Proportional multi-state multiple-cohort life table model

Belen Zapata-Diomedi and Ali Abbas

26 March 2018

Contents

1 Introduction

The proportional multi-state multiple-cohort life table model (PMSLT) is a population level model (macro) approach to simulate health (and economic) implications of changes in exposure to health risk factors (e.g. physical inactivity, air pollution and diet). The PMSLT has been widely used to simulate outcomes for population level interventions for the reduction of chronic diseases.

The model was developed by Jan Barendregt and colleagues and has been widely used in Australia and New Zealand (T. Vos et al. 2010; Blakely et al. 2015).

The basic infrastructure of the model consist of three components: (1) Effect size for the intervention of interest (e.g. intervention to urban design that modifies population levels of physical activity); (2) Calculation of the potential impact fraction (PIF) to derive the change in occurrence of disease (incidence rate/mortality rate) attributable to a change in the distribution of the risk factor (e.g. physical activity); and (3) Use of the PMSLT to simulate health (and economic) outcomes attributable to a change in the distribution of health risk factor/s in the population of interest. Figure 1 summaries the basic infrastructure of the model. ITHIM is included in Figure 1 to show that both approaches share in common steps one and two and differ in the mechanisms of calculation of change in health burden.

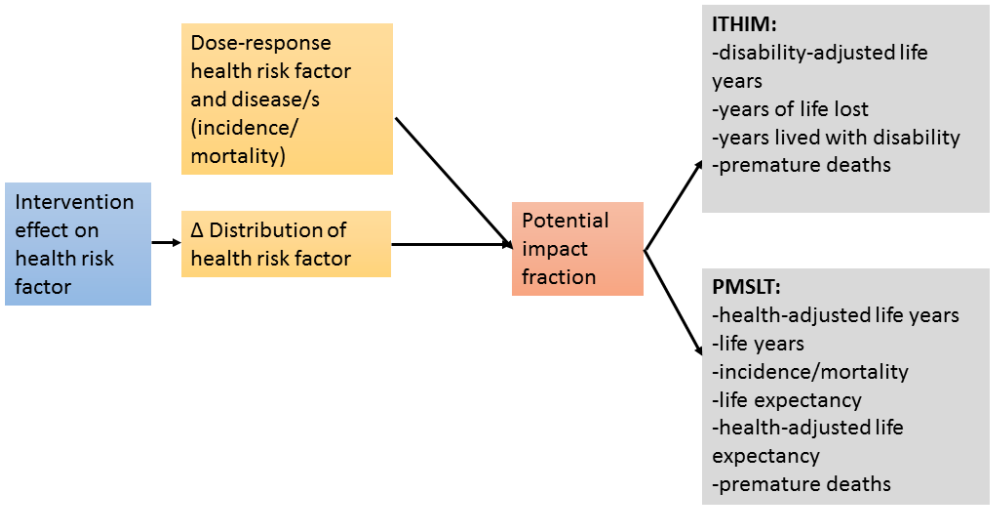


Figure 1: Basic ITHIMR infrastructure

HALYs, QALYs and DALYs

In this model we use the term 'health-adjusted life year' (HALY). As 'summary measure of population health' it measures both quantity and quality of life, where one HALY represent the equivalent of one year in full health (which could be two years with a quality of life of 0.5, for example). Specific types of HALY are the quality-adjusted life year (QALY) and the disability-adjusted life year (DALY). The QALY derives from economics and was first used in the 1960s as a measure of health gain (Gold, Stevenson, and Fryback 2002). The disability-adjusted life-year (DALY) was developed for use in burden of disease studies as a measure of health loss due to disease (Gold, Stevenson, and Fryback 2002). Our calculated HALYs are neither QALYs not DALYs, but something in between. They are similar to QALYs in that they represent health gains. However, the main difference is in the calculation of the health-related quality of life component. QALYs use measures of utility weights that traditionally represent individual experiences of health, whereas our estimated HALYs use disability weights linked to specific diseases, which were developed for the Global Burden of Disease study (Gold, Stevenson, and Fryback 2002). As discussed in past research (L. Cobiac, Vos, and Barendregt 2009; Roux, Pratt, and Tengs 2008) the main advantage of using disability weights over utility weights is that disability weights refer to specific diseases rather than health states. We opted to use the more general terms HALYs given that the use of the DALYs terminology may lead to think that our calculations are similar to those in burden of diseases studies (Murray et al. 2012). In our study, our model does not explicitly separate years of life lost (YLL) and years lived with disability (YLD) components, but instead calculates the total number of life years lived, adjusted for the average health-related quality of life in those years (by age and sex). In burden of disease studies, DALYs are defined as the sum Years of Life Lost (YLL) and Years Lived with Disability (YLD).

1.1 Contribution to ITHIMR

The PMSLT similar to ITHIM is a comparative risk assessment approach (Briggs, Scarborough, and Smith 2016) that consist of calculating the change in the health burden for a population of interest from a change in exposure to health risks factors (e.g. physical inactivity, air pollution and road trauma).As depicted in Figure 1, both methods need estimates of the potential impact fraction (PIF), which indicates the proportion of the disease burden attributable to a risk factor of interest (e.g. physical inactivity) (Barendregt and Veerman 2010). A step further back, is the development of scenarios that bring about change in the distribution of the risk factor of interest. For now, we only focus on calculations from the PIF onward, and provide a hypothetical example of change in population levels of physical activity. Incorporation of additional health risk factor (air pollution, road trauma, NO2 and noise) will be discussed in the relevant code sections.

1.1.1 Difference between ITHIM and PMSLT

- **Time component** The *PMSLT* follows a population of interest over time. For example, as set up here, we simulate sex and age (5 years starting at 18) cohorts over time until they die or reach 100 years of age. This implies that we can include trends for diseases, time lags between change in exposure to risk factors and change in health and demographic changes (e.g. population growth). In addition, we can estimate yearly changes in the burden of diseases over the life course or for a specified number of years. The *ITHIM* approach is a snapshot of change in burden for one year.
- **Interaction between multiple diseases** The *PMSLT* accounts for the interaction between multiple diseases, with proportions of the population being able to be in more than one health state (Briggs, Scarborough, and Smith 2016). This avoids overestimation of outcomes as a result of summing health outcomes attributable to each disease individually as done in *ITHIM*. It is important to note that the *PMSLT* assumes that diseases are independent of each other. That is to say, developing a disease is unrelated to a concurrent diagnoses of another disease).
- **Mortality rate** The *PMSLT* calculations for changes in life years (and health-adjusted life years) and mortality outcomes is based on observed mortality rates for the population of interest. In the *ITHIM* model, if burden of disease estimates from the Global Burden of Disease (GBD) study are used, then, the mortality component is based on the highest attained life expectancy observed in the world.
- **Impact of disability in increased life expectancy** In GBD studies, YLLs are not adjusted for disability; hence, their use in estimating intervention effects results in over-estimation, which the *PMSLT* approach avoids. Another way of seeing this is that estimated changes in morbidity using the *ITHIM* do not allow for how implicit increases in life expectancy impact on morbidity. While the changes in deaths and prevalence using the *PMSLT* are in some ways more accurate than those from the *ITHIM* approach it should be noted that that the average age of death and incident disease will change and thus the disease burden will be on average be shifted later in life (which is a realistic approach).

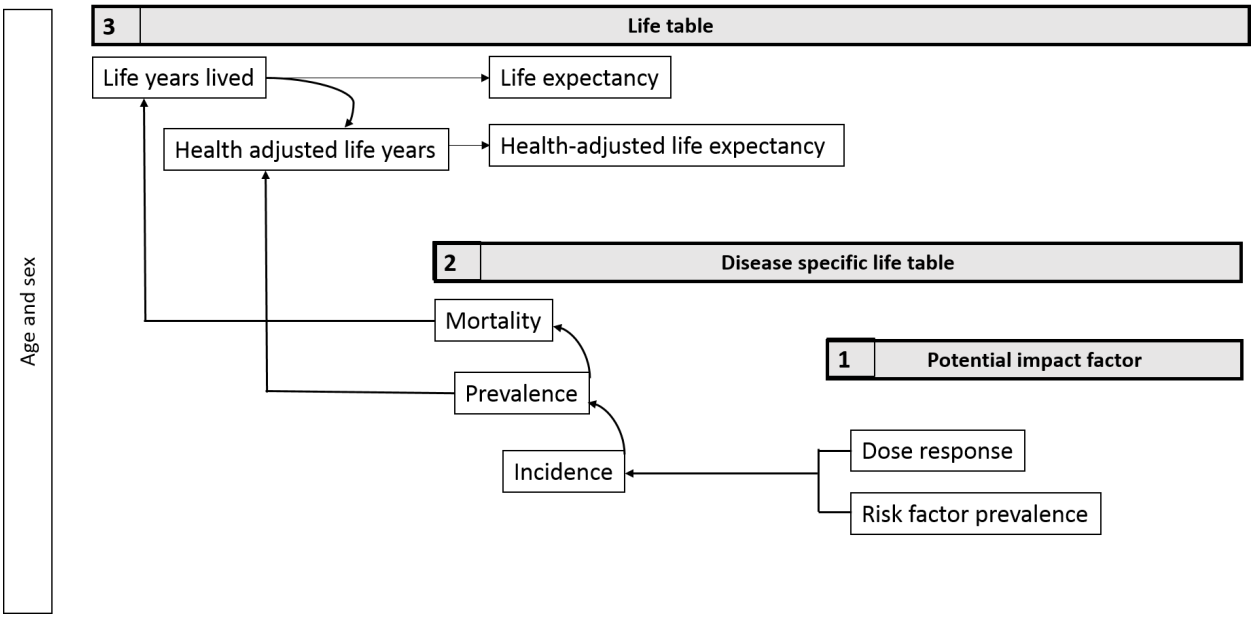


Figure 2: Proportional multi-state life-table simplified framework. *The simplied PMST shows the interaction between the life table, disease life table and potential impact fraction (PIF). The PIF calculations by age and sex group are the same as those generated for ITHIM. The PIF (or 1-PIF) modifies incidence of disease, which changes prevalence and mortality (disease specific life table). Changes in prevalence and mortality rates from the disease specific life tables feed into the life table by changing all-cause mortality, which in turn changes life years. Change in prevalence of diseases changes total years lived with disability, which in turn modifies health-adjusted life years*

2 R development

The model is set up as a long script to perform the required mathematical calculations. Where possible, we wrote functions and loops to avoid repetition. We set up the model with Australian data, for Melbourne. Figure 2 is depicts the PMSLT model framework, which was followed in the code development.

In what follows, first, we specify input parameters. Second, we present the code with explaining notes. Third, we present examples of outcomes and lastly we comment on topics related to implementation. Here we only included the physical activity health pathway. In the comments section, implementation of exposure to air pollution and road trauma is discussed. Note that in the presentation of input parameters, those needed to calculate PIFs are excluded, as these are common to the ITHIM, expect if trends are included (refer to comments section).

2.1 Inputs

We specify data requirements for the life table and disease life tables (Figure 2) and potential sources.

2.1.1 Life table

Inputs of the life table are: population numbers by sex (per 1-year or age grouping of interest), mortality rates or probability of all cause mortality by single age group and sex and total prevalent years lived with disability rate per one year by sex. Disease specific disability weights are presented as inputs here as these adjust the total years lived with disability, hence, the health-adjusted life years.

2.1.1.1 Population numbers

These data will be provided by the synthetic population. In the code presented here, we created 5-year age and sex cohorts from one-year age groups data. I left potential data sources below as a reference.

Data source: (1) National census; (2) Worldwide population and mortality data: http://www.mortality.org/ (mostly high income countries; and (3) Calculate from the Global Buden of Disease by the Institute of Health Metrics and Evaluation (GBD IHME) data (rates and numbers available from (http://ghdx.healthdata.org/gbd-results-tool).

2.1.1.2 Mortality rates

Mortality rates are needed per single year and sex. These data are available from GBD IHME, however, in age groups (1-4, 5-9, etc). Interpolation can be used to derive in between ages rates (cubic spline).

Note that we need data for population numbers and all cause mortality rates for: (1) PMSLT and (2) Dismod II collection (more in Dismod II section). Population data from the synthetic population is used for the PMSLT. For Dismod II, popuation and mortality data should be from the same source (GBD IHME)

2.1.1.3 Total years lived with disability rates per single year and sex.

These data is available from the GBD (http://ghdx.healthdata.org/gbd-results-tool) per 5-year age groups. We can use interpolation to derive between ages rates.

2.1.1.4 Disability weights (quality of life weights)

Disability weights (DW) can be derived from disease specific years lived with disability (YLD) and disease specific prevalence by age group (5 years) and sex. Data for YLDs prevalence can be obtained from the online GBD IHME data tool (http://ghdx.healthdata.org/gbd-results-tool). Our calculations of DW in the example here based on the GBD methods for estimating YLDs as the sum of sequelae prevalence multiplied by sequelae disability weights (REF GBD). The GBD has publicly available data at the cause level (e.g. ischemic heart disease) instead of squelae level (e.g. myocardial infarction, angina and heart failure). However, the GBD disability weights are for health states associated with sequelae, hence, we need to calculate DWs. An age and sex specific-correction was introduced to counteract the effects of accumulating comorbid illnesses in the older age groups (Equation 1).

$$(YLDd/Pd)/(1 - YLDt) = DWadjustedfortotalYLDs \quad (1)$$

Where YLDd is the YLD mean number per age and sex for a given disease, Pd is the prevalence (as reported in GBD) for a given disease by age and sex and YLDt is total YLD rate per age and sex.

2.1.2 Disease life tables

2.1.2.1 Incidence and case fatality

For each of the modeled diseases the PMSLT needs incidence and case fatality rates per sex and one-year intervals. Data from the GBD IHME studies with Dismod II (free at https://www.epigear.com/index_files/dismod_ii.html) can be used to derive internally consistent data and generate missing data. For example, the GBD studies provide data for incidence, prevalence and disease mortality, however, not case fatality. Other national level sources may also be explored/used, and compare with estimates produce from GBD data and Dismod II.

Dismod II inputs are: (1) population numbers and mortality rates and (2) disease specific inputs.

Population and mortality

Within Dismod II, each setting (e.g. country) has a collection that consists of population numbers (preferably the same as used in GBD IHME studies, due to the mortality envelop) and all- cause mortality rates (numbers and calculate rates). The GBD provides 5-year age groups that are acceptable input parameters for Dismod II.

Disease inputs by age group and sex

Each setting collection has a given number of diseases. Dismod II works with at least three of: case fatality, prevalence, incidence, mortality (disease), case fatality, remission, duration and the relative risk for mortality. So far, we have been assuming that remission is zero for chronic diseases, that is to say, when people become diseased, they do not recover. Special care should be taken with this assumption, as the GBD data assumes remission for some diseases, for example cancers, where after 10 years cases recover, except for long term sequelae. Since GBD now provides prevalence, incidence and mortality, it may be best to use all three as Dismod II input parameters to compare the effect of the remission assumption by the GBD for some diseases.

2.2 Code

Following the structure of Figure 2, we developed functions to perform sex and age cohorts calculations for the life table, disease life tables and potential impact fractions: run_life_table, run_disease and and run_pif. We also generated two functions for outputs: plot_outputs and gen_aggregate. The function plot_outputs creates age-group and sex linear plots for specified outcomes (e.g. health-adjusted life years, incidence of diabetes) and gen_aggregate adds up each cohort results. Functions were then used in a code script. In what follows, we explain each step in the development of the script. Here we also include code chunks, however, we also kept them separately in the MSLT folder, in the code file.

In what follows, we start with the **model** script file, and explain the **functions** script file as these are used to perform calculations. The **functions** will be explained in detailed to provide clarity for required inputs.

Table 1: PMSLT inputs

Input	Source	Comments
Life table	Synthetic population per sex and age group	Age grouping in life table to match synthetic population
Life table	Synthetic population per sex and one-year age group	If one year age group is not avabilable it can be derive using interpolation from age groups data
Life table	Global Burden of Disease (GBD) study per one-year age group and sex	GBD data is in five-year age groups, interpolation to derive one-year age groups
Disease life table	GBD data for prevalence, incidence and mortality and DISMOD II	Two step process. First obtain disease and population data from GBD. Second, use Dismod II to derive internally consistent estimates for incidence and case fatality (PMSLT disease life table iputs
Disease life table	Derive from disease prevalence and years lived with disability from GBD	Adjustments for comorbidities in later years of life to be applied

2.2.1 Set up

We start by cleaning the global environment (#1) to keep track of our works and ensure that the code is generating our outcomes. Then, we set up an option to avoid the use of scientific notation (#2) and lastly we load the functions (#3). The code chunks are shown in the rmarkdown ouput.

```
#1
rm (list = ls())
#2
options(scipen=999)
#3
source("code/functions.R")
```

2.2.2 Inputs

Table 1 describes data needs for the PMSLT, here we expand on the data needs and mechanisms (Figure 3) to use the PMSLT approach in ITHIMR (Figure 1). Figure 3 is used as a reference in the development of the code.

2.2.2.1 Data for example

The code below is to load the data used in the example provided in this code for Australia (#4). *idata* includes data by one-year age groups and sex for: population numbers, mortality rates, total years lived with disability rates, incidence, case fatality, disability weights and costs for included diseases (ischemic heart disease, ischemic stroke, diabates, breast cancer (women) and colon cancer), cost for all other diseases

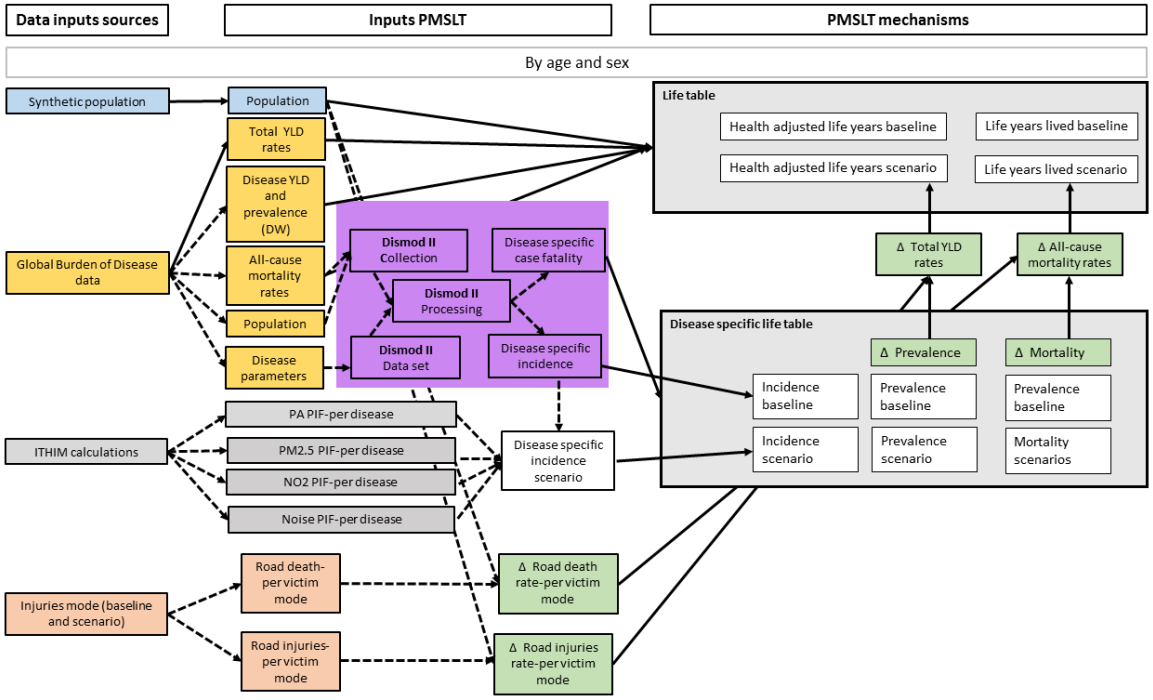


Figure 3: Proportional multi-state life-table model. *Three sections are presented in Figure 3:* ***Data input sources***, ***Inputs PMSLT*** *and* ***PMSLT mechanisms***.

(excludes those for physical inactivity) and trend for incidence and case fatality. Here we do not use cost and trends data. *edata* is for physical activity prevalence per categories (health risk factor), *irr* are relative risks and **ee** is energy expenditure per physical activity categories.

```
#4
idata <- read.csv("data/idata.csv", stringsAsFactors = F)
edata <- read.csv("data/edata.csv", stringsAsFactors = F)
irr <- read.csv("data/irr.csv", stringsAsFactors = F)
ee <- read.csv("data/ee.csv", stringsAsFactors = F)
```

2.2.2.2 Model parameters

We define parameters for our model that are used multiple times. *p_age_cohort* defines the starting age for the cohorts. *p_sex* is for the cohort sex. *p_disease* specifies the modelled diseases. *p_intervention* is a hypothetical increase in MET minutes per week.

```
#5
p_age_cohort <- c(22, 27, 32, 37, 42, 47, 52, 57, 62, 67, 72, 77, 82, 87, 92, 97)
p_sex <- c("males", "females")
p_disease <- c("ihd", "istroke", "diabetes", "colon_cancer", "breast_cancer")
p_intervention_effect <- 100
```

2.2.2.3 Data preparation

The following code applies to the data used as an exmaple here. #6 transforms upper cases to lower cases in the sex variable. #7 and #8 is used to rename mortality_rate to mx (used in life table language). #9 is used to replace male and female in the exposure data to males and females to facilitate combination of data sets.

```
#6
idata$sex <- tolower(idata$sex)
#7
idata$mx <- idata$mortality_rate
#8
idata$mortality_rate <- NULL
#9
edata$sex[edata$sex=="male"] <- "males"
edata$sex[edata$sex=="female"] <- "females"
```

```
##### Create new variable for 5_year population (age cohorts). Depends on the age-cohorts of interest (5

idata$five_year_population <- NA
start_index <- 3
index <- 1
end_index <- 4
for (i in 1:20){
  if (i == 1)
```

```
    index <- 1
  else
    index <- ((i - 1)  * 5) + 1
  if (index == 96){
    end_index <- 5
  }
  cat("population sum ", start_index, "start index ", index, " and end index ", index + end_index, "\n")
  idata$five_year_population[start_index] <- sum(idata$population[index:(index + end_index)])
  start_index <- start_index + 5
}
```

```
## population sum  3 start index  1  and end index  5
## population sum  8 start index  6  and end index  10
## population sum  13 start index  11  and end index  15
## population sum  18 start index  16  and end index  20
## population sum  23 start index  21  and end index  25
## population sum  28 start index  26  and end index  30
## population sum  33 start index  31  and end index  35
## population sum  38 start index  36  and end index  40
## population sum  43 start index  41  and end index  45
## population sum  48 start index  46  and end index  50
## population sum  53 start index  51  and end index  55
## population sum  58 start index  56  and end index  60
## population sum  63 start index  61  and end index  65
## population sum  68 start index  66  and end index  70
## population sum  73 start index  71  and end index  75
## population sum  78 start index  76  and end index  80
## population sum  83 start index  81  and end index  85
## population sum  88 start index  86  and end index  90
## population sum  93 start index  91  and end index  95
## population sum  98 start index  96  and end index  101
```

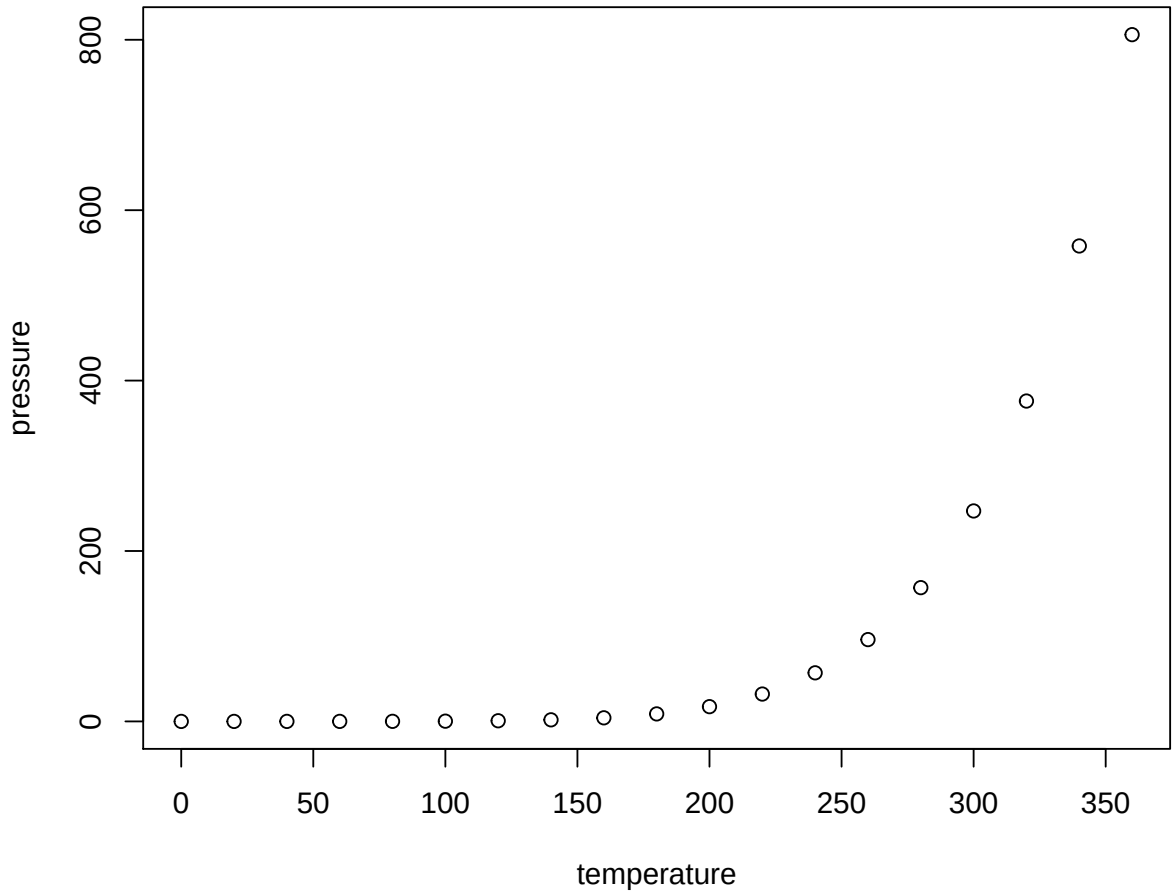
```
end_index <- 4
start_index <- 104

for (i in 1:20){
  if (i == 1)
    index <- 102
  else
    index <- (((i - 1)  * 5) + 1) + 101
  if (index == 197){
    end_index <- 5
  }
  cat("population sum ", start_index, "start index ", index, " and end index ", index + end_index, "\n")
```

```
  idata$five_year_population[start_index] <- sum(idata$population[index:(index + end_index)])
  start_index <- start_index + 5
}
```

```
## population sum  104 start index  102  and end index  106
## population sum  109 start index  107  and end index  111
## population sum  114 start index  112  and end index  116
## population sum  119 start index  117  and end index  121
## population sum  124 start index  122  and end index  126
## population sum  129 start index  127  and end index  131
## population sum  134 start index  132  and end index  136
## population sum  139 start index  137  and end index  141
## population sum  144 start index  142  and end index  146
## population sum  149 start index  147  and end index  151
## population sum  154 start index  152  and end index  156
## population sum  159 start index  157  and end index  161
## population sum  164 start index  162  and end index  166
## population sum  169 start index  167  and end index  171
## population sum  174 start index  172  and end index  176
## population sum  179 start index  177  and end index  181
## population sum  184 start index  182  and end index  186
## population sum  189 start index  187  and end index  191
## population sum  194 start index  192  and end index  196
## population sum  199 start index  197  and end index  202
```

2.3 Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

3 Comments

3.1 Road injuries in the PMsLT

The disease model used in each of the disease life table is not directly applicable to road injuries, however, similar concept can be follow. Firstly, changes in road fatalities impact on the overall mortality rate, hence, by knowing the road fatality rates for baseline and scenarios, we will be able to incorporate changes to mortality attributable to road fatalities. For road injuries, methods developed by Kavi Bhalla and Marko Tanio (REFS) that derive the average YLD attributable to life long and short term injuries can be applied to derive the change in total YLDs (CHECK THAT THESE WERE DEVELOPED AS INCIDENCE YLDs).MT's methods assumes that injuries do not reduce the life expectancy of the injured person.

References

Barendregt, J.J., and J.L. Veerman. 2010. "Categorical Versus Continuous Risk Factors and the Calculation of Potential Impact Fractions." Journal Article. *J Epidemiol Community Health* 64 (3): 209–12. doi:10.1136/jech.2009.090274.

Blakely, T., L. J. Cobiac, C. L. Cleghorn, A. L. Pearson, F. S. Deen, G. Kvizhinadze, N. Nghiem, M. McLeod, and N. Wilson. 2015. "Health, Health Inequality, and Cost Impacts of Annual Increases in Tobacco Tax: Multistate Life Table Modeling in New Zealand." Journal Article. *PLoS Med* 12. doi:10.1371/journal.pmed.1001856.

Briggs, Adam, Peter Scarborough, and Adrian Smith. 2016. "Modelling in Public Health." Book Section. In *Public Health Intelligence: Issues of Measure and Method*, edited by Krishna Regmi and Ivan Gee, 67–90. Cham: Springer International Publishing. doi:10.1007/978-3-319-28326-5_4.

Cobiac, L.J., T. Vos, and J.J. Barendregt. 2009. "Cost-Effectiveness of Interventions to Promote Physical Activity: A Modelling Study." Journal Article. *Plos Med* 6 (7): e1000110–e1000110. doi:10.1371/journal.pmed.1000110.

Gold, Marthe R., David Stevenson, and Dennis G. Fryback. 2002. "HALYs and Qalys and Dalys, Oh My: Similarities and Differences in Summary Measures of Population Health." Journal Article. *Annu Rev Public Health* 23 (1): 115–34. doi:doi:10.1146/annurev.publhealth.23.100901.140513.

Murray, Christopher J. L., Majid Ezzati, Abraham D. Flaxman, Stephen Lim, Rafael Lozano, Catherine Michaud, Mohsen Naghavi, et al. 2012. "GBD 2010: Design, Definitions, and Metrics." Journal Article. *The Lancet* 380 (9859): 2063–6. doi:10.1016/S0140-6736(12)61899-6.

Roux, L., M. Pratt, and T. O. Tengs. 2008. "Cost Effectiveness of Community-Based Physical Activity Interventions." Journal Article. *Am J Prev Med* 35. doi:10.1016/j.amepre.2008.06.040.

Vos, T., R. Carter, J. J Barendregt, Mihalopoulos C., JL. Veerman, A. Magnus, L. Cobiac, Bertram MY., and AL. Wallace. 2010. "Assessing Cost-Effectiveness in Prevention (Ace-Prevention): Final Report." Report.