Appendix 3: Updating CVD risks

Cardiovascular disease risks were updated yearly through prediction algorithms that were generated using the longitudinal National Population Health Survey, the Canadian Community Health Survey and the Canadian Heart Health Survey. The cardiovascular disease risk algorithms were applied separately to each actor on a yearly basis for each cardiovascular disease risk. As actors were added to the model over time (through births or deaths), they also had cardiovascular disease risk algorithms applied each year. Similarly, actors who died or emigrated were no longer part of the model population and were no longer updated. The predicted change for the following year for each cardiovascular disease risk was estimated using the predictor variables contained in the algorithms that were specific to the particular cardiovascular disease risk. Each cardiovascular disease risk algorithm is described in detail below. Risk algorithm development used different statistical approaches depending on the specific risk and data source; however, there was a common approach. The approach included: assessment of predictive risks based on established relationships and empiric findings from Canadian national development data; and the balancing of discrimination and calibration and/or risk factor distribution.

eTable 3.1: DIABETES	
Name	Diabetes Population Risk Tool
Objective/Purpose	To create and validate a population-based risk prediction tool for incident diabetes using commonly collected national survey data.
Process of Development	
- Data source/ development data	1996/7 **National Population Health Survey** for 23,403 Ontario residents collected by Statistics Canada. Survey data was linked to administrative data (**Ontario Diabetes Database**) to ascertain physician diagnosed diabetes status.
- Outcome definition	Outcome was physician diagnosed diabetes defined in administrative data as a hospital admission with a diabetes diagnosis code 250 (International Classification of Disease-Canadian Modification) before 2002 or International Classification of Disease -10 code E10-E14 after 2002; OR a physician services claim with a diabetes diagnosis followed within 2 years by either a physician services claim or a hospital admission with a diabetes diagnosis. Cases of gestational diabetes were excluded.
- Predictive/stratifying/causal variables	The variables used to model diabetes incidence were based on established evidence, easily captured using population surveys and in a consistent manner across surveys and populations and included age, height, weight, chronic conditions diagnosed by a health professional, ethnicity, immigration status, smoking status, educational achievement, household income, alcohol consumption and physical activity. Important predictor variables for men were hypertension, non-white ethnicity, heart disease, current smoking status, education and age/body mass index category. The general form of the male model is outlined below: **Log(Diabetes incidence time)= $\alpha + ß_1HTN + ß_2NW + ß_3HD + ß_4CS + ß_5ED + ß_6age/BMI + \sigma\varepsilon$**

	with ε following the extreme value distribution Important predictor variables for women were hypertension, non-white ethnicity, immigration, education and age/BMI category. The general form of the female model is outlined below: **Log(Diabetes incidence time) = $\alpha + ß_1HTN + ß_2NW + ß_3IMM + ß_4ED + ß_5age/BMI + \sigma\varepsilon$**
- Statistical Method	Weibull accelerated failure time model allows user to predict diabetes probability for range of follow-up periods. Diabetes functions were derived separately for men and women.
Performance Characteristics and Assessment Process	
- Validation/external validation	The model was assessed for discrimination and calibration. Discrimination was measured using a C-statistic modified for survival data. C- statistics ranged between 0.77 and 0.79 in the development and validation datasets, indicating good discrimination. Accuracy or calibration was measured using a Hosmer-Lemeshow chi-squared statistic modified for survival data. Calibrated models in validation cohorts fell below the acceptable cut off of 20, indicating good calibration. The model was externally validated on 2 external datasets from Ontario and Manitoba and performed well both in terms of discrimination and calibration.
- Estimates produced/ valid for which subpopulations?	Estimates of diabetes incidence were produced for age, BMI, ethnicity and educational level subgroups
Plain Language Summary	Diabetes Population Risk Tool is a means to determine diabetes risk estimates, using national survey data to calculate the number of Canadians at risk of developing diabetes and to determine how this disease risk is distributed among the population. The purpose of the study was not only to provide risk estimates but also to inform health policy. Accuracy and discrimination of the model was described by comparing observed diabetes rates with predicted estimates. Results of the study provided predictive risk factors, including body mass index, age, ethnicity and smoking. Through the two external cohorts, the Diabetes Population Risk Tool showed good discrimination and calibration. Models such as the Diabetes Population Risk Tool can inform healthcare planning and disease prevention strategies.
References	L. Rosella, D. Manuel, C. Burchill, T. Stukel and the PHIAT-DM team. A population-based risk algorithm for the development of diabetes: development and validation of the Diabetes Population Risk Tool (DPoRT). J Epidemiol Community Health. June 2010.

<table>
<tr><th colspan="2">eTable 3.2: CHOLESTEROL AND BLOOD PRESSURE</th></tr>
<tr><td>Name</td><td>Cholesterol, high density lipoprotein and hypertension (POHEM:cardiovascular disease)</td></tr>
<tr><td>Objective/Purpose</td><td>To derive the joint probability of changing cholesterol and blood pressure states from one age group to the next.</td></tr>
<tr><td>Process of Development</td><td></td></tr>
<tr><td>- Data source/ development data</td><td>1986 to 1992 Canadian Heart Health Surveys</td></tr>
<tr><td>- Outcome definition</td><td>Cholesterol and high density lipoprotein were categorized in 5 groups, tabulated below for simplicity:
<table>
<tr><th>Cholesterol mmol/L</th><th>Total Cholesterol</th><th>High density lipoprotein</th></tr>
<tr><td>Low</td><td><4.15</td><td><0.90</td></tr>
<tr><td>Low-medium</td><td>4.15-5.17</td><td>0.90-1.16</td></tr>
<tr><td>Medium</td><td>5.18-6.21</td><td>1.17-1.29</td></tr>
<tr><td>Medium-high</td><td>6.22-7.24</td><td>1.30-1.54</td></tr>
<tr><td>High</td><td>7.25+</td><td>1.55+</td></tr>
</table>
Blood pressure (systolic and diastolic) was also categorized into 5 groups. The definition for each group, except the optimal group, depends on reaching a cut off value for either systolic OR diastolic blood pressure tabulated below for simplicity:
<table>
<tr><th>Blood pressure (mm Hg)</th><th>Systolic</th><th>Diastolic</th></tr>
<tr><td>Optimal</td><td><120</td><td>AND <80</td></tr>
<tr><td>Normal</td><td>120-130</td><td>OR 80-85</td></tr>
<tr><td>High normal</td><td>130-140</td><td>OR 85-90</td></tr>
<tr><td>Hypertensive stage I</td><td>140-160</td><td>OR 90-100</td></tr>
<tr><td>Hypertensive stage II-IV</td><td>≥160</td><td>OR ≥100</td></tr>
</table></td></tr>
<tr><td>- Predictive/stratifying/ causal variables</td><td>5 year age group, sex, body mass index, diabetic status</td></tr>
<tr><td>- Statistical methods for imputing cholesterol, HDL and blood pressure from CHHS to the POHEM startup population (CCHS 2001)</td><td>Specifically, using variables common to the Canadian Community Health Survey and Canadian Heart Health Survey, separate initial values of blood pressure and cholesterol categories for each individual were imputed using “hot-deck” methods. In other words, individuals, in the Canadian Heart Health Surveywere matched to those in the Canadian Heart Health Surveybased on 5-year age-group, sex, self-reported hypertension, body mass index category and diabetes status and were assigned the corresponding categories total cholesterol/blood pressure available in the Canadian Heart Health Survey. High density lipoprotein was subsequently imputed based on the total cholesterol level also using Canadian Heart Health Surveyrecords, for persons having the same 5 year age-group, sex, total cholesterol, body mass index and diabetes status.</td></tr>
</table>

- Statistical method for determining the transition probabilities from one cholesterol/ blood pressure group to another	Once blood pressure and cholesterol categories were imputed the joint transition probabilities of changing blood pressure and cholesterol were estimated. The Canadian Heart Health Surveyis a cross sectional data source, therefore the joint transition probabilities were estimated from one 5-year age group to the next. Transport flow methodology, with the SAS NETFLOW procedure, was used to estimate these transition probabilities. Essentially this method generates a probability matrix that allows for change in individual actors' states (of total cholesterol, high density lipoprotein and blood pressure) while minimizing the flow needed to achieve the observed distribution. Diabetes and body mass index are correlated with cholesterol and blood pressure levels and were therefore included as covariates to smooth the transitions of cholesterol and blood pressure. Using a period, actuarial or life table approach it was assumed that the transition probabilities between age groups estimated from the cross-sectional Canadian Heart Health Surveywould also apply over time as individuals in the population aged.
Performance Characteristics and Assessment Process	
- Validation/external validation	During development and subsequently, the accuracy of this prediction method was assessed against observed population-based survey data, however as this was not a classical prediction model, no formal statistical assessment of calibration was made. This study will be the first published external validation of the cholesterol, high density lipoprotein and hypertension module.
- Estimates produced/valid for which subpopulations?	Not reported
Plain Language Summary	Using the Canadian Heart Health Surveyseparate initial values of blood pressure, total cholesterol and HDL were imputed into the POHEM start-up dataset (Canadian Community Health Survey 1.1). In addition, transition probability matrices, that allowed co-evolution of blood pressure, total cholesterol and HDL categories were calculated from the Canadian Heart Health Surveydata and used in POHEM to allow these physiological measures to change over time.
References	None published

eTable 3.3: SMOKING	
Name	Smoking (Canadian Cancer Risk Management Model) integrated into POHEM: cardiovascular disease
Objective/Purpose	To estimate the distribution of smokers in the Canadian population and to generate a probability of moving from one smoking group to another, given a simulated individual's 5-year age group, sex, province and former smoking category.
Process of Development	
- Data source/ development data	1979 **Canadian Health Survey**, 1994 **National Population Health Survey** and 2008 **Canadian Community Health Survey**
- Outcome definition	Smokers were categorized into 3 exposure groups, non-smoker, light smoker (less than 20 cigarettes per day) and heavy smoker (20 or more cigarettes a day). These categories were derived from variables in the Canadian Health Survey which had 4 categories: heavy smoker, light smoker, former smoker, never smoked; and from variables in the Canadian Community Health Survey and National Population Health Survey which had 4 categories: never smoked, daily smoker, occasional smoker and former smoker. For daily smoker, if the number of cigarettes smoked per day is >=20 then the individual is categorized as heavy, else light smoker. For occasional smokers, if the number of cigarettes per months is >=600 then the individual is categorized as heavy, else light smoker. For never smoked or former smoker the individual is categorized as non-smoker.
- Predictive/stratifying/ causal variables	Age, sex and region (province)
- Statistical method for determining the distribution of smokers	Starting with data of persons aged 15-19 years in 1979 from CHS (and 30-34 in 1994 from National Population Health Survey), the percentage of heavy smokers, light smokers and non smokers (pH, pL, pN) in 1979 and in 1994 was computed. For any year (y) between 1979 and 1994 (including these bounds), the following **linear interpolation** was applied: pH(y)=pH(1979) + (pH(1994)-pH(1979))/15*(y-1979), pL(y)=pL(1979) + (pL(1994)-pL(1979))/15*(y-1979), pN(y)=pN(1979) + (pN(1994)-pN(1979))/15*(y-1979) [because 15 is the number of years between 1979 and 1994]. The same process was repeated for each province, each age group, sex and for the time period between 1994 (National Population Health Survey data) and 2008 (Canadian Community Health Survey data).
- Statistical method for determining the transitional probabilities from one smoking group to another	For the time period 1979-1994, it was assumed that the persons could change their exposure group only during the last year of a 5-year period, i.e. in 1984, 1989 and 1994. Therefore, 3 transition periods were considered: 1979-1984, 1984-1989, and 1989-1994. For each of these periods, a transition matrix was built, where the row marginals are the prevalence rates for the beginning year, and the columns marginals are the prevalence observed for the end year (for the age group 5 years older), see example and tables below. The transition matrix was filled in using the North-West method, where the values of the cells are determined in such a way that (i) the marginals are satisfied, (ii) a constraint is to minimize the proportion of persons

<table>
<tr><td></td><td>changing their exposure group, thus reproducing the likely tendency to reduce the probability of changing one's exposure group, (as a consequence, most values will lie in the diagonal) (iii) filling in starts at the cell at the North-West corner of the table. As an additional constraint, we "forbid" any person to transit directly from Non-smoker to Heavy smoker and vice versa. Once the values in the transition matrix have been determined, they are transformed into probabilities of falling in a given group, given their (initial) group.
For example, if, for a given sex, province, age group the row marginals are 0.1, 0.6 and 0.3 (representing the proportion of heavy, light and non-smokers) and the columns marginals (for the age group 5 years older 5 years later) are: 0.2, 0.7, 0.1, then the transitions satisfying the method are indicated in the "Change required" table below. Then these values are transformed into probabilities of falling in a given final group, in the following way (see "Probability matrix" table). For persons who were Heavy smokers at the beginning of the period, the probability of remaining smoker is 1 (100%); for the persons who were Light smokers at the beginning of the period, the probability is .5/.6 of remaining light and .1/.6 of becoming a heavy smoker; for the ones who were non-smokers initially, the probability of becoming light and remaining non-smoker is respectively .2/.3, .1/.3.
Change required Probability matrix
<table>
<tr><td></td><td>To</td><td>H</td><td>L</td><td>N</td><td></td><td></td><td>H</td><td>L</td><td>N</td></tr>
<tr><td>From</td><td></td><td>0.2</td><td>0.7</td><td>0.1</td><td rowspan="4"></td><td></td><td>0.2</td><td>0.7</td><td>0.1</td></tr>
<tr><td>H</td><td>0.1</td><td>0.1</td><td></td><td></td><td>0.1</td><td>1</td><td></td><td></td></tr>
<tr><td>L</td><td>0.6</td><td>0.1</td><td>0.5</td><td></td><td>0.6</td><td>1/6</td><td>5/6</td><td></td></tr>
<tr><td>N</td><td>0.3</td><td></td><td>0.2</td><td>0.1</td><td>0.3</td><td></td><td>2/3</td><td>1/3</td></tr>
</table>
For the period 1994-2008, 3 transitions were also considered: 1994-1999, 1999-2003, 2003-2008 (that is, periods of respectively 4 and 5 years). During this period more data were available and were used in this paper to externally validate the linear interpolation.
Adjustments of the method were made for the cases where the age group was the youngest or the oldest; also, to extend the projection before 1979 and after 2008, some assumptions were made. In particular, for projection after 2008, it was assumed that the marginals for a given age group followed the trend already observed in period 2003-2008.</td></tr>
<tr><td>Performance Characteristics and Assessment Process</td><td></td></tr>
<tr><td>- Validation/external validation</td><td>During development and subsequently the accuracy of this prediction method was assessed against observed population-based survey data, in order to ensure that the distribution of the smoking groups was maintained over time and between data sources.
This study will be the first published external validation of the smoking module.</td></tr>
<tr><td>- Estimates produced/ valid for which subpopulations?</td><td>Estimates of smoking prevalence were produced for 5 year age groups, sex and province.</td></tr>
<tr><td>- Assumptions</td><td>For the CHS 1979, only 5 regions were available (instead of provinces):</td></tr>
</table>

	Atlantic, Québec, Ontario, Prairies and Vancouver. In designing the linear interpolation it was assumed that all 4 Atlantic provinces shared the same prevalence and that all Prairie provinces shared the same prevalence and that British Columbia was similar to Vancouver.
Plain Language Summary	The smoking module of the Canadian Cancer Risk Management Model provided a means to estimate the prevalence of smoking exposure by different age groups, sex and province, using observed population-based data. These prevalences provided the data necessary to calculate the probability of transitioning from one smoking group to another, given 5-year age group, sex, province and former smoking category.
References	None published

eTable 3.4: OBESITY	
Name	Body weight (POHEM: cardiovascular disease)
Objective/Purpose	To estimate equations to model change in weight over time among Canadian adults.
Process of Development	
- Data source/ development data	**Longitudinal National Population Health Survey**
- Outcome definition	The outcome was change in BMI over time. Weight change was assessed by calculating an individual's body mass index from self-reported weight and height, at each of five consecutive cycles of the National Population Health Survey, conducted in two year intervals from 1996–97 to 2004–05. A Box-Cox transformation was applied to normalize the distribution of the outcome variable.
- Predictive/stratifying/ causal variables	Age, sex, income quartile, education, region of residence and previous values of body mass index.
- Statistical Method	Linear regression models, stratified by age, sex and previous body mass index category, were used to estimate change in individual's self-reported body mass index using National Population Health Survey data from 1996–97 to 2004–05. In addition, because previous body mass index may be limited to 1, 2, 3 or 4 occasions in the course of the simulation, 4 versions of the regressions were constructed. Briefly, each of the 4 sets of regression were subdivided into 28 groups according to age, group, sex and previous body mass index category, totalling 112 separate regressions. The general form of the regression equation is outlined below: $\mathbf{(30+\Delta BMI)^{\lambda} = \alpha + ß_1 BMI + ß_2 I + ß_3 E + ß_4 R + \sigma\varepsilon}$ Each of these models predicts a 2 year change in BMI. In this analysis outliers were excluded based on the Student's residuals and cases with missing values for the important variables were removed from the analysis. Because income was frequently missing, imputation was used rather than letting the affected observations go to waste.
Performance Characteristics and Assessment Process	
- Validation/external validation	Regressions were validated by performing within-sample 5-fold cross-validation. The average error between predicted and observed body mass index was 1.60 kg/m^2 for females and 1.25 kg/m^2 for males. Regressions with smaller number of cases (underweight category) had errors as high as

	3.62 kg/m^2. There appeared to be little bias in the error, so that the overall distribution of body mass index in the entire sample was well represented. The 112 R-squares values derived from the regression models ranged from 35% to 85%. After implementing the regression in POHEM:cardiovascular disease further external validation was completed by comparing the projected distribution of body mass index to that observed in the Canadian Community Health Survey in 2003, 2005 and 2007. In addition, further validation was undertaken to evaluate the projections over a longer period.
- Estimates produced/valid for which subpopulations?	Not reported.
Plain Language Summary	Using the longitudinal National Population Health Survey, a set of regression models to estimate weight change over time were estimated for Canadian adults. The outcome modeled was change in body mass index and explanatory variables were age, sex, income quartile, education and region of residence in Canada and previous body mass index value(s). The models were internally validated using 5-fold cross validation and externally validated by comparing projections of body mass index categories to observed data collected in later years (Canadian Community Health Survey).
References	None published