessarily refer to the same mathematical quantities.

Under the assumption that there is a gradual displacement of mortality rates reduction from younger ages towards older ages (see below), we define the annual decrease in the probability of death, $q_{\mathrm{x}}(s, a, .)$, in the following recursive manner:

$$q_{\mathrm{x}}(s, a, 1) = 1, \tag{26}$$
$$q_{\mathrm{x}}(s, a, t) = q_{\mathrm{x}}(s, a, t-1)\,(1 - M_{s,i,j})^{g}\,, \tag{27}$$

with $a = 1, \ldots, nages$ and $t = 2, \ldots, nyears$.

On the one hand, note that the minimum age an individual can have in COMPAS, $startage$, the current age range a, and the real age of an individual are linked by the relation:

$$age = startage + a - 1. \tag{28}$$

Thus, the matrix index i in (27) is linked to the real age of an individual and the population age group in the prospective mortality table $M_{s,n,m}$ through the relationship:

$$i = \sum_{k=1}^{n} k \times \mathbb{1}_{\left(age \in ga_{k,n}\right)} = \begin{cases} 1 & \text{si } age \in ga_{1,n} \\ 2 & \text{si } age \in ga_{2,n} \\ \vdots & \\ n & \text{si } age \in ga_{n,n} \end{cases}$$

On the other hand, the year of initialization of the model *startyear*, the number of years between two demographic transitions g, the current cycle t and the current year *year* in the simulation are linked by:

$$year = startyear + g(t-1). \tag{29}$$

Thus, the matrix index j in (27) is linked to the current year *year* and the projection period in the prospective mortality table $M_{s,n,m}$ by the relation:

$$j = \sum_{k=1}^{m} k \times \mathbb{1}_{\left(year \in pp_{k,m}\right)} = \begin{cases} 1 & \text{si } year \in pp_{1,m} \\ 2 & \text{si } year \in pp_{2,m} \\ \vdots & \\ m & \text{si } year \in pp_{m,m} \end{cases}$$

We then define the probability that an individual remains alive, accounting for his/her individual characteristics and after a demographic transition, as

$$p := p_{s,a,t} = 1 - (1 - p_{i',j'})q_{\mathrm{x}}(s,a,t), \tag{30}$$

with $p_{i',j'} = 1 - \exp(-\exp(\epsilon_{i',j',t}))$ as the distribution function of the Gompertz law where $\epsilon_{i',j',t}$ is a random term specific to individual i' and to illness j', including his/her disability status, at time t; for details, the reader can refer to sections 4.2.1 and 4.2.2. This probability thus depends on sex, projection period, age range and individual health status.

Starting from the observation that mortality among younger individuals has clearly improved, reaching a very low level which can hardly keep declining at the same rate as in the past, the RRQ uses a modified Lee-Carter model in order to account for the gradual displacement of reduced mortality rates from younger ages to older ages. Since our model is mainly based on RRQ forecasts of mortality rate reductions, we have introduced the factor $(1 - M_{s,i,j})^g$ in equation (27) in order to account for this observation. This factor will tend to reduce the rate of decline of the probability of death of younger individuals compared to the elderly.

The extension to the model used by the RRQ does not incorporate opinions on the potential evolution of mortality due to advances in medicine, to the appearance of new illnesses or risk factors or to changing lifestyles. in this respect, it is worth noting that the RRQ implicitly recognizes this through the following: *‘ the repercussions of new technologies, genetics research and new medicines on life expectancy are difficult to evaluate. Until now, lifestyle and the environment seem to be stronger determining factors of individuals’ mortality"*. Now, in COMPAS we model lifestyle-related improvements, but not those linked with technological progress. This last aspect is accounted for by introducing an ‘ exogenous" improvement to mortality, as described elsewhere (Boisclair et al., 2014).

We have slightly modified tables 6.1 and 6.2 in order to account for two specific aspects of COMPAS: *(i)* the simulation ends in 2130

Age group	Observed			Forecast				
	1980-1990	1990-2000	2000-2010	2010-2020	2020-2030	2030-2040	2040-2050	2050-2060
	%	%	%	%	%	%	%	%
0 to 9 years	4.8	6.4	2.5	2.2	1.0	0.5	0.5	0.5
10 to 19 years	2.4	4.4	5.1	1.7	0.7	0.5	0.5	0.5
20 to 29 years	2.0	2.5	3.4	2.3	0.8	0.5	0.5	0.5
30 to 39 years	0.1	3.4	3.1	1.6	0.6	0.5	0.5	0.5
40 to 49 years	2.5	2.4	2.6	2.2	1.2	0.7	0.6	0.5
50 to 59 years	3.2	2.7	2.5	2.8	2.5	1.9	1.4	1.3
60 to 69 years	2.1	2.9	3.1	2.5	2.4	2.1	1.9	1.6
70 to 79 years	1.2	2.1	3.4	1.9	1.8	1.6	1.3	1.1
80 to 89 years	0.6	1.2	2.5	1.0	0.8	0.7	0.6	0.5

Table 6.1: Annual rates of reduction of mortality rates among men (source : Régie des rentes du Québec (2013, tab. 14))

Age group	Observed			Forecast				
	1980-1990	1990-2000	2000-2010	2010-2020	2020-2030	2030-2040	2040-2050	2050-2060
	%	%	%	%	%	%	%	%
0 to 9 years	4.0	7.6	1.0	2.5	1.6	0.8	0.6	0.5
10 to 19 years	4.2	1.7	1.9	2.1	1.5	0.9	0.6	0.5
20 to 29 years	3.4	1.5	1.1	0.9	0.6	0.5	0.5	0.5
30 to 39 years	1.0	1.8	2.4	1.3	1.1	0.8	0.6	0.5
40 to 49 years	2.4	1.1	1.8	1.3	1.2	1.0	0.8	0.7
50 to 59 years	2.7	1.1	1.1	1.4	1.4	1.4	1.4	1.3
60 to 69 years	2.0	1.5	1.5	1.6	1.6	1.5	1.4	1.3
70 to 79 years	1.9	1.6	1.7	1.5	1.5	1.3	1.1	0.9
80 to 89 years	1.2	1.3	1.6	1.1	1.0	0.8	0.6	0.5

Table 6.2: Annual rates of reduction of mortality rates among women (source: Régie des rentes du Québec (2013, tab. 14))

6.2 Immigration

In order to better reflect demographic reality, we should account for individuals moving from one country to another — *international migration* — or from one province to another — *interprovincial migration*.

As indicated in chapter 5, in COMPAS we simply refer to data on net migration from the Régie des rentes du Québec (2013). This is the net migratory balance, i.e., the difference between immigration and emigration over the course of a year. These numbers, which include migration of non-permanent residents, shows that net migration represented about 0.48% of the population between 2001 and 2014. In 2015, forecasts indicate a gradual decline in the balance to 0.42%, which would remain stable up until 2061. In 2062, the net migratory balance will only amount to 0.36% of the population.

Note that the above forecasts assume that (Régie des rentes du Québec, 2013):

- Quebec and Canada's policies on immigration, in particular the December 2012 international immigration targets of the Ministry of Immigration and Cultural Communities, remain stable;
- the interprovincial migratory balance remains stable and corresponds with the average level observed over the 15 years preceding 2012.

6.3 Conclusion

As shown by figure **??**, by integrating the mortality model and the migratory balance forecasts in COMPAS and combining the results with the RRQ's forecasts for those under 30 years old[8], we get a total population of just over 9 million inhabitants in 2050.

7 Health care use

7.1 Overview

The health care use module makes it possible to evaluate, for each simulation year, the quantity of medical resources used by the population. In order to obtain these results, however, we must establish a relationship between health status and resource use.

The models used to do this must account for a particularity of data on use of medical resources, namely, that they are generally characterized by many missing values and an asymmetric distribution. Stated otherwise, the foregoing means that during a reference period, many individuals do not actually use any medical resources while others use much more than the average (Frees et al., 2011). This observation applies to both the number of consultations with a doctor and the number of nights spent in a hospital.

In order to account for such a distribution of data, we use a negative binomial regression. Two of the health care variables considered in COMPAS, namely consumption of at least one medication and use

[8]This operation is required here because COMPAS only simulates the population aged 30 and over.

(or not) of home care services, can take only two values (null or positive). To model them, a logistic regression model is used instead. The econometric theory of the models presented is based on Cameron and Trivedi (2005).

7.2 Econometric models

7.2.1 Negative binomial regression

The negative binomial regression is generally used to analyze discrete and countable data, i.e. data that only takes whole and non-negative values, which corresponds well to health care use data. This type of regression, by allow for overdispersion of data (variance higher than the mean), can capture the asymmetry in the distribution of health care use data. Using the Poisson distribution, which assumes equality between mean and variance, would not have matched the distribution of the data as effectively.

The negative binomial regression assumes that the observations are generated from a negative binomial distribution, the first two moments of which are:

$$\mathbb{E}(y_i|\mu_i, \alpha) = \mu_i = \exp(x_i'\beta)$$
$$\mathbb{V}ar(y_i|\mu_i, \alpha) = \mu_i(1 + \alpha\mu_i)$$

where:

- $\mathbf{x}_i$ is a vector of explanatory variables for individual i;
- β is a vector of coefficients;
- $\boldsymbol{\alpha}$ is the overdispersion parameter.

If $\alpha > 0$, it necessarily follows from the definition of variance that there is overdispersion, given that the variance is greater than the mean (μ_i being positive).This definition of variance is the one used by the *nbreg* command in Stata to perform the negative binominal regression. Estimation of coefficients is done by the maximum likelihood method.

However, the estimated parameters do not directly yield the effects of a variable x on the conditional expected value of y. Thus, it is not the parameters themselves which are interesting when analyzing the explanatory variables' impact, but rather the marginal effects, namely, the change in the conditional expected value of y_i when the value of a variable x_i changes by one unit. As mentioned in chapter 3, in the context of COMPAS, we use average of marginal effects (AME). As a reminder, the AME is an average change, over all individuals, in the expected value of y_i when x_i changes by one unit.

7.2.2 Logistic regression

The logistic regression, as mentioned above, is used with variables that can only take two values, such as consumption (or not) of at least one medication. The dependent variable y thus has a probability

p to take a value of 1 and of $1-p$ to take a value of 0. This probability is modelled as follows:

$$p \equiv \mathbb{P}\left(y_i = 1|x_i\right) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$$

where:

- $\mathbf{x}_i$ is a vector of explanatory variables for individual i;
- β is a vector of coefficients.

Estimation of the parameters is again done by maximum likelihood. As for the negative binominal regressions, the estimated parameters are not amenable to direct interpretation. It is the AME that is used to show the impact of the explanatory variables on the probability that y_i will take a value of 1.

The two types of regressions described in this section allow us to model the relation between use of medical resources and the presence of seven illnesses, two risk factors, and disability.

7.3 Results

This section presents the results of the different regression models discussed above and applied to the NPHS (see chapter 2 on this topic). While the regression type differs by the type of health care considered, the explanatory variables remain the same in all regressions. The regressions are only performed using respondents who are in private households in the current period, because there is not sufficient information on health care use for those who are in an institution. For example, if a respondent is in an institution at the time of the last survey cycle, his/her answers to the preceding cycles are still used (only responses in the last cycle are excluded). The socio-demographic variables used are sex, age, immigration status, education, and an indicator for respondents living in Quebec. This last variable allows us to capture a possible difference between Quebecois respondents and those of the rest of Canada.

The effect of age on resource use does not appear to be linear. Specifically, it appears to change after 50 years of age. In order to capture this effect, a spline was created (Goldman et al., 2005). There are therefore two age variables, one for individuals under the age of 50 and another for those aged 50 years and over. This last variable represents the number of years since the individual turned 50 (e.g., it takes a value of 5 for a 55-year old person). The estimated coefficient for the first age variable thus captures the effect of age up to 50 years on the dependent variable; similarly, the coefficient of the second variable is the effect of age after 50 years on use of health care resources.

The variables for disability and for the presence of the seven diseases studied are included in the regressions. The risk factors (tobacco use and obesity) are also included. Tobacco use is represented by two binary variables: one for former smokers and one for current smokers. Two binary variables are included in the regression to capture the effect of class I ($30 \leq$ BMI <35) and classes II-III (BMI

$\geq$ 35) obesity. The reference categories are thus individuals who never smoked and those with a BMI under 30.

The weighting used to make the sample representative of the Canadian population is the " longitudinal square weight" of the master file, that is the first weight computed by Statistics Canada (2012c). This weight is calculated at the first cycle of the NPHS and never changes. It can be used to make the sample representative of the Canadian population, even if in subsequent cycles some respondents no longer participate in the NPHS or no longer respond to all questions.

7.3.1 Negative binomial regression

The effect of the different illnesses on use of medical resources is presented in table 7.1. The use is calculated over a 12-month period. As such, if an individual consulted with a specialist physician 3 times, this indicates that he consulted with a specialist 3 times over the course of a 12-month period. As expected, illnesses have an important effect on the number of consultations with a generalist physician. The presence of diabetes, hypertension, cancer, heart diseases or lung diseases increases the number of consultations with a generalist by a multiple of 1.2, on average. Classes I and II-III obesity also increase the number of consultations, by 0.4 and 0.7 respectively. However, it is incapacities suffered by individuals that lead to the largest increase in consultations. An individual with at least 2 incapacities had 2.5 more consultations on average than an individual with no incapacity. On average, Quebecers consult physicians less often than individuals in the rest of Canada. The gap is about 1 consultation.

The effect of illnesses on the number of consultations with a specialist is also positive and significant. The presence of diabetes, hypertension, cancer, heart diseases or lung diseases increases the number of consultations. The presence of some type cancer is linked with nearly to additional consultations per year. The effect of dementias is not significant, but this could be due to the small number of individuals suffering from dementias in the sample used. Classes II-III obesity increases the number of specialist consultations by 0.1 visit per year on average, while tobacco use variables have little effect; former smokers have 0.1 more than individuals who have never smoked.

The presence of dementia does not have a significant impact on the number of nights in short-term hospitalization, but all other types of illness have considerable effects. The presence of a cancer or stroke increases the number of nights hospitalized by more than 2 on average. Individuals suffering from 2 or more incapacities spend an average of 4.2 more nights in hospital than those with no incapacity. The coefficient on class I obesity is not significant and is nearly zero, but individuals with a BMI of 35 or over (classes II-III) spend an average of 0.4 more nights in hospital than those of healthy weight.

7.3.2 Logistic regression

Table 7.2 presents the average marginal effects of each explanatory variable on the binary variables for resource use, namely, consumption of at least one medication and use (or not) of home care services.

All illnesses other than dementia have a positive and significant impact on the probability of consuming

Variable	Nb. of consultations: generalist	Nb. of consultations: specialist	Nb. of nights hospitalized
Woman	0.861***	0.272***	0.211*
	(0.05)	(0.02)	(0.10)
Age (if under 50 years)	-0.015***	-0.000	-0.008
	(0.00)	(0.00)	(0.01)
Age (if 50 years or over)	0.013***	-0.005***	0.027***
	(0.00)	(0.00)	(0.01)
Immigrant	0.320***	0.010	-0.216
	(0.07)	(0.03)	(0.12)
Secondary diploma	-0.256***	0.216***	-0.133
	(0.07)	(0.03)	(0.11)
College diploma	-0.255**	0.205***	-0.206
	(0.08)	(0.04)	(0.11)
University diploma	-0.380***	0.395***	-0.242
	(0.08)	(0.05)	(0.16)
Resides in Quebec	-0.931***	0.092**	0.636***
	(0.05)	(0.03)	(0.18)
Presence of diabetes	1.216***	0.179***	0.533***
	(0.11)	(0.04)	(0.15)
Presence of hypertension	1.228***	0.167***	0.297**
	(0.06)	(0.03)	(0.10)
Presence of cancer	1.353***	1.929***	2.851***
	(0.15)	(0.12)	(0.44)
Presence of heart diseases	1.306***	0.636***	1.639***
	(0.10)	(0.05)	(0.19)
Presence of stroke	0.667***	-0.079	2.059**
	(0.19)	(0.06)	(0.75)
Presence of lung diseases	1.307***	0.253***	0.370*
	(0.14)	(0.06)	(0.19)
Presence of dementia	0.057	0.118	0.163
	(0.21)	(0.12)	(0.27)
Class I obesity	0.364***	0.056	-0.037
	(0.07)	(0.03)	(0.11)
Classes II-III obesity	0.698***	0.112*	0.405*
	(0.12)	(0.05)	(0.19)
Current smoker	0.176*	0.001	0.294
	(0.07)	(0.03)	(0.18)
Former smoker	0.194***	0.090***	0.127
	(0.06)	(0.02)	(0.11)
Disability: 1 ADL or cognitive impairment	1.847***	0.713***	2.682***
	(0.12)	(0.07)	(0.41)
Disability: 2+ ADL	2.541***	0.941***	4.166***
	(0.14)	(0.08)	(0.38)
Number of observations	82,715	82,924	76,628
Average number of consultations or nights	3.11	0.75	0.86
Legend	* $p<0.05$; ** $p<0.01$; *** $p<0.001$		

Table 7.1: Negative binomial regression: average marginal effects of the different variables on health care use. Usage is calculated on an annual basis.

at least one medication. The largest effect is for hypertension, and individuals suffering from it have a 14.2% higher probability than non-sufferers of consuming at least one medication. The impact of

Variable	1+ medication	Home care services (yes)
Woman	0.099***	0.007***
	(0.00)	(0.00)
Age (if under 50 years)	-0.000	-0.000
	(0.00)	(0.00)
Age (if 50 years or over)	0.001**	0.002***
	(0.00)	(0.00)
Immigrant	-0.046***	-0.007**
	(0.01)	(0.00)
Secondary diploma	0.029***	-0.002
	(0.00)	(0.00)
College diploma	0.037***	-0.002
	(0.01)	(0.00)
University diploma	0.048***	-0.004
	(0.01)	(0.00)
Resides in Quebec	-0.023***	0.002
	(0.00)	(0.00)
Presence of diabetes	0.117***	0.009**
	(0.00)	(0.00)
Presence of hypertension	0.142***	-0.001
	(0.00)	(0.00)
Presence of cancer	0.062***	0.033***
	(0.01)	(0.00)
Presence of heart diseases	0.130***	0.012***
	(0.00)	(0.00)
Presence of stroke	0.095***	0.015***
	(0.01)	(0.00)
Presence of a lung diseases	0.081***	0.009**
	(0.01)	(0.00)
Presence of dementia	-0.044	0.002
	(0.03)	(0.00)
Class I obesity	0.027***	-0.001
	(0.00)	(0.00)
Classes II-III obesity	0.043***	0.010*
	(0.01)	(0.00)
Current smoker	0.007	0.001
	(0.00)	(0.00)
Former smoker	0.036***	0.001
	(0.00)	(0.00)
Disability: 1 ADL or cognitive impairment	0.064***	0.079***
	(0.01)	(0.00)
Disability: 2+ ADL	0.083***	0.178***
	(0.01)	(0.01)
Number of observations	82,199	83,524
Average proportion of users	84.58%	3.40 %
Legend	* p <0.05 ; **p <0.01; ***p <0.001	

Table 7.2: Logistic regression: average marginal effects of the different variables on health care use. Usage is calculated on an annual basis.

obesity is also significant. Classes II-III obesity increases the probability of consuming at least one medication by 4.3%, while the upward impact of class I obesity is 2.7%. All types of disability increase the probability of consuming at least one medication. Individuals with a secondary, college or university diploma have a higher probability of consuming at least one medication than those without a diploma. Quebec residents are 2.3% less likely to consume at least one medication. Former smokers have a 3.6% higher probability of consuming at least one medication than individuals who have never smoked, but being a current smoker has no impact.

The last health care use variable studied relates to home care services. The effects of health statuses on the probability of receiving home care services are generally quite low, at less than 1%. However, the probability of using home care services increases by an important amount for individuals with at least 2 disabilities. For them, the probability of using home care services increases by 17.8% compared to individuals with no incapacity. The presence of cancer increases the probability of receiving home care services by 3.3%, while the presence of a stroke increases the probability of receiving home care services by 1.5%. We also find that individuals who are classes II-III obese have a 1% higher chance of receiving home care services than those of healthy weight.

What we can take from these analyses is that illnesses considered in the model are important factors in explaining the use of health care resources. Disability also has sizable effects. Finally, respondents who live in Quebec can be distinguished from those in the rest of Canada in terms of consultations with a general practitioner and the number of nights spent in hospital. Quebecers have one more medical consultation per year on average and spend an average of 0.6 more nights in hospital per year than respondents from other provinces.

Conclusion

The creation of the COMPAS microsimulation model is a fairly recent affair: this began in 2013, and its development has been pursued since, benefitting among other things from the experience of a team member who participated in constructing the U.S. *Future Elderly Model*. To date, the team has laid the foundations of a structure which performs well and is flexible, making it possible to progressively refine the different elements already incorporated into the model and to incorporate new ones over the months and years to come.

COMPAS is mainly built in FORTRAN, a simple language which offers great flexibility and adaptability to the various needs inherent to modelling. The design of the model is modular, which makes it possible to modify the parameters and components separately and without necessarily – or immediately – affecting the entire model.

At present, COMPAS mainly uses 2010 CCHS data to buld the model's initial population, and the different cycles of the NPHS to compute the different transition models used to generate the dynamics of health care use. Data from other surveys, like the LFS, are also used to build different hypotheses and perform imputations.

The model is largely based on official demographic assumptions, in particular those of the RRQ, and uses a standard demographic model to generate the population dynamics. Some trend adjustments are also implemented by the team, in order to account for recent trends in the evolution of certain variables such as risk factors (tobacco use and obesity). These adjustments can themselves be easily modified at the user's discretion.

The behaviour of COMPAS appears adequate in its present state of development. While mortality appears overestimated relative to the RRQ's projections, for example, these gaps are not unexpected. They can be the object of research and improvements in the medium term, but various results of the reference scenario in terms of health (health status, life expectancies, health care use) seem reasonable and consistent with expectations (Boisclair et al., 2014).

The next major development of COMPAS, already in progress, consists of incorporating the costs of health care used by individuals. The subsequent development, already in progress as well, consists of linking COMPAS to the socioeconomic miscrosimulation model previously constructed by the team (see Clavet et al., 2012). This last will also feature improved modeling of savings and retirement income, which opens the door in the near future to refined analyses integrating the dimensions of health status, health costs and socioeconomic behaviours — including sociodemographic behaviours, work, consumption, savings and retirement — and the impact of these variables on the living conditions of families, demand for public services and public finances.

Finally, COMPAS mostly sets itself apart from existing dynamic microsimulation models in regards to:

- its emphasis on health sates and health services, including costs associated with the latter;
- its Quebec primary ambit (the model can eventually be extended to the whole of Canada);
- its development carried out within a university environment, but in collaboration with different non-university partners;

- its foreseen linkage with a socioeconomic microsimulation model developed for Quebec, which includes a detailed tax simulator that accounts for the specifics of the tax system and the main benefits programs in place in Quebec.

The team hopes that, in its current and future versions, this first Quebecois microsimulation model focused on health will prove a useful tool for both the scientific community and for policy analysis.

References

Audet, N. (2007). L'évolution de l'excès de poids chez les adultes québécois de 1990 à 2004 : mesures directes. Technical report, Institut de la statistique du Québec.

Bohnert, N. (2013). Mortality: Overview, 2008 and 2009. Technical report, Statistics Canada.

Boisclair, D., Côté-Sergent, A., Duclos, J.-Y., Lekina, A., Marchand, S., and Michaud, P.-C. (2014). Compas: Un modèle de microsimulation santé pour le Québec. Working paper 14-07 of the Chaire de recherche Industrielle Alliance sur les enjeux économiques des changements démographiques.

Brouhns, N., Denuit, M., and Vermunt, J. K. (2002). A poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, 31(3):373 — 393.

Cameron, A. and Trivedi, P. (2005). *Microeconometrics: Methods and Application*. New York: Cambridge University Press.

Canadian Institute for Health Information (2012). National Health Expenditure Trends, 1975 to 2012. Technical report.

Clavet, N.-J., Duclos, J.-Y., Fortin, B., and Marchand, S. (2012). Le Québec, 2004-2030 : une analyse de micro-simulation. Project Report CIRANO 2012RP-16.

Clavet, N.-J., Duclos, J.-Y., Fortin, B., Marchand, S., and Michaud, P.-C. (2013). Les dépenses en santé du gouvernement du Québec, 2013-2030: projections et déterminants. Scientific booklet CIRANO 2013s-45.

Daouli, J., Davillas, A., Dermoussis, M., and Giannakopoulos, N. (2014). Obesity persistence and duration dependence: Evidence from a cohort of US adults. *Economics & Human Biology*, 12:pp.30–44.

Décarie, Y., Boissonneault, M., and Légaré, J. (2012). An inventory of Canadian microsimulation models. SEDAP Research Paper No. 298.

Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1998). An analysis of sample attrition in panel data: The Michigan panel study of income dynamics. *The Journal of Human Resources*, 33(2):251–299.

Frees, E., Gao, J., and Rosenberg, M. (2011). Predicting the frequency and amount of health care expenditures. *North American Actuarial Journal*, 15(3):377–382.

Godbout, L., St-Cerny, S., Arseneau, M., Dao, N. H., and Fortin, P. (2014). La soutenabilité budgétaire des finances publiques du gouvernement du Québec. Working document 2014/01 of the Chaire de recherche en fiscalité et en finances publiques de l'Université de Sherbrooke.

Goldman, D., Cutler, D., Rowe, J., Michaud, P.-C., Sullivan, J., Peneva, D., and Olshansky, S. (2013). Substantial health and economic returns from delayed aging may warrant a new focus for medical research. *Health Affairs*, 32(10):pp.1698–1705.

Goldman, D., Shekelle, P., Bhattacharya, J., Hurd, M., Joyce, G., Lakdwalla, D., Matsui, D., Newberry, S., Panis, C., and Shang, B. (2005). Health status and medical treatment of the future elderly. Technical report, RAND. Récupéré le 12 janvier 2014 de http://www.rand.org/content/dam/rand/pubs/technical_reports/2005/RAND_TR169.pdf.

Gutterman, S. and Vanderhoof, I. T. (1998). Forecasting changes in mortality: a search for a law of causes and effects. *North American Actuarial Journal*, 2(4):135–138.

Health Canada (2012). Smoking Prevalence 1999 — 2012. Technical report.

Human Mortality Database (2005). Complete data series. Technical report.

Institut de la statistique du Québec (2009). Perspectives démographiques du Québec et des régions, 2006-2056. Technical report.

Institut national de santé publique du Québec (2006). Les maladies du coeur et les maladies vasculaires cérébrales: prévalence, morbidité et mortalité au québec. Technical report, Canada.

Institut national de santé publique du Québec (2011). Surveillance de l'hypertension au Québec: incidence, prévalence et mortalité. Technical report, Québec.

Institut national de santé publique du Québec (2012). Surveillance du statut pondéral chez les adultes québécois : portrait et évolution de 1987 à 2010. Technical report.

Kapteyn, A., Michaud, P., Smith, J., and Soest, A. (2006). Effects of attrition and non-response in the health and retirement study. WR-407.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting U. S. mortality. *Journal of the American Statistical Association*, 87(419):pp. 659–671.

Macdonald, A. (1997). *The Second Actuarial Study of Mortality in Europe: The Companion Disk*. Groupe consultatif des associations d'actuaires des pays des Communautés Européennes.

McDonald, A., Cairns, A., Gwilt, P., and Miller, K. (1998). An international comparison of recent trends in mortality. *British Actuarial Journal*, 4:3–141.

Michaud, P.-C., Lakdawalla, D., Goldman, D., Zheng, Y., and Gailey, A. (2012). The value of medical and pharmaceutical interventions for reducing obesity. *Journal of Health Economics*, 21(4):pp.630–640.

National Research Council (2012). Improving health care cost projections for the medicare population. Technical report.

Payeur, F. F. (2013). La mortalité et l'espérance de vie au Québec en 2012. Technical report, Institut de la statistique du Québec.

Public Health Agency of Canada and Canadian Institute for Health Information (2011). Obesity in Canada, Determinants and Contributing Factors. Technical report.

Régie des rentes du Québec (2013). Évaluation actuarielle du Régime de rentes du Québec au 31 décembre 2012.

Renshaw, A. and Haberman, S. (2003a). Lee– Carter mortality forecasting: a parallel generalized linear modelling approach for england and wales mortality projections. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):119–137.

Renshaw, A., Haberman, S., and Hatzopoulos, P. (1996). The modelling of recent mortality trends in United Kingdom male assured lives. *British Actuarial Journal*, 2:449–477.

Renshaw, A. E. and Haberman, S. (2003b). Lee–Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics*, 33(2):255–272.

Renshaw, A. E. and Haberman, S. (2003c). On the forecasting of mortality reduction factors. *Insurance: Mathematics and Economics*, 32(3):379–401.

Réseau de recherche en santé et en sécurité du travail du Québec (2014). National Population Health Survey. Technical report.

Statistics Canada (2000a). Canadian Community Health Survey, 2010, Wave 1.1, Annual component. Data sets.

Statistics Canada (2000b). Labor Force Survey, 2000. Data sets.

Statistics Canada (2010). Canadian Community Health Survey, 2010, Wave 5.1, Annual component. Data sets, Statistics Canada.

Statistics Canada (2012a). Canadian Community Health Survey, 2012, Annual component. Data sets.

Statistics Canada (2012b). Labor Force Survey, 2012. Data sets.

Statistics Canada (2012c). National Population Health Survey—Household component — longitudinal (NPHS). Technical report.

Statistics Canada (2012d). National Population Health Survey, 1994-2010, waves 1-9, Household component- longitudinal. Data sets.

Statistics Canada (2013). Life tables, Canada, provinces and territories 2007 to 2009. Technical Report 003, Statistics Canada.

Statistics Canada (2014a). Table 102-0218 -Life expectancy, abridged life table, by age group and sex, three-year average, Canada, provinces, territories, health regions and peer groups, *Terminated* occasional (years).

Statistics Canada (2014b). Table 102-0512 - Life expectancy, at birth and at age 65, by sex, Canada, provinces and territories annual (years).

United Nations (2008). World Population Prospects – The 2008 Revision. Technical report, United Nations - Population Division.

World Bank (2014). Data: By Country: Canada [Table]. Data sets, World Bank.

World Health Organization (2006). BMI classification. Technical report, United Nations.