Planning Tool Guidance

Project Big Life

2019-08-09

Contents

Chapter 1

Welcome to Project Big Life's Planning Tool

TO DO: Insert the image for the PBL planning tool

1.0.1 What is the Project Big Life Planning Tool

To do: Develop video showcasing the platform including what it is and why someone should use the platform

1.0.2 Who made the Project Big Life Planning Tool?

The Project Big Life Planning Tool was developed by the Project Big Life Team. The Project Big Life Team is part of the ICES. The below video explains what ICES is.

```
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```

Chapter 2

Introduction

The Project Big Life Planning Tool was developed in order to support health professionals: research, plan, develop, and evaluate evidence-based health interventions.

For instance Project Big Life Planning Tool helps:

- Public health professionals: assess the impact of preventative interventions that target health behaviours
- Health planners: assess the need for palliative care

What types of questions can it answer?

The Project Big Life Planning Tool can answer the following types of questions:

- What is the burden of smoking on life expectancy?
- How many deaths would be prevented if everyone met their daily exercise requirements?

How does it work?

- This tool provides health planners with access to multivariable predictive risk algorithms, created and housed by the Project Big Life Team.
- The multivariable predictive risk algorithms use distinct characteristics and health profiles of groups of people to assess the risk of a health outcome (e.g. Life Expectancy).
- The multivariable predictive risk algorithms are developed and validated using data routinely collected by Statistics Canada and provincial health agencies, and the algorithms have been published in various journals.
- More information about multivariable predictive risk algorithms can be found in the key concepts (Chapter 6).

Why should I used it?

- It is **easy** and **flexible** to use.
  - The user only needs to upload their data and choose which calculation to run.
  - It can be used to assess the future risk of a health outcome.
  - It can be used to assess the effectiveness of different intervention scenarios (e.g. policy) on a health outcome.

- It generates **accurate** predictions.
  - It can be used to accurately assess the risk of a health outcome in populations that were not used in its developement, and groups of people that account for only a fraction of the population.
- It is **private**.
  - Loaded data remains on your computer and is not uploaded or sent anywhere.

Chapter 3

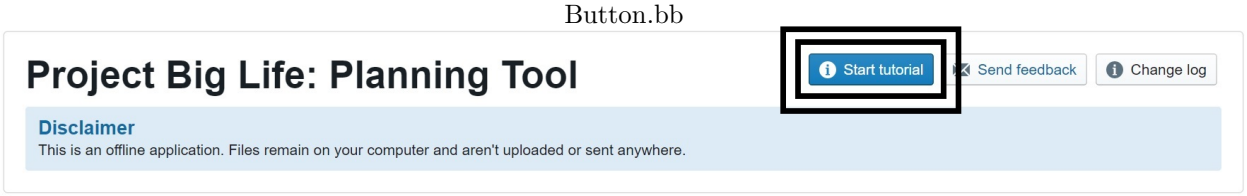
Getting Started

To help you get started quickly with Project Big Life's Planning Tool, we built a Tutorial directly onto the platform.

The tutorial takes you through step-by-step how to use Project Big Life's Planning Tool. The tutorial will not explain the steps in detail (Chapter 4) nor will it provide reference material (Chapter 7), but it will give you an understanding of how easy it is to use the Planning Tool!

To access the tutorial, go onto Project Big Life's Planning Tool (http://policy.projectbiglife.ca/) and click on the Tutorial button in the top right corner!

Button.bb



Chapter 4

How To

These guides will cover the topics covered in the tutorial but in greater detail.

- Customize data
- Load data
- Select calculation
- Filter data
- Stratify data
- Run scenarios: Intervention and Cause-deleted
- Calculate results
- Visualize data (TBD)
- Export data (TBD)
- Resolve error messages (TBD)

4.1 Customize Data

Prior to using the Project Big Life Planning Tool you may want to manipulate your data set. Reasons include: customized filter(s) and/or customized stratification(s).

Data manipulation can occur on any programming software: R, SAS, STATA, etc, as long as you output your data set as a '.csv' file.

An example of customizing your data set is converting the variable: Body Mass Index (CCHS 2013 variable HWTGBMI) from a continuous variable into four distinct categories:

- Underweight: BMI less then 18.5
- Normal or Healthy Weight: BMI of 18.5 to 24.9
- Overweight: BMI of 25.0 to 29.9
- Obese: BMI greater or equal to 30.0

Steps

The following steps show the R code that would be used to create these strata:

1. Convert observations "Not stated" from 999.99 to NA

```
data[data == 999.99] <- NA
```

2. Load the R package dpylr. This package is used for data manipulation.

```
library(dpylr)
```

3. Create a new column that contains four categories for BMI

```
data$newcolumn <- cut(data$HWTGBMI, breaks = c(0,18.5,25,30,Inf),  labels=c("Underweight", "Healthy
```

4. The output will be your data set + a new column with the corresponding category ("Underweight", "Healthy", "Overweight", "Obese") for that individual.

```
  HWTGBMI   newcolumn
1   22.68     Healthy
2   26.99  Overweight
3      NA        <NA>
4   34.44       Obese
5   23.77     Healthy
6   17.23 Underweight
```

This new column can be now be used with the Project Big Life Planning Tool for the purpose of filtering or stratification.

4.2 Load data

Only one data set can be used on the platform for each calculation; a calculation cannot be preformed across multiple data sets.

Note: Data sets loaded to the Project Big Life Planning Tool remains on your computer and is not uploaded or sent anywhere.

There are two options for your data: use a sample file or load your own file.

Note: The Project Big Life Planning Tool can currently only support .csv data files from the 2013/2014 Canadian Community Health Survey. Both the PUMF and Shared 2013/2014 Canadian Community Health Survey files are accepted.

- More information on the Canadian Community Health Survey and the types of accepted data can be found in Key concepts (Chapter 6) - "Data and Sample Files".

4.2.1 Load sample files

If you don't have your own data or want to explore the platform's capabilities before using your data, you can use the sample files on the Project Big Life Planning Tool. There are **X** sample files you may use:

- **data.sample.csv** this is fabricated dataset that includes all variables required for calculation, recommend for calculation, and needed for the real life application examples (found in Chapter 5). This data set is based on the 2013/2014 Canadian Community Health Survey Public Use Microdata File from, however this data set was created only for example purposes and **cannot** be used for analysis.

Steps

1. Click on the file name under "Sample files" to select it.

4.2.2 Load your own file

Steps

1. Click the **Browse** button under "Select a file to use in calculations".
2. Locate the file on your computer, select, and open.

- If the loaded file has all of the variables required and recommended for calculation, you will be able to continue with the planning tool.
- If the loaded file does **not** have all the variables **required** for the calculation you will not be able to continue with the planning tool.
- If the loaded file does **not** have all the variables **recommended** for calculation you will be able to continue with the planning tool, however the calculations may be less accurate.

4.3 Select calculation

There are two general types of calculations: summary measures and by row measures.

- **Summary measures:** When selected, the result will be a single measure for the entire dataset. For instance when Summary Measure - Life Expectancy (Summary Measure) is selected the result is a single life expectancy at birth for the given for the population.
- **By row measures:** When selected, the result will be the measurement for each individual (e.g. row) in the dataset.

Summary measures must be selected for calculations that have stratifications, intervention scenarios, and cause-deleted scenario.

Steps

1. Check the box beside the calculation's name under "Calculations" to select it. A single calculation or multiple calculations may be selected.

- Once selected the name(s) of the calculation(s) will appear to the right of "Calculations".

More details of what the calculations are and how they are preformed can be found in key concepts (Chapter 6) - "Calculations".

4.4 Filter data

Use filters when you want to analyze only a subset of your data.

Steps

1. Click on the + **New** button under "Filters".
2. Select the variable that you want to filter on by typing its variable name into the "search variables" text bar.

3. Filter **in** the categories/levels within the variable by:

- **Categorical:** Clicking on the "Search categories"" text bar. Scroll and select the category you want to **keep** in your data. Repeat this step to add additional categories.
- **Continous:** Click the "cycle button" found under the variable you have selected. Two new boxes will appear.
  - Select the minimum value for your subset data in the box on the left by typing the values or using the arrows.
  - Select the maximum value for your subset data in the box on your right by typing the value or using the arrows.

4. To add another filter repeat the steps above. A maximum of three filters are recommended to maintain statistical power (added filters reduce sample sizes and reduces statistical power).

- Once selected, the name(s) of the filtered variable(s) will appear to the right of "Filters".

You are able to filter on all types of variables: required for calculation, recommended for calculation, and ignore variables (includes customized variables).

4.4.1 Remove a filter

- To remove a filter entirely, click on the trash can beside the variable you want to delete.
- To remove a level within a filtered variable - categorical, click on the 'x' beside the variable level.

4.5 Stratify data

Use stratifications when you want to get a result for multiple strata (levels or classes).

A summary measure must be selected for stratifications, as only a summary measure will be outputted for each strata. By row measurements may also be selected but they will not be stratified.

Steps

1. Click the box beside either "Life Expectancy (Summary)" or "Deaths (Five Years)" under the "Calculation" drop down.
2. Select the variables you want to stratify on under the "Stratifications'. You are only able to stratify on categorical variables.
3. To add another variable for stratification repeat the steps above. A maximum of 3 stratifications are recommended to maintain statistical power (added strata reduce strata sample size and reduces statistical power).

- Once selected, the name(s) of the stratified variable(s) will appear to the right of "Stratifications".

You are able to stratify on all types of categorical variables: required for calculation, recommended for calculation, and ignore variables (includes customized variables).

4.5.1 Remove a stratification

- To remove a stratification variable click on the 'x' beside the variable level.

4.6 Run scenarios

Scenarios can be used to predict the health outcomes when unhealthy behaviours:

- are modified in the population: **Intervention**, or
- were never present in the population: **Cause-deleted**

Scenarios can be used to inform potential health policies or programs.

4.6.1 Intervention scenarios

Interventions provide you with the ability to customize the scenarios. For example you can answer the questions:

- what if we only had 15% of the population smoked rather then the current 20%?
- what if everyone increased their physical activity by 10%?
- what if everyone ate 4 fruit servings each day?
- what if everyone drank 2 fewer drinks per week?

These intervention scenarios allow you to predict and compare the effectiveness of policies.

There are 3 types of intervention scenarios that you can select:

- **Absolute:** each individual in the population **changes** their health behaviour **by a value of x**.
- **Relative:** each individual in the population **changes** their health behaviour **by a ratio of y**.
- **Target:** each individual in the population **has a set value of z.**

More information on the specifics of each type of intervention scenario and how they are calculated can be found in Key Concepts (Chapter 6) - "Scenarios: Interventions".

Steps

1. Click the box beside either "Life Expectancy (Summary)" under the "Calculation" drop down.
2. Check the button beside: "Intervention" under the "Scenario" drop down.
3. Click on the health behaviour you want to modify: e.g. Diet.

- A drop down menu of all the possible variables of that health behaviour you can modify will appear.

4. Click the box beside the variable you want to modify: e.g. Daily consumption - fruit - (D).

- A drop down menu of the scenario types: absolute, relative, and target, will appear.

5. Check the button beside the type of intervention you want to modify: e.g. Target.
6. Use the arrows beside the the text box: "Decrease by" or "Increase by" to add the value you are modifying: e.g. The value of 4 is used for the scenario what if everyone ate 4 fruit servings each day?

Multiple health behaviours and variables within the health behaviours can be selected for a single calculation.

- Once selected, the name(s) of the health behaviour(s) that have been selected for the intervention will appear to the right of "Scenario".

4.6.2 Cause-deleted scenarios

Cause-deleted scenarios provide you with the ability to see the best case scenario for the population. For example:

- what if no one in the population ever smoked?
- what if everyone in the population met their recommended physical activity levels (3.00 METs/week)?

More information on cause-deleted calculations and how they are calculated can be found in key concepts (*Chapter 6*) - "Scenarios: Cause-deleted".

Steps

1. Click the box beside either "Life Expectancy (Summary)" under the "Calculation" drop down.
2. Check the button beside: "Cause-deleted" under the "Scenario" drop down.
3. Check the box beside the health behaviour that you want to have a cause-deleted calculation.

Multiple health behaviours can be selected for a single calculation.

4.7 Calculate results

1. Name your calculation in the text box: Calculation name.

- Be specific when naming the calculation as it will make it easier to distinguish after running multiple calculations.

Note: the larger the data set is the longer the calculations will take. Depending on the size of the data set and the type of calculation being preformed it could take an hour or more.

4.8 Visualize Data

TBD: Need plots on the platform to work through the steps below - export - create your own(?)

4.9 Download results

Click on the **Download results** button under the **Results** section.

Select which calculations you'd like to download.

To Do: Screenshot of all the calculation options once the platform is fixed.

4.10 Resolve warning or error messages

There are different types of warning and error messages that may appear.

Below describes some of the messages that are likely to occur and steps to resolve them.

Waiting for the branch with the error messages to be finalized and merged

4.10.1 Invalid category

4.10.2 Out of range

"Out of range" is when there are observation(s) in the data set that are not is beyond the limit of

4.10.3 Not a number

4.10.4 Sample size is too small

Chapter 5

Applications

This chapter provides you with examples of how Project Big Life's Planning Tool can be used in your day-to-day operations. The examples will cover:

- analyses that can be included in a health status report,
- a national scenario: What if Canada biked like the Dutch?
- a local scenario: turning Ottawa into the healthiest Canadian region

5.1 Health Status Report

In this example we will highlight statistics that could reported in health status reports. Health status reports are a way to report the health state for a population and the factors that influence the population's health. Information from health status reports are used to inform policy, planning, and resource allocation.

In this example we will calculate:

- The predicted number of deaths by strata
- The impact of eliminating unhealthy behaviours on life expectancy

For this example we will focus on the population of Alberta.

5.1.1 Predicted number of deaths stratified by sex and level of education

By showing the number of deaths by strata, the reader can see the distribution of deaths across specific population characteristics. Any categorical variable can be used for stratification but in this example, we will use sex and level of education.

Steps

1. Select the sample file data.sample.csv under "Sample files".

Note: Although the data.sample.csv is based on the 2013/2014 Canadian Community Health Survey Public Use Microdata File, data.sample.csv is a completely fabricated data set and can only be used for examplary purposes.

2. Select the calculation: Summary Measure – Deaths (Five years)

3. Add filter: GEOGPRV – 48, which is the corresponding code for Alberta.
4. Add two stratification: DDH_SEX and EDUDR04
5. Title the calculation: Deaths by sex and education level
6. Click the calculate button

7. To do: Results – walk through the results

5.1.2 Impact of eliminating unhealthy behaviours: physical inactivity and poor diet, on life expectancy

To show how much an unhealthy behaviour impacts life expectancy we use the scenario: cause-deleted. Cause-deleted scenarios can be used for the health behaviours: alcohol consumption, diet, physical activity, and smoking, individually or in any combination. In this example we will evaluate the impact of physical inactivity and poor diet, in combination, on life expectancy.

Steps

1. Select the sample file data.sample.csv under "Sample files".

Note: Although the data.sample.csv is based on the 2013/2014 Canadian Community Health Survey Public Use Microdata File, data.sample.csv is a completely fabricated data set and can only be used for examplary purposes.

2. Select initial calculation: Summary Measure – Life Expectancy (Summary)
3. Add filter: GEOGPRV – 48, which is the corresponding code for Alberta.
4. Click the text: Scenario.
5. Select Cause-deleted.
6. Select the causes to delete: physical activity and diet
7. Title the calculation: Alberta: Cause-deleted - physical activity and diet
8. Click the calculate button

9. To Do: Walk through results

5.2 Canada bikes like the Dutch

Cycling as a form of transportation, is an effective way of increasing individuals' daily physical activity levels which leads to a decrease in the risk of disease and death.

In this example we will determine the impact on life expectancy if Canadians were to cycle like the Dutch, who are world-renowned for their cycling.

In this example we determine:

A. How much Canadians cycle per day for transportation purposes,

B. How much Dutch cycle per day for transportation purposes,

C. The difference in daily cycling levels for transportation of Canadians and Dutch, and

D. Predict how Canadian life expectancy would change if Canadians biked like the Dutch.

For this example we will use the sample data set **data.sample.csv**, which can be downloaded at https://github.com/Big-Life-Lab/PBL-Planning-Tool---Applications-Vignettes. This link also contains all of the R code for this vignette.

The following steps include the R coding for each of the steps but you can preform the steps in any other software program.

5.2.1 Part A: How much do Canadians cycle per day for transportation purposes

We will determine how much Canadians cycle per day by calculating the average daily energy expenditure from cycling in the current Canadian population.

We will use 3 variables in our sample data set which are based off of the 2013/2014 CCHS PUMF and measure cycling as a form of active transportation:

- PAC_8: In the past 3 months did you cycle to and from work or school?
- PAC_8A: How many times did you cycle to and from work or school, in the past 3 months?
- PAC_8B: How much time did you spend on each occasion?

Step 1: Import your data set to R.

```
PA.data <- read.csv("data.sample.csv")
```

Note: Although the data.sample.csv is based on the 2013/2014 Canadian Community Health Survey Public Use Microdata File, data.sample.csv is a completely fabricated data set and can only be used for examplary purposes.

Assumption #1: we assume individuals 65 or older are retired and therefore not cycling to/from work/school. Therefore, we will only evaluate individuals that are < 65 years old.

Step 2: Filter the data: < 65. The CCHS data set has categories for age, therefore we will filter out all age categories that are >=65 years old.

```
library(dplyr)
PA.data <- PA.data %>%
  filter(DHHGAGE <= 12)
```

We will first evaluate the data for individuals that cycle to/from work/school (PAC_8, answer 1) only:

Step 3: Find the average number of times, that an individual cycled to/from work/school in the past 3 months

a. Exclude the missing data - all individuals that did not cycle in the past 3 months: do not cycle to/from work school, do not know, refused to answer, or did not answer

```
PA.data$PAC_8A[PA.data$PAC_8A == 996] <- NA
PA.data$PAC_8A[PA.data$PAC_8A == 997] <- NA
PA.data$PAC_8A[PA.data$PAC_8A == 998] <- NA
PA.data$PAC_8A[PA.data$PAC_8A == 999] <- NA
```

b. Calculate the average number of times that an individual cycled to/from work/school (mean of PAC_8A)

```
freq.cycle <- mean(PA.data$PAC_8A, na.rm = TRUE)
```

Step 4: Find the average time spent cycling to/from work/school, in the past 3 months

a. Exclude the missing data - all individuals that did not cycle in the past 3 months: do not cycle to/from work school, do not know, refused to answer, or did not answer

```
PA.data$PAC_8B[PA.data$PAC_8B == 6] <- NA
PA.data$PAC_8B[PA.data$PAC_8B == 7] <- NA
PA.data$PAC_8B[PA.data$PAC_8B == 8] <- NA
PA.data$PAC_8B[PA.data$PAC_8B == 9] <- NA
```

b. Calculate the mean time spent cycling (mean of PAC_8B).

Note: Since the PAC_8B variable has 15 minute time periods (1 = 1 to 15 minutes, 2 = 16 to 30 minutes), we will convert the mean into minutes by multiplying the mean of PAC_8A with 15.

```
time.cycle <- 15 * mean(PA.data$PAC_8B, na.rm = TRUE)
```

There are other ways you can do this step. You can take the average of the medians for each time period or use both the minimum and maximum for each time period.

c. Convert the mean time spent cycling to/from work/school from minutes to hours.

```
time.hour.cycle <- time.cycle/60
```

Step 5: Calculate the average duration of cycling. Duration is calculated with the following formula:

Duration = [(Frequency/3 months) $*$ (Time/trip)]/(Days/3 months)

```
duration.cycle <- (time.hour.cycle*freq.cycle)/(30.42*3)
```

Step 6: Calculate the average daily energy expenditure (MET-hours) for cycling to/from work/school

MET-hours = Duration $*$ MET value for walking

Note: For Canadians to cycle like the Dutch, we need to use the Dutch MET cycling value of 5.8 and not the Canadian MET cycling value of 4.0. The Dutch MET value accounts for the speed of cycling of the Dutch.

```
MET.hours.cycle <- (duration.cycle*5.8)
```

We now have our average daily energy expenditure (MET-hours) for Canadians that **DO** cycle to/from work/school.

Now we need to account for the rest of the population that does not cycle to/from work school.

Step 7: Determine the average daily expenditure for all individuals in the population not only those that cycled to/from work/school

a. Calculate the proportion of individuals that cycled to/from work/school in the total population

- Numerator: all individuals that answer Yes to they walked to/from work/school (1 for PAC_7)
- Denominator: total population

```
PA.table.cycle <- table(PA.data$PAC_8)
indiv.cycle <- PA.table.cycle[1]/(sum(PA.table.cycle))
```

b. Calculate the average daily energy expenditure from cycling to/from work/school (MET-hours) for the entire population

- Multiply the average daily energy expenditure for those that did cycle to/from work/school with the proportion of individuals that did cycle to/from work/school

```
MET.hours.cycle.all <- MET.hours.cycle*indiv.cycle
```

On average a Canadian only gets 0.024 MET-hours/day from cycling to/from work/school.

5.2.2 Part B: How much do the Dutch cycle/day for transportation purposes

Using data collected as part of the Dutch National Travel Survey (2010 – 2012), Fisherman, 2015 (doi: 10.1371/journal.pone.0121871 - Table S1) reported the average additional daily energy expenditure from cycling as a form of transportation for the Dutch: males: 1.3 MET-hours, and females: 1.4 MET-hours.

The average Dutch daily energy expenditure from cycling as a form of transportation for both sexes is then 1.35 MET-hours.

5.2.3 Part C: Calculate the difference between Dutch and Canadian daily cycling levels

Canadians need to increase their daily energy expenditure from cycling from 0.07 MET-hours to 1.35 MET-hours, in order to be like the Dutch.

```
Abs.change <- 1.35 - MET.hours.cycle.all
```

5.2.4 Part D: Predict how Canadian life expectancy would change if Canadians biked like the Dutch.

Use the Project Big Life Planning Tool for the following steps:

1. Load your data file: data.sample.csv to the Project Big Life Planning Tool.
2. Select initial calculation: Summary Measure – Life Expectancy (Summary)
3. Add Filter: DDHGAGE – 1,2,3,4,5,6,7,8,9,10,11,12
4. Click: Scenarios, and select Intervention
5. Click Physical Activity then Select “Average daily leisure time energy expenditure in METs”

Assumption #2: Although the scenario for physical activity is for leisure energy expenditure, we assume 1) individuals that are active in their leisure time also use active transportation, and 2) cycling as a form of active transportation is minor part of their energy expenditure. Therefore we will adjust the average daily leisure time energy expenditure in METs.

6. Select Absolute
7. Type in the absolute change calculated in Part 3, into the text box.
8. Name your calculation: Canada bikes like the Dutch
9. Click Calculate
10. Interpret results

5.3 Healthy Cities

What would be Ottawa's life expectancy, if the region had the same health behaviours as the healthiest region in Canada?

In this example, we will determine:

A. who is the healthiest region in Canada,

B. what are the health behaviours of Ottawa vs the healthiest region, and

C. what would be the mortality outcomes be if Ottawa's health behaviours were the same asthe healthiest region.

Here we will use the average life expectancy to determine the health of regions: the higher the life expectancy for the region the healthier that region is.

For this example we will use the sample data set **data.sample.csv**, which can be downloaded at https://github.com/Big-Life-Lab/PBL-Planning-Tool---Applications-Vignettes. This link also contains all of the R code for this vignette.

The following steps include the R coding for each of the steps but you can preform the steps in any other software program.

5.3.1 Part A: Find the healthiest region in Canada

Using the Project Bid Life Planning Tool, we will find the healthiest region in Canada: the region with the highest life expectancy.

1. Load the data file: data.sample.csv.

Note: Although the data.sample.csv is based on the 2013/2014 Canadian Community Health Survey Public Use Microdata File, data.sample.csv is a completely fabricated data set and can only be used for examplary purposes.

2. Select calculation: Summary Measure – Life Expectancy (Summary)
3. Add Stratification: GEODPMF. This variable represents the health regions in Canada
4. Title the calculation: Life expectancy by health region
5. Click the calculate button

6. Download the results - *TO DO*

Using your preferred statistic software (Excel, SAS, STATA, R Studio, etc) identify the healthiest region and the health of the Ottawa health region. The following code is for R.

7. Import the data that you just downloaded to R.

```
data.summary <- read.csv("data.sample_Life_expectancy_by_health_region.csv")
```

8. Sort the data in descending order to find the healthiest region.

```
data.summary[data.summary$SummaryLE == max(data.summary$SummaryLE),]
```

The healthiest region is XXX.

5.3.2 Part B: Determine the health behaviours of the healthiest region and the Ottawa region.

Calculate the average of each health behaviour in the healthiest region: XXX and the Ottawa region.

1. Import your original data set to R.

```
data <- read.csv("data.sample.csv")
```

2. Subset your data for the healthiest region XXX and the Ottawa region.

```
Ottawa <- data %>%
  filter(GEODPMF == 35951)

XXX <- data %>%
  filter(GEODPMF == XXXXX)
```

3. Calculate the average/prevalence of each health behaviour for the XXX and Ottawa region.

Smoking

Calculate the prevalence of current smokers in the XXX and Ottawa region.

```
Ottawa.smokers <- table(Ottawa$SMKDSTY)
Ottawa.smokers.prev <- round(prop.table(Ottawa.smokers),2)
Ottawa.smokers.prev[1]

XXX.smokers <- table(XXX$SMKDSTY)
XXX.smokers.prev <- round(prop.table(XXX.smokers),2)
XXX.smokers.prev[1]
```

Physical Activity

Calculate the average of physical activity (variable PACDEE) in the XXX and Ottawa region.

a. Exclude the missing data: 'not stated'.

```
Ottawa$PACDEE[Ottawa$PACDEE == 99.9] <- NA
XXX$PACDEE[XXX$PACDEE == 99.9] <- NA
```

b. Calculate the average daily energy expenditure for individuals in each region who do have a value for PACDEE

```
Ottawa.PA <- mean(Ottawa$PACDEE, na.rm = TRUE)
XXX.PA <- mean(XXX$PACDEE, na.rm = TRUE)
```

Alcohol

Calculate the average of weekly alcohol consumption (variable ALWDWKY) in the XXX and Ottawa region.

a. Exclude the missing data: 'not stated'

```
Ottawa$ALWDWKY[Ottawa$ALWDWKY == 999] <- NA
XXX$ALWDWKY[XXX$ALWDWKY == 999] <- NA
```

b. Change observations: not applicable to 0. These are individuals who did not drink alcohol in the past year (represented by observations coded as 996 - NA).

```
Ottawa$ALWDWKY[Ottawa$ALWDWKY == 996] <- 0
XXX$ALWDWKY[XXX$ALWDWKY == 996] <- 0
```

c. Calculate the average alcohol consumption for individuals in each region

```
Ottawa.Alcohol <- mean(Ottawa$ALWDWKY, na.rm = TRUE)
XXX.Alcohol <- mean(XXX$ALWDWKY, na.rm = TRUE)
```

Diet

An individual's diet is composed of multiple components including the number of daily servings of: carrot, fruit, juice, potato, salad, and vegetable.

The more daily servings an individual has of:

- carrots, fruits, salad, and vegetables = the more healthy the individual,
- juice = the less healthy the individual,
- potato < 1 = the more healthy the individual, and
- potato > 1 = the less healthy the individual.

To compare the average diet of each region

Calculate the average of each diet component for the XXX and Ottawa region.

a. Eliminate observations: not stated. To reduce the number of coding steps, we will create a data set with only the diet variables.

```
Ottawa.diet <- Ottawa %>%
  select(FVCDCAR, FVCDFRU, FVCDJUI, FVCDPOT, FVCDSAL, FVCDVEG)
Ottawa.diet[Ottawa.diet == 999.9] <- NA

XXX.diet <- XXX %>%
  select(FVCDCAR, FVCDFRU, FVCDJUI, FVCDPOT, FVCDSAL, FVCDVEG)
XXX.diet[XXX.diet == 999.9] <- NA
```

b. Calculate the average of each diet component for XXX and the Ottawa region.

```
summary(Ottawa.diet)
summary(XXX.diet)
```

Comparing health behaviours

4. Compare the averages of all health behaviours in the Ottawa region to the XXX. When the health behaviour for Ottawa is unhealthier then XXX, record the avg/prevalence of that health behaviour for XXX.

5.3.3 Part C: What if Ottawa acted like XXX?

We'll now run a scenario where the Ottawa region acts like XXX. Ottawa will now have the health behaviours of XXX, when the health behaviour of XXX is healthier then Ottawa's.

1. Load your data file: data.sample.csv
2. Select initial calculation: Summary Measure – Life Expectancy (Summary)
3. Add Filter: GEODPMF – 35951 (the code for the Ottawa Health region)
4. Click: Scenarios
5. Click on each health behaviour where Ottawa is unhealthier then XXX and input XXX's value in the target scenario for that health behaviour

- *TO DO:* Examples of where that is

6. Title the Calculation: Ottawa - Intervention - HB1, …..
7. Click calculate.
8. *TO DO:* results

Chapter 6

Key Concepts

This section explains some key concepts in Project Big Life's Planning Tool. This section will explain how it works rather then how to do things.

- Data and sample files
- Multivariable predictive risk algorithms
- Calculations: general
  - Summary vs. By Row
  - Weighted vs. unweighted
- Calculations: specific health outcomes
  - Risk of health outcome
  - Number of health outcomes
  - Life expectancy
- Scenario: Intervention
- Scenario: Cause-deleted
- Limitations

6.1 Data and sample files

The Project Big Life Planning Tool currently accepts **2013/2014 Public Use Microdata File and Shared File of the Canadian Community Health Survey (CCHS) in '.csv' format.**

6.1.1 What is the Canadian Community Health Survey?

The CCHS is an annual cross-sectional survey preformed by Statistics Canada. The CCHS collects information related to health status, health care utilization, and health determinants for the Canadian population. Data is shared at the the sub-provincial geographic level (health region or combination of health regions).

- Details about the survey and its design can be found on Statistic Canada website (http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=144170).
- Details and access to the Public Use Microdata file (PUMF) can be found through the Odesi website (https://search2.odesi.ca/#/details?uri=%2Fodesi%2Fcchs-82M0013-E-2013-2014-Annual-component.xml)

6.1.2 Why can I only use the 2013/2014 Canadian Community Health Survey data?

Each year the CCHS changes a few variables it captures. This makes CCHS data sets from every year other then 2013/2014 incompatible with the algorithms used by the Project Big Life Planning Tool.

The Project Big Life Planning Team is currently working to adjust the algorithms so that they accept other CCHS years, and we will update the guidance when we are done.

6.1.3 Sample Files

- data.sample.csv this is fabricated data set that includes all variables required for calculation, recommend for calculation, and needed for the real life application examples (found in Chapter 5). While this data set was model on the response rates of the 2013/2014 CCHS PUMF (e.g. approximately 80% of responders did not have high blood pressure (CCC_071)), the data is fabricated and random. It **cannot be used for real analysis**.

6.2 Multivariable predictive risk algorithms

Multivariable predictive risk algorithms predict the future risk of health outcomes (e.g. Life Expectancy) for a population using routinely collected health data.

Multivariable predictive risk algorithms can be used to:

- Project the number of new cases of the health outcome
- Estimate the contribution of specific risk factors of the health outcome
- Evaluate effectiveness of health interventions
- Describe the distribution of risk in the population (diffused or concentrated)

Multivariable predictive risk algorithms are able to assess equity issues compared to competing population risk methods (e.g. World Health Organization Global Burden of Disease).

More information on what multivariable predictive risk algorithms are and how they can be used can be found the journal article: *Predictive risk algorithms in a population setting: an overview* (Manuel D, 2012)

6.2.1 Development of multivariable predictive risk algorithms

Data:

- Multivariable predictive risk algorithms are created using routinely collected data that includes information about risk factors (exposure) and health events (outcomes).
- Data is collected at an individual level through population health surveys (e.g. Canadian Community Health Survey) and administrative databases (e.g. Vital Statistics). Data sources are linked together when the individual has given permission too.
- Individuals are followed overtime until the health event (e.g. death or disease) occurs.
- Separate data is collected to create a derivation cohort and validation cohort(s).
  - Note: The risk factors that are collected are from population health surveys and are self-reported; no clinical data (e.g. blood pressure) is collected. Risk factors focus on health behaviours (e.g. smoking) and sociodemographic factors, commonly associated with health outcome.

Algorithm generation:

- Multivariable predictive risk algorithms are cox proportional hazard models that analyze time to health outcome (e.g. death) *Question for Carol - The models are not cox-porportional hazard models but they are similar?*
- Multivariable predictive risk algorithms are developed and validated in 4 stages:
  - Algorithm derivation: the predictive risk algorithm is created using data from the derivation cohort
  - Algorithm validation: the predictive risk algorithm is applied to the validation cohort
  - Final algorithm generation: validation and derivation cohorts are combined to estimate the final application of the predictive risk algorithm
  - Derivation of the application algorithm: creation of a parsimonous (fewer predictors) algorithm that maintained discrimination, calibration, and overall algorithm performance
- In each stage of the algorithm development and validation, algorithm performance is assessed using measures of discrimination and calibration.

6.2.2 Multivariable predictive risk algorithms built in Project Big Life Planning Tool

- There is currently 1 multivariable predictive risk algorithm is built into to Project Big Life planning tool.

Title

Outcomes

Information

Mortality Population Risk Tool

5 year risk of death, Life Expectancy, Cause deleted

Appendix A

6.3 Calculations

6.3.1 Summary vs By Row

There are two general types of calculations Summary Measures and By Row Measures.

Summary measures: When selected, the result will be a single measure for the entire dataset. For instance when Summary Measure - Life Expectancy (Summary Measure) is selected the result is a single life expectancy at birth for the given for the population.

By row measures: When selected, the result will be the measurement for each individual (e.g. row) in the dataset.

Note: Summary Measures are not the same as taking the average of By Row Measures. Summary measures account for the survey weights in their calculations. Only averaging the By Row Measures does not account for the survey weights and will result in an incorrect outcome.

6.3.2 Weighted vs unweighted

Weights are used in complex surveys like the Canadian Community Healthy Survey (CCHS). A weight is given to each respondent in the survey and the weight corresponds to the number of individuals in the population the respondent represents.

When a data set has weights, like the CCHS PUMF data set, the weights are used to calculate the population's outcome e.g. number of deaths.

When a survey does not have weights, the the population's outcome is not calculated with weights.

In either case, weights are not used in the calculation of an individual's outcome (e.g. an individual's 5 year risk of mortality).

6.4 Calculations: specific health outcomes

6.4.1 Risk of health outcome

Risk of the health outcome (e.g. risk of dying) is the outcome of a multivariable predictive risk algorithm. An example of the mutlivariable risk algorithm is:

$$\text{Risk} = \sum_t h_0(t) * e^{\beta_{pred.smoking} * x_{smoking} + \beta_{pred.cancer} * x_{cancer} + \beta_{pred.age} * x_{age} + \ldots}$$

Where:

- t = survival time
- $h_0(t)$ = the baseline hazard
- β_{pred} = predictive hazard ratios for the exposures
- x = the exposure. The exposure can be continuous (e.g. age) or categorical (e.g. smoking status).

Categorical exposures are represented by dummy/factor variables. Each category has its own beta and the exposure is binary. For example smoking status is categorical variable with categories: current, former <= 5 years, former >5 years, or never smoked. For $\beta_{pred.current.smoker}$ the exposure: $x_{current.smoker} = 1$ if the individual currently smokes or 0 if the individual is another type of smoker.

6.4.2 Number of health outcomes

The number of health outcomes (e.g. summary - deaths) is calculated through the following steps:

1. Risk of the health outcome is calculated for each individual (row) in the data set using the mutlivariable predictive risk algorithm.

2. Each individual's (row) risk is weighted with their corresponding survey weight (CCHS PUMF = WTS_M and CCHS shared file = WTS_S).

3. The weighted mean of the health outcome (e.g. mean risk of death) is calculated.

4. The weighted mean is then multiplied with the total number of individuals in the population to generate the number of health outcomes (e.g. number of deaths in 5 years).

6.4.3 Life expectancy

Life expectancy is calculated using abridge life tables *using a modified Chaing approach.(link to reference for Chiang and Hsieh and also one of our papers).* Life expectancy is calculated by two methods: one for summary life expectancy, and a second for by row life expectancy.

6.4.3.1 Summary Life Expectancy

Life expectancy is calculated separately for males and females.

Males:

1. The mortality risk for each male individual is calculated using the male mortality mutlivariable predictive risk algorithm for mortality (MPoRT). Details about the MPoRT can be found in Appendix A.
2. Male individuals are grouped into the 5-year age groups that are used in the 5-year abridge life tables (e.g. 40-44 years old).
3. The weighted average risk of death for each age group is calculated.
4. A male 5-year abridge life table is created using the weighted average risks of death (q(x)) for each age group and the median age for the age group.

Females

Steps 1-4 used to calculate life expectancy for males, are repeated for females using the female MPoRT and a female 5-year abridge life table.

Summary life expectancy

5. The summary life expectancy, or life expectancy of the entire population, is calculated by adding the male life expectancy with the female life expectancy, and taking its average.

Summary life expectancy by strata

Steps (1-4) are repeated for each strata. There will be strata specific weighted risk of death and strata specific life tables.

Step 5 is repeated with the average life expectancy calculated across all strata.

6.4.3.2 By row life expectancy

An individual's life expectancy is calculated by creating a new life table specific to that individual.

These life tables are 1-year abridge life tables, and begin at the individual's age (e.g. an individual that is 43 years old, will have the life table start at 43 years).

1. The probability of death for a person's current age is calculated using the respective MPoRT (male or female), and the individual's health profile (e.g. never smoked, 15 drinks weekly, has hypertension, is a Canadian Citizen, etc) (e.g. q_x, where $x = 43$).
2. The probability of death is recalculated for incremental older ages (additional rows of the life table) up to age 90 years ($q_{(x+1)}, q_{(x+2)} ... q_{90}$). For each life table row, age is the only variable that is changed for MPoRT risk calculation.

Not sure what Doug is trying to get at here…

Note: Life expectancy calculation is for the respondent's age (eg. 43). To calculate age of death, add the respondent's age to their life expectancy estimate.

Note: By row life expectancy does not account for the individual's weight, and therefore can not be used to estimate the summary life expectancy.

6.5 Scenario: Intervention

How would life expectancy change if everyone increased their physical activity levels by 10%?

The health intervention scenario allows you to predict how changing the health behaviours: alcohol consumption, diet, physical activity, and smoking, of a population will affect the population health outcome (e.g. life expectancy). This feature can be used to predict the effectiveness of proposed policies or programs.

There are three types of scenarios: **absolute, relative,** and **target**.

- **Absolute:** each individual in the population **changes** their health behaviour **by a value of x**.
- **Relative:** each individual in the population **changes** their health behaviour **by a ratio of y**.
- **Target:** each individual in the population **has a set value of z.**

For target scenarios if the individual's value is already at the target value or beyond the target value then their value is not changed. E.g. If the target value for physical activity is 2.5 METs/day, then any individual that already has METs/day >= 2.5 METs/week then there value will not be adjusted.

The changes for each type of scenario for **alcohol, physical activity,** and **diet** are described in the following table:

Health.Behaviour

Absolute.change

Relative.change

Target

Each individual changes …

Each individual changes …

Each individual has the value …

Alcohol Consumption

the number of drinks they have per week by x

the number of drinks they have per week by y %

z drinks per week

Physical Activity

their physical activity level by x METs per day

their physical activity level by y % METs per day

z METs per day

Diet

the number of servings of fruits and vegetables they eat by x per day

the number servings of fruits and vegetables they eat by y % per day

Rel, Target cropped.bb

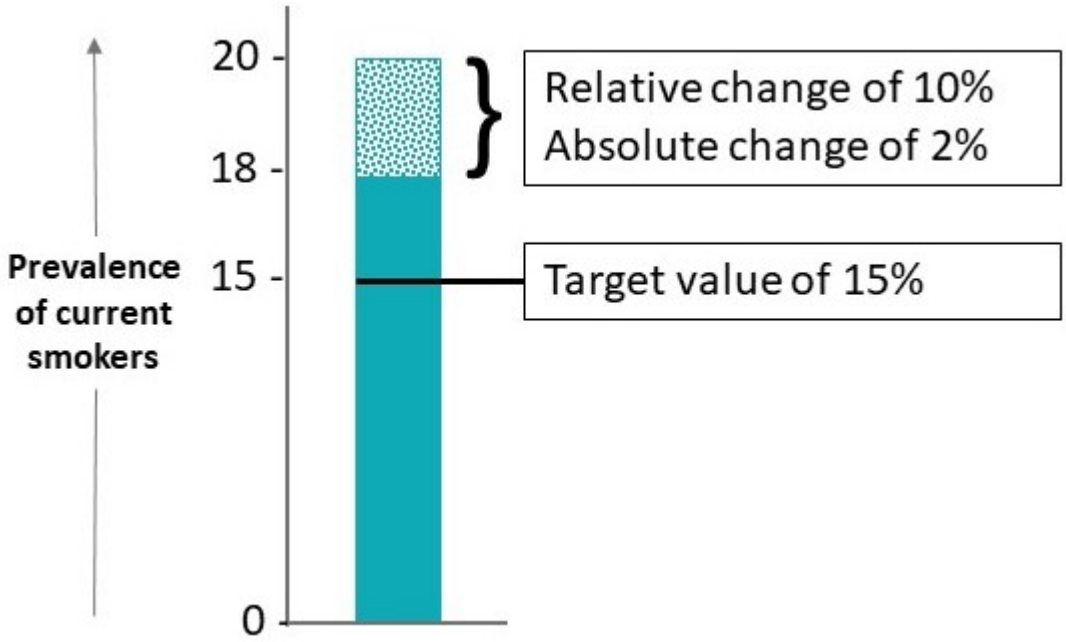


Figure 6.1: Comparison of health intervention scenario types

z fruits and vegetables per day

the number of glasses of juice they drink by x per week

the number of glasses of juice by y % per day

z glasses of juice per day

the number of servings of potatoes the eat by x per week

the number of potatoes they eat by y % per day

z potatoes per day

The **smoking** health intervention scenario is different then the other types of health intervention scenarios as they adjust the prevalence of the health behaviour.

Health.Behaviour

Absolute.change

Relative.change

Target

Smoking

The prevalence of smokers decreases by x %

The prevalence of smokers decreases by y %

The prevalence of smokers is z %

Although each type of health intervention for smoking: absolute, relative and target, changes the prevalence of current smokers they are different. The following figure shows how each is different from one another.

To adjust the prevalence of smokers, the change is applied to every current smoker in the population; individuals are not selected at random. ….

6.6 Scenario: Cause-deleted

What would be the life expectancy of a population if be no one in the population ever smoked? This scenario is a cause-deleted scenario.

There are two distinct terms in cause-deleted calculations:

- **Cause-deleted life expectancy** is the estimated life expectancy for the counterfactual population where a specific cause (e.g. smoking) never existed. For instance, everyone in the counterfactual population were always never smokers.
- **Cause-deleted effect on life expectancy** or **life years lost due to the cause** is the full effect of the cause (e.g. smoking) in the population. For instance, 3 years of life are lost in a population due to individuals smoking: either currently or previously.

If multiple causes are selected (e.g. smoking and physical inactivity) the:

- cause-deleted life expectancy calculated accounts for both of these effects
- cause-deleted effect on life expectancy (or life years lost due to the cause) is calculated for each individual cause. There will be both the cause-deleted effect of smoking on life expectancy and cause-deleted effect of physical inactivity on life expectancy.

Note: Cause-deleted life expectancy of smoking and physical inactivity $\neq$ Cause-deleted effect of smoking+ Cause-deleted effect of physical inactivity

This is because individuals in the population may be both smokers and physically inactive.

6.6.1 Cause-deleted references

The following table describes the reference exposures used in the counterfactual populations for each of the health behaviours in the cause-deleted calculations:

Health.Behaviour

Reference

Smoking

Never smoker

Alcohol Consumption

0 drinks/week

Physical Activity

3.0 METs/day

Diet

A total of: 5 fruit and vegetables, 0 juice, and 0 potato, servings/day

6.6.2 Cause-deleted calculation overview

The cause-deleted

6.7 Assumptions and Limitations

6.7.1 Limitation: Included population

The CCHS survey used to develop the mortality algorithm in the Project Big Life Planning Tool, include only individuals living in the community setting. Therefore individuals that live: in long-term care facilities, First Nations reserves, and full-time members of the Canadian Forces were excluded.

6.7.2 Limitation: Under-reporting of health behaviours

The algorithms used in the Project Big Life Planning Tool use population healht surveys with self-reported data to predict risk and health outcomes. With self-reported surveys there is the possiblitiy of social desirability bias, where respondents over-report what they percieve to be healthy behaviours and under-report what they percieve to be unhealthy behaviours.

This is particularly the case with alcohol, where there is a consistent under-reporting of alcohol consumption in most population health surveys. In Ontario, the sum of self-reported alcohol consumption is about half the volume of alcohol sold {Rehm J (2006)}.

In addition the full spectrum of the behaviour has not been captured due to limited questions about the health behaviour in surveys. With diet there were few questions on the CCHS survey and the questions did not capture diet other diet factors like sodium intake, trans fat, or other healthy/unhealthy behaviours. It is likely that the burden of poor diet is under-estimated.

6.7.3 Limitation: Use of the tool

Can the tool be used for historic analysis e.g. OPH proposed scenario #2

Chapter 7

Glossary

5-year mortality risk

The probability that an individual will die in the next 5 years.

Body Mass Index (BMI)

A weight-to-height ratio used as an indicator of obesity and underweight. BMI is calculated by dividing an individual's body weight in kilograms by the square of height in metres (kg/m2).

Burden

The impact or size of a health problem in an area, measured by cost, mortality, morbidity or other indicators. The burden of unhealthy behaviour is calculated by the differences in life expectancy based on individuals' exposure to four health behavioural risks for poor health relative to the healthy category.

By Row Measures

When selected, the result will be the measurement for each individual (e.g. row) in the dataset.

Calibration

The agreement between predicted risk generated from the model and observed risk generated from the data.

Canadian Community Health Survey

An annual survey preformed by Statistics Canada that collects information related to health status, health care utilization and health determinants for the Canadian population. Details about the survey can be found on Statistic Canada website (http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=144170).

Cause-deleted life expectancy

A cause-deleted health outcome is the estimated health outcome of a population if a specific cause (e.g. smoking) did not exist in that population.

Discrimination

The ability of the model to differentiate between high risk individuals and low risk individuals.

Error Message

Error messages will occur when variables that are **"Required for Calculation"** are missing in the data. If the entire column for the variable is missing then the calculation cannot be performed on the data. If there are missing row entries for the variable then the entire row will not be used in the calculation.

Filter

Chooses part of your dataset for analysis. If you filter on 'Sex' and then 'Male', calculations will only be performed on individuals that are 'Male' and 'Females' will be excluded. For example, when calculating Life Expectancy on the filter variable 'Sex' then 'Male' there will be a Life Expectancy estimate for 'Males' and *no* Life Expectancy estimate for 'Females'.

Health Behaviour

Actions people do that may affect their health, positively or negatively. Health behaviours are among the determinants of health and are influenced by the social, cultural and physical environments in which people live and work.(Statistics Canada, 2010) They are also shaped by individual choices and external constraints.(Statistics Canada, 2010) The four health behaviours of **smoking, alcohol consumption, diet,** and **physical activity** are specified in Project Big Life's planning tool.

Ignored Variables

Are not included in the calculation. It does not matter if your dataset includes these variables or not. Ignored variables can used for filter and stratification.

Life Expectancy (LE)

Life expectancy is a calculation of how long a person or population would be expected to live, on average, given unchanging risk of death from a specific point in time.

Metabolic Equivalent of Task (MET)

The metabolic equivalent of task (MET) is a measure of the rate of energy expenditure from an activity; a measure of calories burned by type, duration and frequency of physical activity. The reference value of 1 MET is defined as the energy expediture rate at rest which is equal to 1kcal/kg/day.

Predictor

A variable that is used in the algorithm to predict the outcome.

Recommend for calculation

Variables that are included in the calculation but not necessary for the calculation to run. Rather these variables increase the accuracy of the results.

Required for calculation

Variables that are included in the calculations and are necessary for the calculation to run. If a dataset does not have these variables then the calculation will not run.

Risk

The probability ofa health event occuring at some point of time in the future.

Socioeconomic Position

People in poorer socioeconomic circumstances generally have poorer health. Deprivation measures identify those who experience material or social disadvantage compared to others in their community. The Deprivation Index for Health in Canada developed by the Institut national desanté publique du Québec (INSPQ)(Pampalon R, Raymond G, 2000) is used in this plannning tool. The index includes education, employment and income as measures of material deprivation; and single-parent families, living alone, or being divorced, widowed or separated as measures of social deprivation. The deprivation index was used to assign geographical areas into socioeconomic position groups (low, middle and high) based on material and social quintiles. High-deprivation neighbourhoods were those in the top two quintiles for both social and material deprivation. Low-deprivation neighbourhoods were those in the bottom two quintiles.

Stratification

The seperation of data into smaller strata (levels or classes which individuals are assigned too). If the variable 'Sex' is stratified it creates two strata: 'Male' and 'Female'. Calculations are performed on each strata (level or class) and the outcome will be specific to that strata. For example, when calculating Life Expectancy

on the stratified variable 'Sex' there will be a Life Expectancy estimate for 'Males' and a different Life Expectancy estimate for 'Females'.

Stratification can only occur on categorical variables.

Summary Measures

When selected, the result will be a single measure for the entire dataset. For instance when Summary Measure - Life Expectancy (Summary Measure) is selected the result is a single life expectancy at birth for the given for the population.

When stratifications are selected, the summary measure will be given for each strata. For instance when Summary Measure - Life Expectancy and Stratification - Sex are selected, then two life expectancy measures will be given one for males and one for females.

Warning Message

Warning messages will occur when variables that are **"Recommended for Calculation"** are missing in the data. If the entire column for the variable is missing the calculation will still be performed on the data. If there are missing row entries for the variable the row will still be used in the calculation

Weights

A weight is given to each respondent in the survey and the weight corresponds to the number of individuals in the population the respondent represents.

Appendix A

Mortality Population Risk Tool (MPoRT)

A.1 Overview

Outcomes: 5-yr risk of death, Life Expectancy, Cause-deleted Life Expectancy

Calculations

Using MPoRT you are able to calculate:

- 5 year mortality risk
- Number of deaths
- Life Expectancy (individual and population)
- Cause-deleted deaths and life expectancy
- Burden of health behaviour in deaths and on life expectancy

Types of Questions

- What is the burden of smoking on life expectancy?
- How many deaths would be prevented if everyone met their daily exercise requirements?

Description

A multivariable predictive risk model that estimates the future risk of all-cause death in Canada. It adjusts for health behaviours: smoking, unhealthy alcohol consumption, poor diet, and physical inactivity, and a wide range of other risk factors.

A.2 MPoRT version description

Versions of MPoRT have been developed since 2012 and used in various studies. Each version of MPoRT (v1.0, v1.2, v2.0) used the Ontario subset of the Canadian Community Health Survey (CCHS) for development and the survey respondents were linked to personal death records. In later versions of MPoRT (v1.2, v2.0) the following changes were made:, (a) algorithm variables were adjusted to improve predictions, and (b) the

algorithms were validated using: the Ontario subset of CCHS of the years that were not used in development and the National CCHS data set (excluding Ontario).

MPoRTv1.0

Was used in the "Seven More Years" report, a joint report with Public Health Ontario and IC/ES (https://www.ices.on.ca/Publications/Atlases-and-Reports/2012/Seven-More-Years). In summary, the algorithm estimated the risk of death associated with health behaviours: smoking, unhealthy alcohol consumption, poor diet, physical inactivity and stress. There were approximately 550,000 person-years of follow up and over 6000 deaths in the development data set. The algorithm used categorical predictor variables for health behaviours and sociodemographic factors.

MPoRTv1.2

Was published in PLoS (https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002082). In summary, the algorithm estimated the risk of death associated with health behaviours: smoking, unhealthy alcohol consumption, poor diet, and physical inactivity (stress was removed due to its low prediction ability). There were approximately 1 million person-years of follow up and over 9000 deaths in the development and validation data sets. The algorithm used multiple continuous predictor variables, and added chronic disease predictor variables and interaction terms.

MPoRTv2.0 - Version in Project Big Life's Planning Tool

This version of MPoRT has not yet been published.

Development: This predictive risk model was developed using Ontario subsets of the 2001 to 2008 CCHS and participants were linked to personal health records. There were approximately 1.3 million person-years of follow-up and over 15,000 deaths in the developmental data set.

Validation: This predictive risk model was validated using three different data sets: Ontario subset of the 2009 to 2012 CCHS, National data set (except Ontario) of the 2003 to 2008 CCHS, and the National data set of the 2000 and 2005 National Health Interview Survey in the United States of America. In all validation data sets individuals were linked to personal health records.

Model: Two MPoRTs have been created: one for males and one for females. Each model is a cox-proportional hazard model that looks similar to:

$$\text{Mortality risk} = \sum_t h_0(t) * e^{\beta_{pred.smoking} * x_{smoking} + \beta_{pred.cancer} * x_{cancer} + \beta_{pred.age} * x_{age} + \ldots}$$

Where:

- t = survival time
- $h_0(t)$ = the baseline hazard
- β_{pred} = predictive hazard ratios for the exposures
- x = the exposure. The exposure can be continuous (e.g. age) or categorical (e.g. smoking status).

Parameters: The parameters (betas and exposures) used in this multivariable predictive risk model are:

Category

Variable

Scale

Description

Demographic

Age*

Continous

5 knot spline. Valid range 20 to 102

Sex

Dichotomous

Stratified Female/Male

Health Behaviour

Pack years of smoking

Continous

3 knot spline. Valid range: 0 to 78 (Female), 0 to 112.5 (Male)

Smoking Status

Categorical

Non-smoker

Current Smoker

Former Smoker <= 5 years

Former > 5 years

Alcohol (number of drinks per week)

Continous

4 knot spline (Females) and 3 knot spline (Males). Valid range: 0 to 25 (Female), 0 to 50 (Male)

Former/non-drinker

Dichotomous

Yes/No

Simplified diet score

Continous

3 knot spline. Valid range: -18.9 to 20.7 (Female), -16.8 to 18.4 (Male)

Leisure physical activity (MET)

Continous

3 knot spline. Valid range: 0 to 12.4 (Female), 0 to 16 (Male)

Socio-demographic

Ethnicity

Categorical

White

Black

Chinese

Arab; South Asian; West Asian

Filipino; Japanese; Korean; Southeach Asian

Other; Indigenous; Latin American; Multiple origin; unknown

Immigrant

Dichotomous

Yes/No

Fraction of lifetime in Canada

Continous

3 knot spline†. Valid range: 0 to 1

Education

Categorical

Less than secondary

Secondary School Graduation

Some Post-Secondary

Post-Secondary Graduation

Neighbourhood social and material deprivation

Ordinal

Low (1st or 2nd quantile

High (4th or 5th quantile)

Moderate (all others)

Chronic Conditions

Diabetes

Dichotomous

Yes/No

High Blood Pressure

Dichotomous

Yes/No

Chronic Respiratory Disease

Dichotomous

Yes/No

Mood Disorder

Dichotomous

Yes/No

Cancer

Dichotomous

Yes/No

Dementia

Dichotomous

Yes/No

Heart Disease

Dichotomous

Yes/No

Stroke

Dichotomous

Yes/No

Epilepsy

Dichotomous

Yes/No‡

BMI

Continous

3 knot spline. Valid range: 8.9 to 47.2 (Female), 8.6 to 43.7 (Male)

* Age interaction included for all variables exept immigrant, fraction of time in Canada, and ethnicity † Excluded in the male model, remains in the female model ‡ Excluded in the female model, remains in the male model

A.2.1 Derived parameters

In MPoRTv2 there are some parameters that are derived from varialbes within the data set. These include: pack-years of smoking, and the simplified diet score.

Pack-years of smoking

Pack-years of smoking is generated from: the type of smoker, age in which they started smoking daily, how many cigarettes they smoke daily, and if applicable the age in which they stopped smoking daily. Pack-years of smoking also includes: their age, the age of thier first cigarette, and whether throughout thier life they have smoked 100+ cigarettes.

In summary, the more an individual smokes or the longer they smoke, the greater the pack-years of smoking.

Simplified Diet Score

The simplified diet score essentially adds the healthy diet variables together (daily servings of carrots, fruit, salad, vegetables) and subtracts the unhealthy diet variables (daily servings of juice and potato).

Appendix B

Cause-deleted calculations

This section will explain the math behind the Project Big Life Planning Tool cause-deleted calculations and additional considerations.

B.1 Calculation

There are two parts to the calculations preformed in cause-deleted scenarios: (A) calculate the risk, and (B) calculate the health outcome: life expectancy or number of deaths.

B.1.1 Part A: Risk calculations

The original multivariable predictive risk algorithm is:

$$\text{Risk} = \sum_t h_0(t) * e^{\beta_{pred.smoking} * x_{smoking} + \beta_{pred.cancer} * x_{cancer} + \beta_{pred.age} * x_{age} + \ldots}$$

Step 1. Modify the original algorithm to include the external coefficient(s). This means replacing all predictive hazard ratios/betas related to the health behaviour to the causal hazard ratios/betas.

- Remove the original regression coefficient(s) for the health behaviour.
- Add the new external coefficient(s) to the algorithm. External coefficients are generated from either: causal models, or from systematic reviews or meta-analysis.

$$\text{External coefficient risk} = \sum_t h_0(t) * e^{\beta_{\textbf{causal.smoking}} * x_{smoking} + \beta_{\textbf{causal.cancer}} * x_{cancer} + \beta_{pred.age} * x_{age} + \ldots}$$

Step 2. Risk is calculated using the modified algorithm created in Step 1 and the respondent's original profile (e.g. current smoker). This is the "external coefficient risk".

$$\text{External coefficient risk} = \sum_t h_0(t) * e^{\beta_{causal.smoking} * (\textbf{current smoker}) + \beta_{causal.cancer} * x_{cancer} + \beta_{pred.age} * x_{age} + \ldots}$$

Step 3. "Cause-deleted risk"' is calculated by setting an exposure to a reference (non-exposed) value (all other risk exposures remain unchanged).

only -cbf.bb

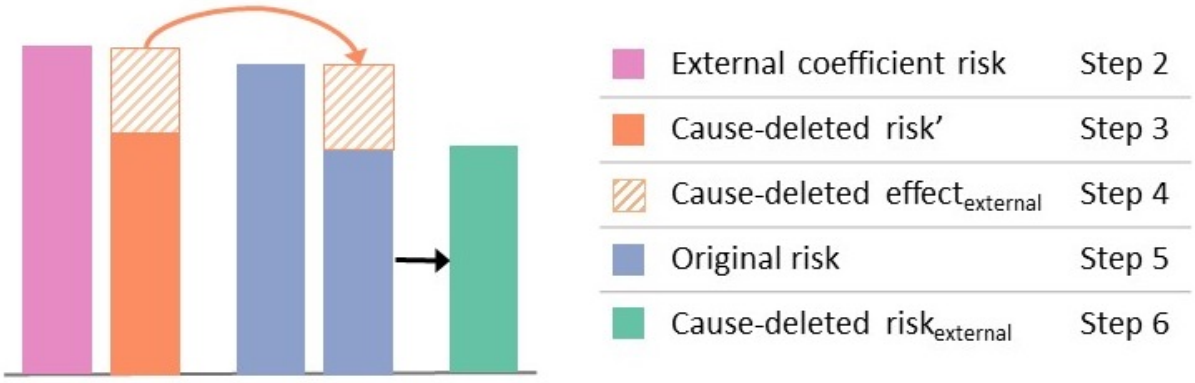


Figure B.1: Risk portion of the cause-deleted calculations

$$\text{Cause-deleted risk'} = \sum_t h_0(t) * e^{\beta_{causal.smoking} * (\textbf{never smoker}) + \beta_{causal.cancer} * x_{cancer} + \beta_{pred.age} * x_{age} + \ldots}$$

Step 4. The "cause-deleted effect external" is calculated as "external coefficient risk" (Step 2) minus the "cause-deleted risk'."(Step 3).

$$\text{Cause-deleted effect}_{external} = \text{External coefficient risk} - \text{Cause-deleted risk'}$$

Step 5. Original risk is calculated using the original algorithm and the original respondent's profile.

$$\text{Original risk} = \sum_t h_0(t) * e^{\beta_{\textbf{pred.smoking}} * (\textbf{current smoker}) + \beta_{\textbf{pred.cancer}} * x_{cancer} + \beta_{pred.age} * x_{age} + \ldots}$$

Step 6. The "cause-deleted risk external" is calculated by "original risk" (Step 5) minus the "cause-deleted effect external" (Step 4).

$$\text{Cause-deleted risk}_{external} = \text{Original risk} - \text{Cause-deleted effect}_{external}$$

B.1.2 Part B: Health outcome calculations

Using risks generated above you can then calculate:

- cause-deleted life expectancy or life years lost attributable to a health behaviour (exposure)
- cause-deleted number of deaths or number of deaths attributable to a health behaviour (exposure)

Life expectancy calculations

Step I: Calculate the original life expectancy by using the original risk (Step 5 above) in the sex-specific 5-year abridge period life tables.

Step II: Calculate the cause-deleted life expectancy by using the cause-deleted risk external (Step 6 above) in the sex-specific 5-year abridge period life tables.

Step III: Calculate life years attributable to a health behaviour by: original life expectancy (Step I) minus cause-deleted life expectancy (Step II):

$$\text{Life years due to exposure} = \text{Original life expectancy} - \text{Cause-deleted life expectancy}$$

If the life years attributable to a health behaviour are negative then the value represents the life years lost due to smoking.

Number of deaths calculations

Step I: Calculate the number of deaths that would occur using the original risk (Step 5 above).

Step II: Calculate the number of deaths that would occur using the cause-deleted risk external (Step 6 above).

Step III: Calculate the number of deaths that are attributable to a health behaviour (exposure) by: original number of deaths (Step I) minus cause-deleted number of deaths (Step II):

$$\text{Deaths due to exposure} = \text{Original number of deaths} - \text{Cause-deleted number of deaths}$$

B.2 Additional considerations

B.2.1 Exposures with non-monotonic relationships

Exposures (risk factors) with relationships that are non-monotonic (always increasing or always decreasing) can be used in cause-deleted calculations but special consideration may be warranted for both policy and analytic reasons.

An example of exposure with a non-monotonic relationship is alcohol. Some studies suggest that there is a "J" shaped risk relationship for outcomes such as cardiovascular disease.

Is this still a consideration if we are using 0 as the reference value? Need to complete the thought at the end of the paragraph Alcohol drinking guidelines usually do not recommend people: non-drinkers or former-drinkers, initiate drinking. In this situation, the target population for cause-deleted calculations can be restricted to for respondents with moderate or higher drinking, or multiple reference exposures could be created…

B.2.2 Exposures with interactions or multiple coefficients

The and cause-effect and cause-deleted risk estimates can be calculated for risk factors with interaction terms, including complex interactions or multiple coefficients. Examples include:

- risk factors with age interaction;
- spline functions with terms for each knot point; or,
- composite risks such as smoking which may coefficients for smoking status (current, former, never) and pack-years.

All coefficients that related to a common risk factor should be simultaneously considered.

B.3 Exposures not in the original algorithm

Need to edit the text The cause-effect and cause-deleted risk estimates can be calculated for risk factors that are not in the original algorithm. Method 2 must be used. Coefficient(s) are added to the model. Following cause-effect is calculated in the same way as other external risk factors. There is an assumption that external risks do not change or influence the risk estimate when cause-effect and cause-deleted risk

estimates are calculated using the approach. This means that there is no expectation that the external risk changes discrimination or calibration. If external risk is expected to change risk prediction, then the external coefficient can first be added to the algorithm. Following and Method 1 is used to estimate cause-effect and cause-deleted risk. If possible, the new coefficient should be centered when adding to the algorithm to reduce the likelihood of poor calibration (bias or prediction that over estimates risk). This approach can be considered when the external risk in an independent risk with effect that is not correlated with other coefficients.

Bibliography

Manuel D, e. a. (2012). Predictive risk algorithms in a population setting: an overview. *Journal of Epidemiology & Community Health*, 66(10):859–865.

Pampalon R, Raymond G (2000). A deprivation index for health and welfare planning in quebec. *Chronic Dis Can*, 21(3):104–13.

Rehm J, e. a. (2006). Alcohol-attributable mortality and potential years of life lost in canada 2001: implications for prevention and policy. *Addiction*, 101:373–84.

Statistics Canada (2010). Healthy people, healthy places. Technical report, Statistics Canada.