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.

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).

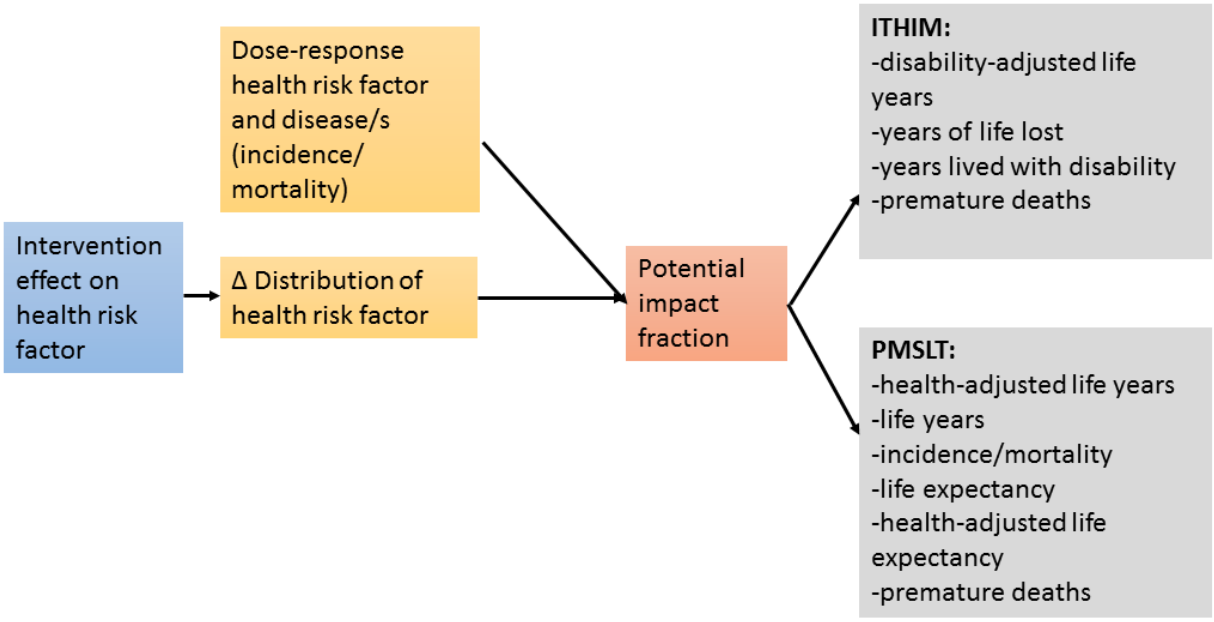

Figure 1: Basic ITHIMR infrastructure

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).

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.

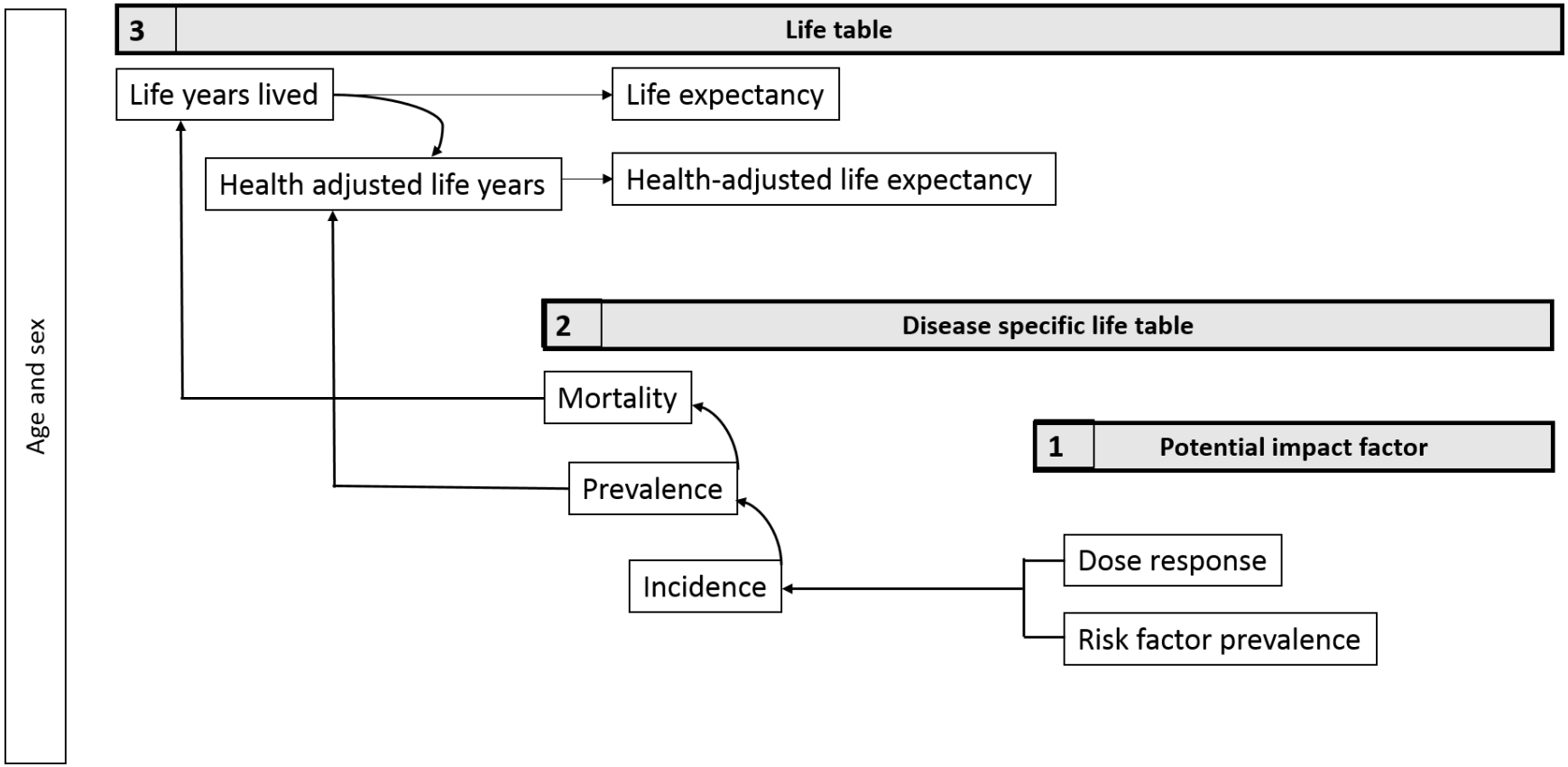

Figure 2: Proportional multi-state life-table simplified framework. *The simplied PMST shows the interaction between the life table, disease life table and potential impact fraction (PIF). The PIF calculations by age and sex group are the same as those generated for ITHIM. The PIF (or 1-PIF) modifies 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*

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 Burden 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, population 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.

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 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.

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 output.

1) Clearn Global Environment

```
rm (list = ls())
```

2) Avoid scientific notation

```
options(scipen=999)
```

3) Load functions

```
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).

Initial case studies for the ITHIMR are: London, Sao Pablo, Delhi, Accra, Los Angeles and Edinburgh. Here, we will start with **Greater London** given the availability of disease epidemiology data from the GBD IHME study. For the rest of the case study cities data is available at the country level, hence, a scaling method is needed to reflect the local burden of disease.

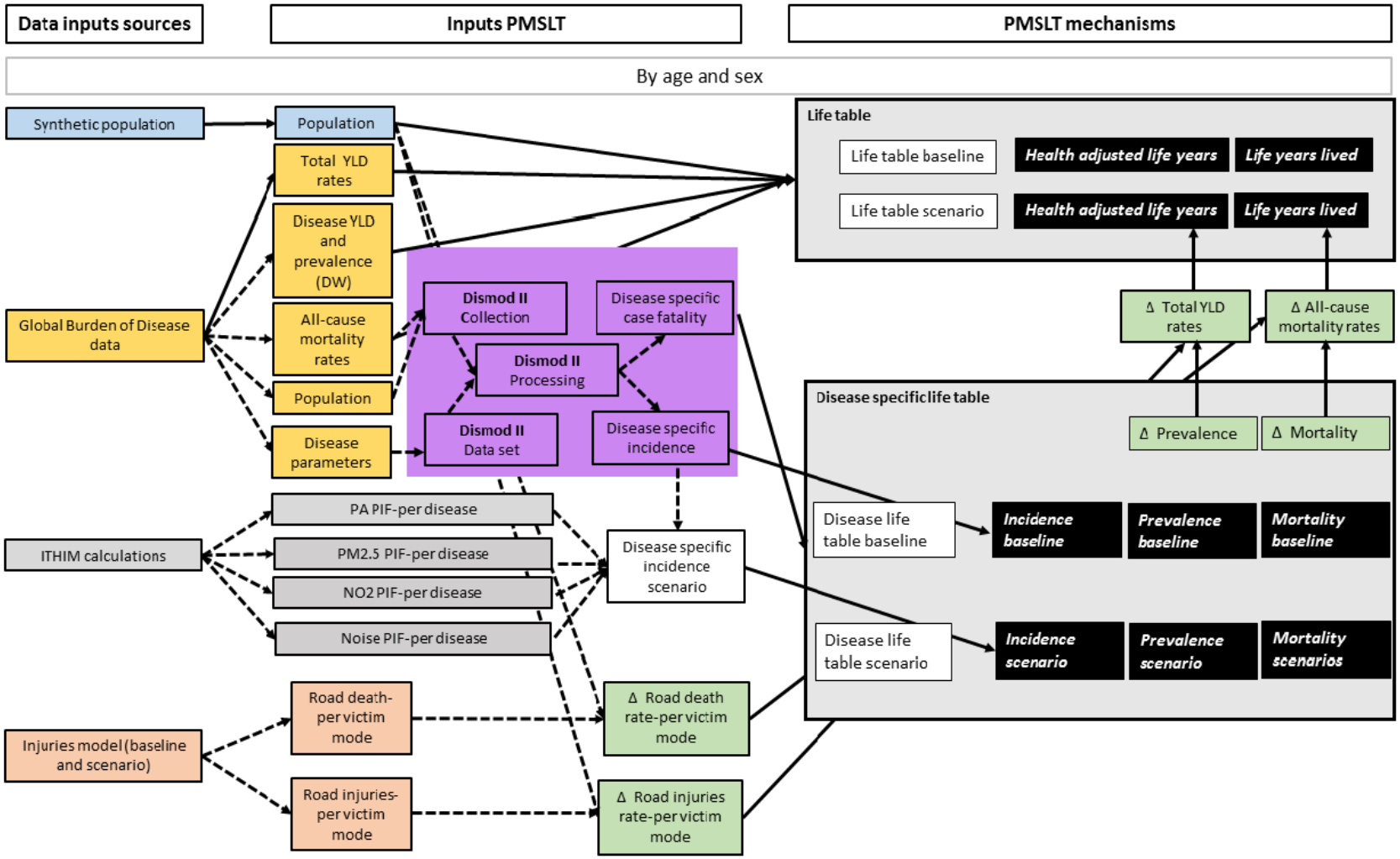

Figure 3: Proportional multi-state life-table model. *Three sections are presented in Figure 3:* ***Data input sources***, ***Inputs PMSLT*** *and* ***PMSLT mechanisms***. *The color coding from Data inputs sources to Inputs PMSLT link sources with inputs for the PMSLT. Solid arrows represent final inputs and dashed-arrows represent intermediate inputs that need further processing. Purple coding means a process and green coding represent change in mortality and disability prevalence rates to modify the life table parameters. Black color coding with white color font represent final model outcomes. For both, the life table and disease life table, two sets of each are simulated, one for the baseline and the other for the scenario.*

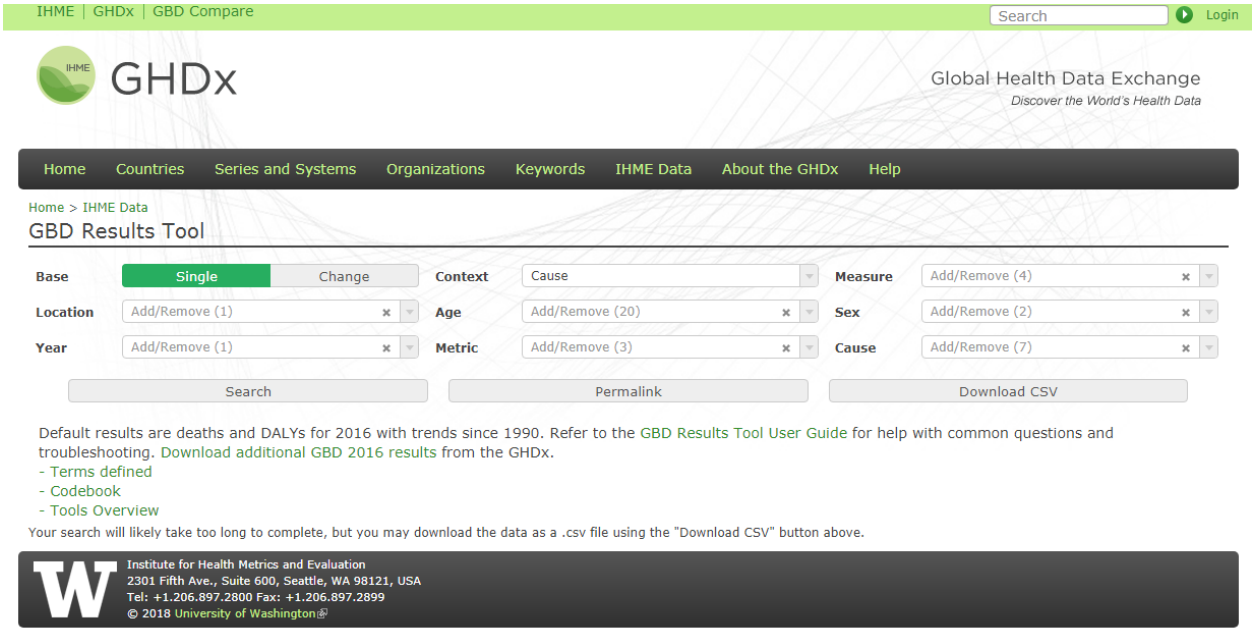

Figure 4: Global Burden of Disease data results tool.

2.2.2.1 Global Burden of Disease data

First, we explain how to obtain the data, second additional processing to derive data not reported (population) and one-year age groups (original data is in five-year age groups) and last procedure to use Dismod II. Data from the *Global Burden of Disease data* in Figure 3 can be download from here: http://ghdx.healthdata.org/gbd-results-tool. Figure 4 is a screenshot of the GHDx.

Table 2 specifies the selections to do for each of the tabs in Figure 2.

Once the selections described in Table 2 are made, the option *Download CVS** in the GHDx website is selected. A prompt comes up asking for an email address. The data is sent to the designated email address (within minutes) in ZIP format, unzip and use the code below to read the data (4). Here, we selected data for Greater London and England. The aim is to compare and derive scaling factors as for most cities the data is not available from the GBD and country level data may be used and scaled to the city level. Note that the data input requirement for the PMSLT, except population numbers, is in rates. Therefore the scaling is to better reflect the burden of an area, this is a different issue than working with numbers (e.g. total mortality numbers, total YLDs numbers) as in the ITHIM approach.

4) Read GBD data

```
GBDdata <- read.csv("data/UK/englandandgreaterlondon.csv", stringsAsFactors = F)
```

The following codes serves to sort out the GBD data to the inputs required for the life table, disease life table and Dismod II

These data should be used to generate the general life table and disease life tables (Figure 3).

5) Change all uper cases to lower case

Table 2: Global burden of disease data

Tab	Selection
Base	Single
Location	Case study city
Year	Latest available
Context	Cause
Age	Under 5, 5 to 9, 10 to 14, 15 to 19, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 49, 50 to 54, 55 to 59, 60 to 64, 65 to 69, 70 to 74, 75 to 79, 80 to 84, 89 to 89, 90 to 95, 95 plus
Metric	Number, Rate
Measure	Deaths, YLDs, Prevalence, Incidence
Sex	Male, Female
Cause	Total All causes, ischemic heart disease, etc

```
GBDdata <- mutate_all(GBDdata, funs(tolower))
```

6) Create age categories index in GBDdata (rounding mid age), total of 20 age groups. These are the age cohorts to simulate.

```
GBDdata$age_cat [GBDdata$age =="under 5"] <- 2
GBDdata$age_cat [GBDdata$age =="5 to 9"] <- 7
GBDdata$age_cat [GBDdata$age =="10 to 14"] <- 12
GBDdata$age_cat [GBDdata$age =="15 to 19"] <- 17
GBDdata$age_cat [GBDdata$age =="20 to 24"] <- 22
GBDdata$age_cat [GBDdata$age =="25 to 29"] <- 27
GBDdata$age_cat [GBDdata$age =="30 to 34"] <- 32
GBDdata$age_cat [GBDdata$age =="35 to 39"] <- 37
GBDdata$age_cat [GBDdata$age =="40 to 44"] <- 42
GBDdata$age_cat [GBDdata$age =="45 to 49"] <- 47
GBDdata$age_cat [GBDdata$age =="50 to 54"] <- 52
GBDdata$age_cat [GBDdata$age =="55 to 59"] <- 57
GBDdata$age_cat [GBDdata$age =="60 to 64"] <- 62
GBDdata$age_cat [GBDdata$age =="65 to 69"] <- 67
GBDdata$age_cat [GBDdata$age =="70 to 74"] <- 72
GBDdata$age_cat [GBDdata$age =="75 to 79"] <- 77
GBDdata$age_cat [GBDdata$age =="80 to 84"] <- 82
GBDdata$age_cat [GBDdata$age =="85 to 89"] <- 87
GBDdata$age_cat [GBDdata$age =="90 to 94"] <- 92
GBDdata$age_cat [GBDdata$age =="95 plus"] <- 97
```

7) Create age and sex categories to obtain population numbers. Population numbers from GBD are used in Dismod II. For the Life table (Figure 3), the numbers may be from the synthetic population.

```
GBDdata$sex_age_cat <- paste(GBDdata$age_cat,GBDdata$sex, sep = "_"  )
```

8) Conver string variables to numberic to do calculations.

```
GBDdata$val <- as.numeric(as.character(GBDdata$val))
```

9) Generate population numbers for England and Greater London in a new data frame ("GBD_population") and then separate in dataframes for England and Greater London (this avoids repeating the same code). Population numbers are derived from rates per 100,000 and total numbers of cases.

```
GBD_population <- filter(GBDdata, measure == "deaths", cause == "all causes", metric == "rate" | metric
```

10) Generate population numbers from given number of cases and rates per 100,000 people.

```
 for (i in 1:nrow(GBD_population)) {
    if (GBD_population$metric[i] == "number") {
      GBD_population$val_pop[i] <- GBD_population$val[i] * 100000/ GBD_population$val[i + 2]}
   else {GBD_population$val_pop[i] <- NA}
 }
```

11) Remove rows with zero

```
GBD_population <- GBD_population[!is.na(GBD_population$val_pop),]
```

12) Keep relevant variables

```
GBD_population <- filter(GBD_population) %>% select(sex_age_cat, val_pop, location)
```

13) Create data frames for Greater London and England population to be later used as inputs for PMSLT calculations

```
GBD_population_GL <-  filter(GBD_population, location == "greater london") %>% select(sex_age_cat, val_

GBD_population_England <-  filter(GBD_population, location == "england") %>% select(sex_age_cat, val_po
```

12) Check population total numbers

```
GreaterLondon <- sum(GBD_population_GL$val_pop)
England <- sum(GBD_population_England$val_pop)
```

13) Generate data frames for England and Grater London with per person rates (per 100,000 in original data).

```
GBDEngland <- filter(GBDdata, location == "england" & metric == "rate") %>% select(measure, location, s
GBDEngland$one_rate <- GBDEngland$val/100000
GBDGL <- filter(GBDdata, location == "greater london" & metric == "rate") %>% select(measure, location,
GBDGL$one_rate <- GBDGL$val/100000
```

Health outcomes included in leisure time meta-analysis: all-cause mortality, cardiovascular disease, stroke, coronary heart disease, total cancer, colon, lung, endometrial, and breast cancer.

2.2.2.2 Model parameters

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.