

▶▶
UHASSELT



Maastricht University

KNOWLEDGE IN ACTION

Faculty of Sciences
School for Information Technology

Master of Statistics

Masterthesis

Development of a generic and open source IncidencePrevalenceMortality model for the assessment of chronic disease epidemiology

Jade Vincent Membrebe

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

SUPERVISOR :

Dr. Brecht DEVLEESSCHAUWER

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.



UHASSELT

KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2016
2017



Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics

Masterthesis

Development of a generic and open source IncidencePrevalenceMortality model for the assessment of chronic disease epidemiology

Jade Vincent Membrebe

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Ziv SHKEDY

SUPERVISOR :

Dr. Brecht DEVLEESSCHAUWER

Acknowledgment

When one thinks of the epidemiology of a particular disease, one does not limit their concept to a single variable. As such, disease incidence does not completely describe the disease and this thesis was not created without help. This section was then written to give credit to those who gave significant contributions to this thesis.

The conception of this thesis is largely due to Dr. Brecht Devleesschauwer, and I am very grateful to him for letting me explore the possibilities of his ideas. Brecht, you have influenced me in more ways than one. Your passion towards this thesis topic is something that I have always found to be infectious.

I owe a lot to Prof. Dr. Ziv Shkedy. He offered me a fresh perspective on the project, which pushed me towards completing a work which I can be proud of. I have always been eager to hear his ideas on what I should and can do to make the thesis better. Ziv, I have learned a great deal from you.

My gratitude goes towards Mr. Ewoud De Troyer for teaching me programming techniques which helped me with my coding problems. Ewoud, I was truly ecstatic upon learning new skills from you. Since our first meeting, I have already installed a hundred programs in my computer upon your recommendation (fine, that was an exaggeration). But, indeed, the programs helped a lot!

This thesis would not be possible without VLIR-UOS whose financial support gave me the opportunity to finish, not just this thesis, but a master's program in Belgium. You really do change lives.

I am in awe of the kindness of Cesar, ate Chella, Jed, kuya Johan, Louraine, ate Nolen, Pia, ate Rochelle, and Zebs. They have offered me their homes when I was complaining of how I could not concentrate at my place; they provided me with internet connection when mine was not working; they answered my dumb questions about the thesis and other practical matters; and lastly, but in a way most importantly, they made sure that I ate a home-cooked meal at least once a month.

I would not have finished this thesis without Harry and Percy. They took my mind off things when I badly needed a break. They gave me brooms and blue cakes to cheer me up, and I could not have asked for anything better.

To Tatay Dodong, Nanay Susan, Kuya Peter, Kuya Bobong, Ate Suzette, JR, Lawrence, and Eurecka, you have inspired me to work better and do better.

Above all, to God, I owe you everything.

Jade Vincent Membrebe
Diepenbeek, 8 September 2017

Abstract

It is often a problem for researchers when the epidemiology of a disease is unreliable or altogether lacking. This thesis was created in order to make an R package called `DisModR` that accounts for this problem. In particular, the package was made for applications on chronic diseases.

The compartmental model used was taken from previous iterations of a software called `DisMod`. This model gives the logical relationship between different epidemiological parameters of a disease. In particular, inputs on disease incidence, remission, and case fatality were needed to determine the disease-specific mortality and prevalence. However, an iterative optimization procedure called the Nelder-Mead method was employed in order to accept more inputs. This was done through the `fminsearch()` function of the `neldermead` package. Likewise, in order to make the values of the input variables vary continuously as a function of age, cubic smoothing splines were incorporated in `DisModR`.

The `DisMod()` function in `DisModR` allows the user to compute the different epidemiological parameters of a disease. It accepts data on three of the following five inputs: incidence, remission, case fatality, mortality, and prevalence. Likewise, it allows users to manipulate options on the cubic smoothing splines and the tolerance limits of the optimization procedure. Using the breast cancer data set of the Netherlands, the properties of the `DisMod()` function was shown. Additional outputs on case fatality, prevalence, and relative risk mortality were produced using only inputs on incidence, remission, and mortality. In addition, the function also produced various objects that contained lists of input and output values, as well as their corresponding graphs.

There are several uses to `DisModR`. The first, which is perhaps the most obvious, is that it provides values for epidemiological parameters, which are otherwise unavailable such as mortality and prevalence of particular diseases. The second is that it allows users to compare its results with existing data. This is to check if the present data is consistent with the logical relationship of the other input variables. The third gives researchers the opportunity to work in the R environment, which is a unique feature of this package compared to other versions of `DisMod`. Furthermore, users are advised that they should not only rely on the R package alone, but they must combine the results with expert knowledge of the epidemiology of the disease.

Contents

Acknowledgment	ii
Abstract	iii
1 Introduction	1
1.1 Incidence-Prevalence-Mortality (IPM) Models	1
1.2 Objectives of the Study	3
1.3 Structure of the Thesis	3
2 The Epidemiology of a Disease: Data	5
3 Methodology	7
3.1 The Compartmental Model	7
3.2 Estimation of the Parameters	8
3.3 Data Smoothing	9
3.4 Accommodation of More Inputs	10
4 The R Package: DisModR	11
4.1 The Disease Modeling function	11
4.1.1 Data Inputs	11
4.1.2 Additional Options	12
4.2 Output	13
5 Analysis of the Breast Cancer Data	15
5.1 Uploading the Data	15
5.2 The Input Data	16
5.3 Cubic Spline Interpolation	17
5.4 Parameter Estimates of Disease Epidemiology	19
6 Discussion and Some Ideas for Further Research	25
6.1 Discussion	25
6.2 Further Research	26

List of Tables

2.1	Incidence, Remission, and Mortality Data for Breast Cancer in the Netherlands	6
5.1	An Exhaustive Epidemiology of Breast Cancer	21

List of Figures

1.1	The conceptual disease model	2
2.1	Total Mortality for Females in the Netherlands	6
5.1	Graph of the Input Data From DisMod R	17
5.2	Breast Cancer Data with Cubic Spline Interpolation From DisModR (default options are used)	18
5.3	Breast Cancer Data with Cubic Spline Interpolation From DisModR (<code>spar=0.4</code> was used)	19
5.4	Output variables generated with DisModR (with cubic spline interpolation)	20
5.5	Output variables generated by DisMod R (summarized into age groups)	22
5.6	Output variables generated with DisModR (no interpolation)	23

Chapter 1

Introduction

1.1 Incidence-Prevalence-Mortality (IPM) Models

The estimation of disease-specific incidence, prevalence, and mortality, specified by age and sex, are essential inputs to burden of disease calculations and cost-effectiveness analyses (Kruijshaar et al., 2002). However, gathering data on these variables is often hard. Although primary data on these variables are the obvious choice for analyses in order to draw reliable conclusions, this is not often feasible. Data on incidence, prevalence, and disease specific mortality are usually incomplete, unreliable, or altogether lacking (Barendregt et al., 2003).

There is a clear need to address the problems caused by the limitations of the data. A way to circumvent this problem is to exploit the logical relationship between the variables. This comes from the fact that although incidence, prevalence, and mortality each gives a different description of the epidemiology of a disease, they are related and are not independent. Specifically, a prevalent case was an incident case at some earlier time and age. Likewise, one has to have the disease first before recovering or dying from the said disease (Barendregt et al., 2003).

The relationship between the said epidemiological parameters can be expressed as a mathematical, Markov-type model of a generic disease process (Figure 1.1). In this model, incidence, case fatality, and remission are the variables that determine the prevalence and mortality of a disease. Because of this causal relationship, they are interrelated and one variable is constrained by the values of the other variables. Barendregt and Ott (2005) describe consistent estimates as those that comply to the constraints given by the causal framework that describes the disease process.

It should be noted that the model in Figure 1.1 works well for chronic diseases. This is because the three transition hazards may be present in chronic diseases. For chronic diseases such as diabetes only, where remission can be ignored, remission rates may be set to zero. Other examples of chronic diseases are cardiovascular diseases, cancer, and chronic respiratory diseases. Using the model for chronic diseases is especially relevant since chronic diseases are a huge global burden. In fact, it is the major cause of death and disability worldwide (World Health Organization, 2005) and even makes up for 86% of deaths in Europe (European Commission, 2014)

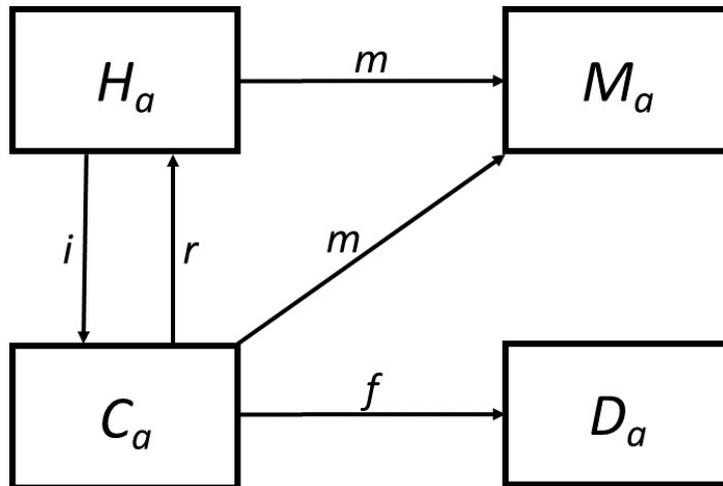


Figure 1.1: The conceptual disease model. H_a refers to the number of healthy people (i.e., without the disease under consideration) at age a ; C_a refers to the number of people with the condition/disease at age a ; D_a refers to the number of people dead from the disease at age a ; and M_a refers to the number of people dead from all other causes at age a . There are four transition hazards: i for incidence, r for remission, f for case fatality, and m for all other mortality.

Moreover, such a generic disease model (Figure 1.1) allows calculation of a complete and internally consistent description of disease epidemiology from partial data. Software implementations of generic disease models for descriptive epidemiology have been proposed since the 1990s. For the original Global Burden of Disease study, the Harvard Incidence Prevalence (HIP) Model was developed. This was a spreadsheet model that took as input a set of instantaneous incidence, remission, and excess mortality rates and produced estimates of disease prevalence and duration. This model later developed into 'DisMod', which has continued to develop in line with the further iterations of the Global Burden of Disease project. DisMod II comes with an interactive graphical interface and allows for a wider range of disease inputs. DisMod MR, the latest iteration of the DisMod family, complements the system dynamics model with a statistical data model, leading to an integrative metaregression framework for descriptive epidemiology (Flaxman et al., 2015)

DisMod MR and DisMod II have their own advantages and disadvantages. DisMod II, in particular, requires interaction with the tool only via the graphical user interface. This is a drawback for most researchers as it requires manual interactions and thus limiting usability. On the other hand, DisMod MR was designed for simultaneously estimating disease epidemiology across all countries and regions of the world (Flaxman et al., 2015). In fact, DisMod-MR has been used for the Global Burden of Disease 2010 and 2013 studies and proved to be very useful. However, this is also less usable for individual researchers who aim to study specific diseases for a certain country. Despite

the development of DisMod MR, recent papers have still surfaced that uses DisMod II (McLean et al., 2015; Sözmen et al., 2015; Wang et al., 2015; Yang et al., 2015). For this reason, there is a need to develop a software that is accessible for researchers and that builds on the weaknesses and strengths of the two iterations of DisMod.

1.2 Objectives of the Study

To overcome the limitations of DisMod II and DisMod MR, this thesis aimed to develop an R package that could be used to estimate chronic disease epidemiology. More specifically, its objective was to develop a model to be used in the R package from existing Incidence-Prevalence-Mortality models. Moreover, another specific objective was to apply a technique for interpolating grouped data, which is common in these type of studies. Finally, the project aimed to accommodate a wider array of input variables, as done by DisMod II.

1.3 Structure of the Thesis

In the subsequent chapter, a detailed description of the data in studies on disease epidemiology is provided. Likewise, the data used in the analyses were also discussed. Chapter 3 provides a discussion on how the model was developed and what its properties were. The fourth chapter provides an overview of the developed R package, which was named `DisModR`. The content of Chapter 5 deals with the analysis of the data using `DisModR`. In this chapter, sample codes are given to simulate discussion. Finally, Chapter 6 gives a discussion on important conclusions and ideas for further research are provided.

Chapter 2

The Epidemiology of a Disease: Data

There are several variables that determine the epidemiology of a disease. Such variables include incidence, prevalence, duration, remission, survival, mortality, and case-fatality, among others. Barendregt and Ott (2005) claim that while each of these variables provides a different view of the epidemiology, they all describe the same disease process, and hence, they are not independent.

These variables are key components in most epidemiological studies. In fact, in 2017 alone, several studies were already found and that used datasets for such variables (Ageno et al., 2017; Gilsanz et al., 2017; Harari et al., 2017; Raheel et al., 2017; among others). However, gathering and compiling these components together are usually a problem. Aside from unavailability, these variables usually come from different sources. Combining these data to estimate disease epidemiology leads to data that are summarized into age groups. This study then uses such kind of data.

In order to demonstrate the features and results of the R package created in this study, a sample dataset from DisMod II was used. The dataset was on breast cancer in the Netherlands. It contained data on incidence, remission, and breast cancer mortality. These were also grouped based on 20 age groups. The data are tabulated in Table 2.1. Likewise, the mortality data for the Dutch population, also taken from DisMod II, were also used. In contrast, there is mortality for every age. This is presented in Figure 2.1.

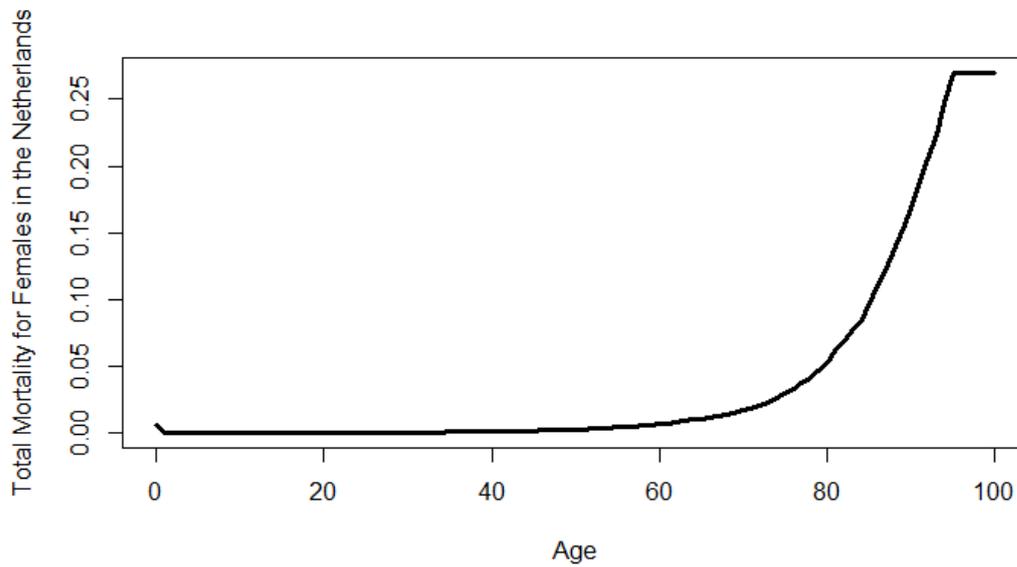


Figure 2.1: Total Mortality for Females in the Netherlands

Table 2.1: Incidence, Remission, and Mortality Data for Breast Cancer in the Netherlands

AgeLB	AgeUB	Incidence	Remission	Mortality
0	4	0.0000	0	0.0000
5	9	0.0000	0	0.0000
10	14	0.0000	0	0.0000
15	19	0.0000	0	0.0000
20	24	0.0000	0	0.0000
25	29	0.0001	0	0.0000
30	34	0.0003	0	0.0001
35	39	0.0007	0	0.0001
40	44	0.0013	0	0.0003
45	49	0.0021	0	0.0004
50	54	0.0028	0	0.0006
55	59	0.0029	0	0.0008
60	64	0.0034	0	0.0010
65	69	0.0039	0	0.0012
70	74	0.0032	0	0.0013
75	79	0.0036	0	0.0016
80	84	0.0038	0	0.0022
85	89	0.0035	0	0.0030
90	94	0.0029	0	0.0041
95+		0.0017	0	0.0061

Chapter 3

Methodology

3.1 The Compartmental Model

In studying and analyzing basic epidemiological parameters such as incidence of, prevalence of, and mortality from a certain disease, it is wise to look at compartmental models. In this case, as in the previous iterations of DisMod (I, II, and MR), the conceptual model used is depicted in Figure 1.1. Three states were considered in relation to a particular disease: Healthy, With Condition, and Dead. Together with the three states, the state of Dead from Other Causes completed the states of the conceptual model. In this model, people who do not have the studied disease are considered healthy and they belong to the Healthy state. Healthy people are subject to an incidence hazard, which means that they could have the condition (or disease) and they could transfer to the With Condition state. Alternatively, people who have the disease are also subject to a remission hazard. This implies that they could recover from the disease and could go back to the Healthy state. In the same sense, people with the condition are also subject to a case fatality hazard. That is, the disease might kill them and they could be transferred to the Dead state. In addition, both healthy people and people with the disease are subject to a mortality hazard from all other causes. This indicates that both subjects in the two states could die from causes other than the disease and might transfer to the state of Dead from Other Causes.

As pointed out by Barendregt et al. (2003), the usual problem encountered in relation to the model was that the mortality from causes other than the disease was often unknown, despite this as an input to the disease model. Of course, this could be calculated from the total mortality rate and the disease specific mortality, but the latter is also often unknown. Barendregt et al. (2003) further proposed a solution to this problem. They used the property that hazards are unaffected by the presence or absence of other hazards that act on the same population. This means that by assuming that mortality from causes other than the disease was independent of the disease, this transition hazard would not affect the value of the transition hazards for incidence, remission, and case fatality. If the value of mortality from all other causes was then set to 0 (and consequently, left out of the derived equations), the right values for the disease rates would still be derived. In addition, the values for the disease prevalence would also be unaffected, when

reported as a proportion of the total population (Barendregt et al., 2003; Barendregt et al., 1993).

Moreover, the assumption that mortality from causes other than the disease was independent of the disease means that this mortality is the same for healthy people and people with the condition. Barendregt et al. (2003) further explained that this assumption also implies that the disease-specific mortality in the model is not necessarily the same as the disease-specific mortality reported by national statistical offices. That is, the disease-specific mortality in the model, which is the excess mortality caused by the disease, is in compliance with the methodology of burden of disease studies, which aims to attribute all excess mortality to the disease.

Furthermore, the model in this paper followed the starting values given in DisMod II. The model starts the analysis with a cohort of 1000 people at age 0, and studies it until the highest age considered. Likewise, the model also assumes that transition hazards are constant within an age interval. The impact of this assumption, however, is minimized by using 1-year age intervals.

The compartmental model in Figure 1.1 was then translated to the following differential equations when mortality from causes other than the disease was ignored:

$$\frac{dS_a}{da} = -i_a S_a + r_a C_a \quad (3.1)$$

$$\frac{dC_a}{da} = -(f_a + r_a) C_a + i_a S_a \quad (3.2)$$

$$\frac{dD_a}{da} = f_a C_a \quad (3.3)$$

where the three model parameters are the incidence hazard (i), remission hazard (r), and case fatality hazard (f); and the three states are S_a , which is the number of healthy people at age a , C_a , which is the number of people with the condition or disease at age a , and D_a which is the number of people who died from the disease at age a .

3.2 Estimation of the Parameters

In lieu of using differential equations to obtain estimates for the number of people in each of the states S , C , and D , Barendregt et al. (2003) derived new equations from equations (3.1), (3.2), and (3.3) using Maple V (Char et al., 1991). These new equations provided reasonable estimates for the said numbers. They were also used in the creation of the R package in order to obtain the estimates at a much faster time.

For a much simpler view of the derived equations, the following intermediate variables were used:

$$l_a = i_a + r_a + f_a$$

$$q_a = \sqrt{(i_a^2 + 2i_a r_a - 2i_a f_a + r_a^2 + 2f_a r_a + f_a^2)}$$

$$w_a = e^{\frac{-(l_a + q_a)}{2}}$$

$$v_a = e^{\frac{-(l_a - q_a)}{2}}$$

With these intermediate variables, the new equations are then:

$$S_a = \frac{2(v_a - w_a)[S_{a-1}(f_a + r_a) + C_{a-1}r_a] + S_{a-1}[v_a(q_a - l_a) + w_a(q_a + l_a)]}{2q_a} \quad (3.4)$$

$$C_a = \frac{(v_a - w_a)\{2[(f_a + r_a)(S_{a-1} + C_{a-1}) - l_a S_{a-1}] - C_{a-1}l_a\} - C_{a-1}q_a(v_a + w_a)}{2q_a} \quad (3.5)$$

$$D_a = \frac{(v_a - w_a)\{2f_a C_{a-1} - l_a(S_{a-1} + C_{a-1})\} - q_a(S_{a-1} + C_{a-1})(v_a + w_a)}{2q_a} + \frac{2q_a(S_{a-1} + C_{a-1} + D_{a-1})}{2q_a} \quad (3.6)$$

As a further step, the age specific prevalence proportion and mortality rates are also calculated using the number of people in each of the three states S , C , and D , which were obtained from equations 3.4, 3.5, 3.6. The R package calculates the age interval person years at risk (PY_a) first using the following formula:

$$PY_a = 0.5(S_a + C_a + S_{a+1} + C_{a+1}) \quad (3.7)$$

The prevalence proportion c_a and the mortality rate b_a are then:

$$c_a = 0.5\left(\frac{C_a + C_{a+1}}{PY_a}\right) \quad (3.8)$$

$$b_a = \frac{D_{a+1} - D_a}{PY_a} \quad (3.9)$$

3.3 Data Smoothing

One main concern of this study was to appropriately represent how the variables in disease modeling varied with age. That is, the study was concerned with how the hazards or rates varied continuously as a function of age. This stemmed from the knowledge that although a large age gap is expected to have large differences in their disease parameters, populations in close age levels do not. For instance, mortality rates are expected to be widely different between age 30 and age 80, but is expected to be somewhat similar between age 30 and age 31. This concern is more evident when the data is grouped according to age groups. It may result to the same mortality rates for age 29 and age 30, but very different rates between age 30 and age 31, provided that the two latter ages are in different groups.

A solution for this problem was to use spline models. In particular, cubic smoothing splines were employed in this study. In this case, the data were thought of as a set of coordinates $(x_0, y_0); (x_1, y_1); \dots; (x_1, y_1)$ where x is the Age parameter, which was already ordered in ascending order, and y is the value of the disease input. The main idea

behind cubic smoothing splines is bridging the gap between adjacent points (x_i, y_i) , using the cubic functions T_i so as to piece together a curve with continuous first and second derivatives (Pollock, 1993). The following is the cubic polynomial used:

$$T_i = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i$$

where x ranges from x_i to x_{i+1} ; $i = 0, \dots, n - 1$; and n is the maximum age number.

In addition, this technique employs some conditions. The first condition is that the adjacent functions T_{i-1} and T_i for $i = 1, \dots, n$ should meet at the point (x_i, y_i) . The second condition is that the first derivatives must be equal at the junction, while the third condition is that of equality between the second derivatives (Pollock, 1993). These requirements must be met in order to calculate the parameters of the cubic spline, and consequently apply a cubic smoothing spline to the data (see Pollock, 1993 for a more detailed derivation).

3.4 Accommodation of More Inputs

A closer look at the compartmental model in Figure 1.1 and equations (3.1), (3.2), and (3.3) shows that the model only accepts three inputs: incidence, remission, and case fatality. As in DisMod II, one of the desired features of the R package was to allow for a wider range of input variables other than the three transition hazards. This is especially useful for situations where not all three transition hazards are observed, but other data such as prevalence and/or disease-specific mortality are available. With that, one solution for this is to manipulate equations (3.1), (3.2), and (3.3), but this requires too much work and may not be tractable.

An iterative optimization method which is the downhill simplex method could be used to solve this problem (Barendregt et al., 2003). The downhill simplex method, which is also known as amoeba method or the Nelder-Mead method, provides direct search algorithms that are based on the update of a simplex (Bihorel and Baudin, 2015). At age 0, the values for the three transition hazards are put in equations (3.4), (3.5), and (3.6) if all are present. If another set of three input variables are present, a loss function evaluates how different the input variables are to the corresponding output. A new transition hazard is then inserted to the missing transition hazard. This procedure iterates until the loss function reaches an unconstrained minimum. The optimization then moves to the next age level.

Currently, the function accepts five inputs, namely: incidence, remission, case fatality, mortality, and prevalence.

Chapter 4

The R Package: DisModR

This chapter provides a discussion on the R package `DisModR`. The inputs for the disease modeling function (`DisMod()`) is first discussed. These inputs range from data from other sources and from subjective inputs of the researchers. In addition, a way to call the outputs of the function is also presented.

4.1 The Disease Modeling function

In this section, an overview of the `DisMod()` function within the package `DisModR` is provided. It is the main function which produces several output plots and tables, which are discussed in the subsequent sections. Its main purpose is to provide other outputs that describe the epidemiology of a disease. The said function is then formulated as follows, with the results being stored in the R object `Epi`.

```
Epi = DisMod(Data, AgeLB, AgeUB, Incidence = NULL, Remission = NULL,  
            Case_Fatality = NULL, Mortality = NULL, Prevalence = NULL,  
            Total_Mortality = NULL, Total_Mortality_Group = NULL,  
            birth_prevalence = 0, Cubic_spline = TRUE, spar = NULL,  
            cv = TRUE, Tol = 1e-09, MaxIt = 1000)
```

The inputs for the `DisMod()` function can be divided into two. The first class is composed of the data needed to run the algorithm. This is discussed in Subsection 4.1.1. The second comprises of options that can be modified to allow the user to control the run of the model. This is presented in Subsection 4.1.2.

4.1.1 Data Inputs

The `DisMod()` function requires the user to upload data needed for the run of the program. The following part of the program is then codes that accept the data into the function:

```
Epi = DisMod(Data, AgeLB, AgeUB, Incidence = NULL, Remission = NULL,  
            Case_Fatality = NULL, Mortality = NULL, Prevalence = NULL,  
            Total_Mortality = NULL, Total_Mortality_Group = NULL, ... )
```

In the function, `Data` must contain the object with the age bounds and the input variables. That is, the subsequent inputs `AgeLB`, `AgeUB`, `Incidence`, `Remission`, `Case_Fatality`, `Mortality`, `Prevalence`, and `Total_Mortality_Group` are stored as columns in the object that is given to `Data`. `AgeLB` and `AgeUB` are the corresponding columns for the lower bounds and upper bounds of the Age groups. Note that should the data not be given in an age schedule, but rather given for every age (in years), the data for age must be placed in both columns. That is, `AgeLB` and `AgeUB` must contain the same dataset of say, 0, 1, 2, ..., etc.

Furthermore, the input variables for disease modeling must be inputted in either `Incidence`, `Remission`, `Case_Fatality`, `Mortality`, and or `Prevalence`. Note that only three inputs are needed for the program to run. `Incidence`, `Remission`, `Case_Fatality`, and `Mortality` accept the incidence hazard of, remission hazard of, case fatality hazard of, and mortality from the selected disease, respectively. Likewise, `Prevalence` asks for the prevalence proportion of the selected disease in the population.

Note that the `DisMod()` function primarily estimates the parameters for disease epidemiology in equations (3.1), (3.2), and (3.3) through equations (3.4), (3.5), and (3.6), respectively. In order to accommodate more inputs such as `Mortality` and `Prevalence`, the `fminsearch()` function of the `neldermead` package is used. This function "searches for the unconstrained minimum of a given cost function" (Bihorel and Baudin, 2015). The cost function evaluates how different the input variables are to the corresponding output, and once this reaches a minimum, the `fminsearch()` function will provide a new value for the missing transition hazard. The procedure will then move to the next age group.

In addition to these variables, the R package needs total mortality rates to calculate specific variables. The `DisMod()` function can accommodate two forms of the total mortality rates of a population. The data must be inputted to `Total_Mortality` if the total mortality have an accompanying data for age (in years). In this case, the input must come from a separate object other than the `Data` object. On the other hand, if data for total mortality are in group intervals like the other input variables, then these can be stored through the `Total_Mortality_Group`. In this case, the input must be a column in the `Data` object.

4.1.2 Additional Options

The `DisMod()` function also allows for additional options to control the behavior of the data or the algorithm. In particular, the following part of the `DisMod()` function allows the user to access the options:

```
Epi = DisMod(..., birth_prevalence = 0, Cubic_spline = TRUE,
             spar = NULL, cv = TRUE, Tol = 1e-09, MaxIt = 1000))
```

The option `birth_prevalence` can be based on the data and/or the subjective decision of the researcher. This refers to the prevalence at birth (at age 0) and its default value is 0, meaning there is no prevalence at birth.

Additionally, in order to perform data interpolation, cubic smoothing splines can be employed in the package. In `DisMod()`, there are also options for cubic smoothing

splines. Setting `Cubic_spline=TRUE`, which is also the default, triggers the program to smooth the data by cubic polynomials before estimating the parameters of the disease epidemiology. More specifically, two options are also available for controlling the smoothing technique. One could set `cv=TRUE`, the default, to use the "ordinary leave-one-out" technique. Setting it to `cv=FALSE` would allow a "generalized" cross-validation. Note that this is used for smoothing parameter computation only when `spar` is not specified (R Core Team, 2017). When `Cubic_spline=FALSE` is specified, the non-interpolated data are used in the function.

Finally, the last options are connected to the nelder-mead algorithm used in this study. It contains optimization options, which drive the behavior of `fminsearch()`. The `fminsearch()` performs the iterative optimization procedure mentioned in the previous chapter. Note that `Tol` is the absolute tolerance on the simplex size and the default value is `1.0E-9`. `MaxIt` is the maximum number of iterations and the default is 1000.

4.2 Output

The output of the `DisMod()` function is composed of a list for the inputs, cubic-spline adjusted inputs, final estimates in yearly intervals, final estimates in their group intervals, as well as their corresponding graph functions.

Suppose the `Epi` object contains the results of the `DisMod()` function. The following commands will call the input table, which is converted to yearly intervals, as well as a plot for the input data.

```
Epi$Input
Epi$Plot_input
```

Should the researcher decide to employ cubic smoothing splines on the data, issuing the following calls will give the input data cubic splines smoothed data and the plot of the said data, respectively:

```
Epi$Cubic_spline
Epi$Plot_cubic_spline
```

Note that when cubic splines are not employed, the above codes will return a `NULL` set.

The main goal of the `DisMod()` function is to estimate parameters of the epidemiology of a disease. The final estimates for this output can be generated by:

```
Epi$Final_estimates
```

Note that these estimates are still in yearly interval form. In order to call the plot of this function, use

```
$Plot_Final_estimates
```

Furthermore, the `DisMod()` function also allows the researchers the option to look at the data in tabulated form grouped in the age schedule of the original data. In order to do this, the following codes should be followed:

```
Epi$Final_Table
```

Likewise, a plot of the final estimates in summarized form can be obtained by:

```
$Plot_Final_groupestimates
```

Other outputs such as `RR_Mortality` can be obtained by calling the list name where it belongs and then calling its object name. `RR_Mortality` refers to the relative risk mortality of the disease. For instance,

```
Epi$Final_estimates$RR_Mortality
```

This is particularly useful when one wants to use the graphical tools in R or when there is a need to compute other parameters using the output data.

Chapter 5

Analysis of the Breast Cancer Data

The chapter gives a detailed example of the various functionalities in `DisModR`. It uses the breast cancer data as an example. Various options are explored and discussed. Likewise, the chapter also provides some examples on how the results are interpreted.

5.1 Uploading the Data

Three input variables were readily available for breast cancer in the Netherlands: incidence, remission, and mortality. The data were in twenty different age groups, most having a range of five except for the last age group. Note that, in order to avoid errors in reading the data, an upper limit of 100 was given to the last age group. The data were first read through the R software. Similarly, the total mortality data were also saved in an object.

```
> bc=read.table(file.choose(),header=T)
> Totmor=read.table(file.choose(),header=T)
> head(bc)
  AgeLB AgeUB Incidence Remission Mortality
1  0  4  0.0e+00  0  0
2  5  9  0.0e+00  0  0
3 10 14  0.0e+00  0  0
4 15 19  0.0e+00  0  0
5 20 24  1.2e-05  0  0
6 25 29  8.0e-05  0  0
> head(Totmor)
  Age Males Females
1  0  0.0071  0.00542
2  1  0.0006  0.00052
3  2  0.0004  0.00033
4  3  0.0002  0.00020
5  4  0.0002  0.00018
6  5  0.0002  0.00020
```

The breast cancer data were saved on the object `bc`, while the total mortality data were saved in the object `Totmor`. It is worth noting that although the breast cancer data were summarized into age schedules, the total mortality data are more detailed, having corresponding data for each age (in years). For this reason, the two datasets were separately saved in different objects. Otherwise, total mortality could just become one column in `bc`.

After reading the data in the R software, these were inputted in the `DisMod()` function. In this situation, the default values for the additional options were used. That is, a cubic smoothing spline using the "ordinary leave-one-out" technique was used. The optimization options were also retained. The Nelder-Mead algorithm was particularly important in this dataset since data for case fatality were not available, but data for breast cancer mortality were.

```
> DM = DisMod(Data = bc, AgeLB = 'AgeLB', AgeUB = 'AgeUB', Incidence =
+   'Incidence', Remission = 'Remission', Mortality = 'Mortality',
+   Total_Mortality = Totmor$Females, birth_prevalence = 0,
+   Cubic_spline = TRUE, spar = NULL, cv = TRUE, Tol = 1e-09,
+   MaxIt = 1000)
```

After running the above program, the results of the `DisMod()` function were automatically saved in the object `DM`. This contained lists of the outputs generated and their corresponding plots.

5.2 The Input Data

One of the plots produced by the `DisMod()` function was a graph of the input data. The resulting graph is presented in Figure 5.1. This and the data used in the plot can be produced by calling two of the list components of `DM`.

```
> DM$Plot_input
> str(DM$Input)
List of 8
 $ Age : num [1:101] 0 1 2 3 4 5 6 7 8 9 ...
 $ Incidence : num [1:101] 0 0 0 0 0 0 0 0 0 0 ...
 $ Remission : int [1:101] 0 0 0 0 0 0 0 0 0 0 ...
 $ Case_Fatality : NULL
 $ Mortality : num [1:101] 0 0 0 0 0 0 0 0 0 0 ...
 $ Prevalence : NULL
 $ RR_Mortality : NULL
 $ Total_Mortality: num [1:101] 0.00542 0.00542 0.00052 0.00033 0.0002
0.00018 0.0002 0.00015 0.00013 0.00011 ...
```

The purpose of graphs such as Figure 5.1 is to show the researchers the behavior of the input variables. Indeed, it can be seen from the graph that there is an increasing trend in mortality due to breast cancer as the age of the females in the Netherlands increases. Similarly, an increasing trend can also be observed for the incidence of breast

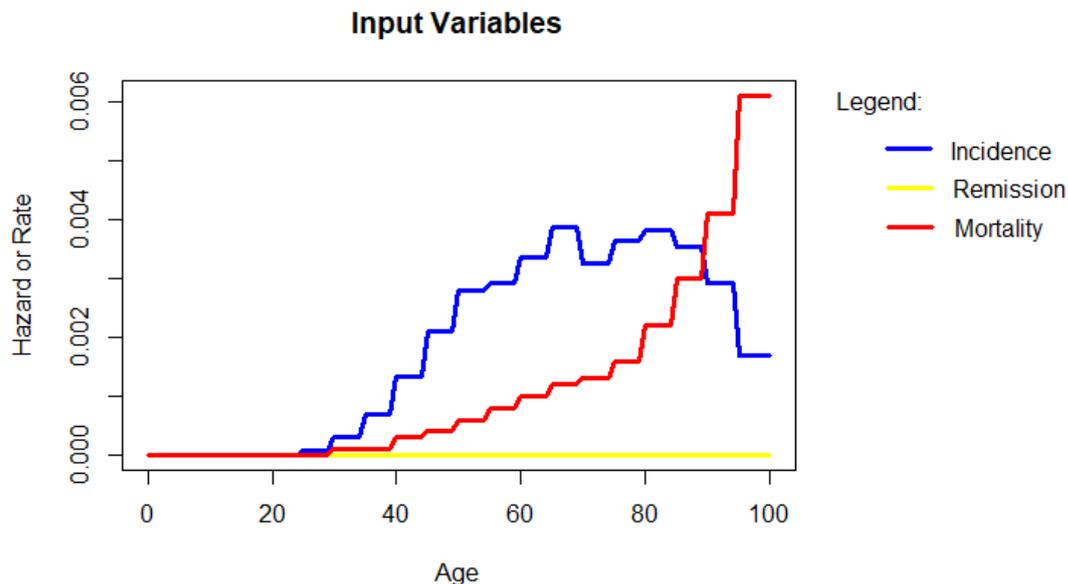


Figure 5.1: Graph of the Input Data From DisMod R

cancer as age increases for women up to 70 years old. This trend starts to go down for the oldest part of the population. Finally, there is no recorded remission in the dataset. Based on this, one may decide if the data behave relatively well, and may perhaps, decide if a technique such as cubic smoothing splines is appropriate.

5.3 Cubic Spline Interpolation

One of the features of DisModR is the interpolation of the grouped data using cubic splines. In fact, this was done in Section 5.1, when the DM object was created, through the following options:

```
> DM = DisMod(..., Cubic_spline = TRUE, spar = NULL, cv = TRUE,
+   Tol = 1e-09,...)
```

An interpolation of the breast cancer data was done using cubic polynomials with default option of "ordinary leave-one out" technique used. Note that the technique can be used in this case since there are no duplicate points per variable for every age value. A plot of the interpolated values, which is presented in Figure 5.2, is called by invoking:

```
> DM$Plot_cubic_spline
```

A closer inspection of the plot shows smoothed curves instead of the usual look of a step function. It should also be noted that DisModR prevents any of the interpolated values to go below zero. This is to avoid any problem during the calculation of the output values, and to make consistent interpretations of the input values. Should one of

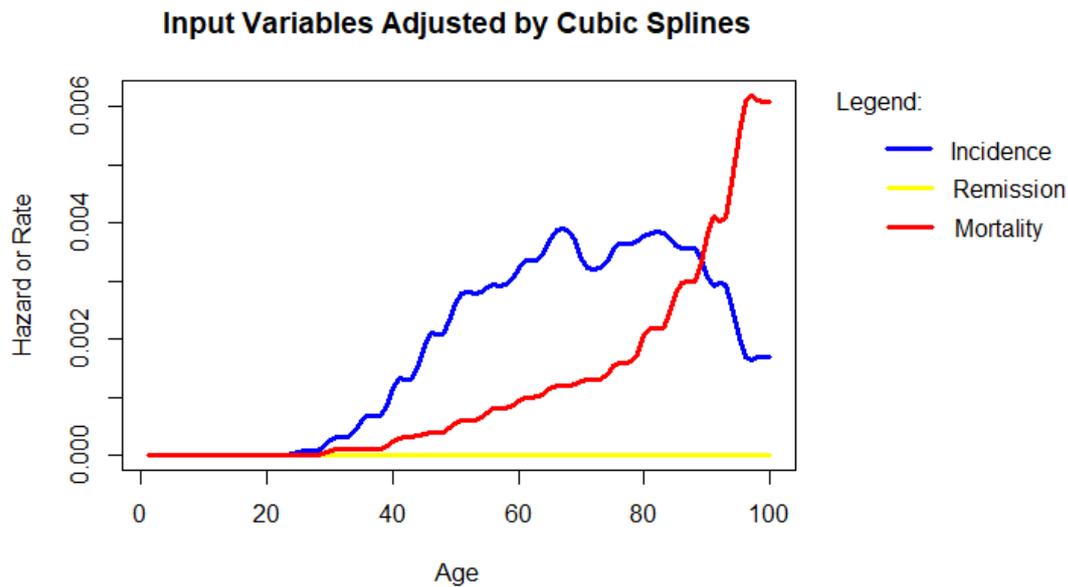


Figure 5.2: Breast Cancer Data with Cubic Spline Interpolation From DisModR (default options are used)

the interpolated values be negative, it would instead be assigned the value in the previous age.

Moreover, researchers could use options in the R package to manipulate the interpolated data to what they deemed acceptable. In DisModR, this option could be done in two ways. The first was by manually calling out the cubic spline interpolated values and editing them.

```
> DM_cs = DM$Cubic_spline
> DM_cs
$Age
[1] 0 1 2 3 4 ...
$input_inci
[1] 0.000000e+00 1.397604e-13 1.397604e-13 9.671499e-13 ...
$input_remi
[1] 0 0 0 0 ...
$input_cfat
NULL
$input_mort
[1] 0.000000e+00 0.000000e+00 2.118183e-15 2.118183e-15 ...
$input_prev
NULL
$input_rrmort
NULL
$totmor
```

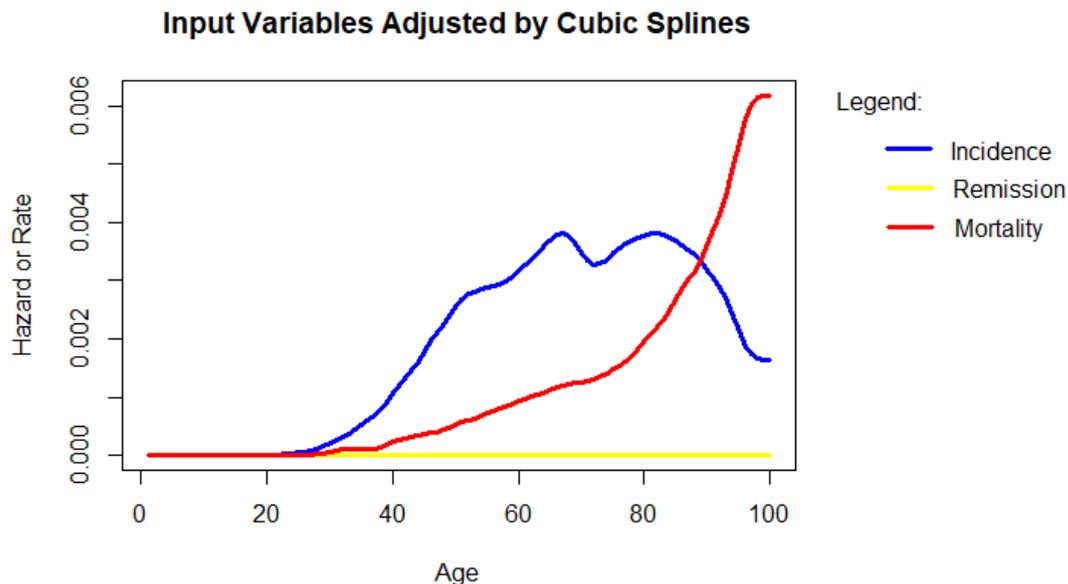


Figure 5.3: Breast Cancer Data with Cubic Spline Interpolation From DisModR (spar=0.4 was used)

```
[1] 0.00542 0.00542 0.00052 0.00033 ...
```

The new values might then once again be used as inputs in `DisMod()` with `DM_cs$Age` being the new input for both `AgeLB` and `AgeUB`. Another method was by making use of the options `spar` and `cv`. For instance, instead of using the `cv` option, one could specify `spar=0.4` instead.

```
> DM2=DisMod(Data=bc, AgeLB='AgeLB', AgeUB = 'AgeUB', Incidence='Incidence',
+   Remission='Remission', Mortality = 'Mortality', Total_Mortality =
+   Totmor$Females, Cubic_spline = TRUE, spar = 0.4, cv = TRUE)
```

The only difference between `DM` and `DM2` is the `spar` parameter. Because the smoothing parameter was specified, the `cv` option was not used by the program. Extracting the plot for the cubic spline interpolated data,

```
> DM2$Plot_cubic_spline
```

leads to Figure 5.3. It can be noticed that compared to Figure 5.2, the lines in the plot is much smoother. In fact, the effect of earlier groupings in the data is barely noticeable.

5.4 Parameter Estimates of Disease Epidemiology

Recall that the interpolated data from Figure 5.2 were used in `DisModR`, where the results are stored in the object `DM` from Section 5.1. The main goal of the package was to produce complete estimates of disease/breast-cancer incidence, remission, case fatality, mortality,

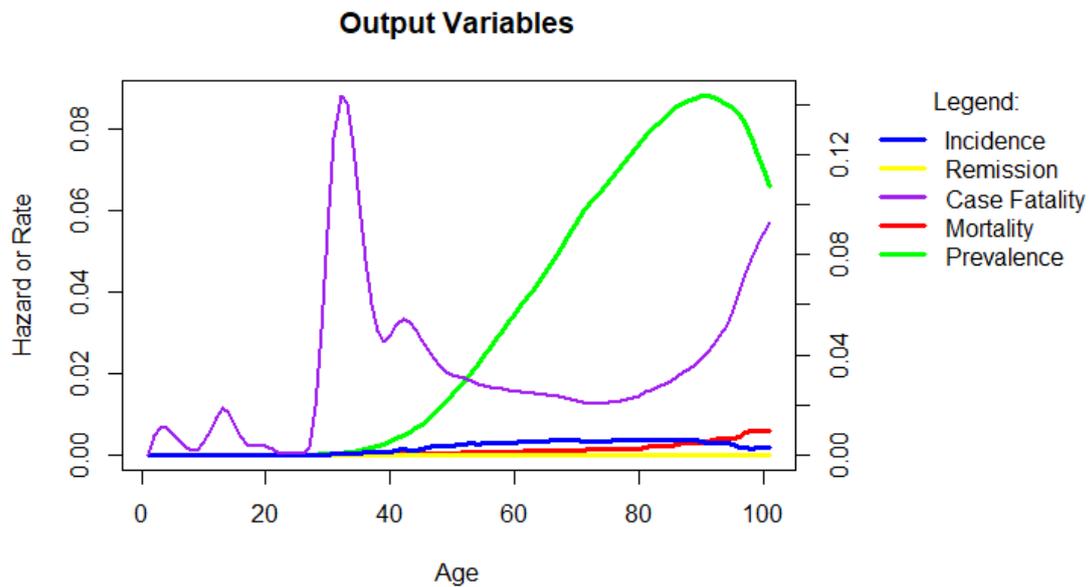


Figure 5.4: Output variables generated with DisModR (with cubic spline interpolation)

prevalence, and relative-risk mortality. The estimates could simply be generated by calling a list in DM:

```
> DM$Final_estimates
$Incidence
[1] 0.000000e+00 1.397604e-13 1.397604e-13 9.671499e-13 ...
$Remission
[1] 0 0 0 0 ...
$Case_fatality_downhill
[1] 0.0000000000 0.0080749961 0.0109020050 0.0110805090 ...
$Prevalence_proportion
[1] 0.000000e+00 0.000000e+00 6.988854e-14 2.065336e-13 ...
$Mortality_rate
[1] 0.000000e+00 0.000000e+00 0.000000e+00 6.233864e-15 ...
$RR_Mortality
[1] 1.000000 1.000000 59.086527 31.495344 ...
```

Consequently, researchers could also prefer a graph. Figure 5.4 is generated by the following codes:

```
> DM$Plot_Final_estimates
```

In the figure, there is peak for the back-calculated case fatality at age 32. This means that the rate of women dying from breast cancer is highest at this age. Similarly, the prevalence of breast cancer is highest for females aged 92.

DisModR also produces the estimates of the different components of the epidemiology, as they are grouped by the original age schedule. Although it is produced in a list

format, a slight manipulation could easily produce the table of estimates by the original age schedule.

```
> DM = DM$Final_Table
> DM_table = cbind(DM$AgeLB,DM$AgeUB,DM$Incidence,DM$Remission,
+                 DM$Case_Fatality, DM$Mortality, DM$Prevalence, DM$RR_Mortality)
> colnames(DM_table)=c("AgeLB","AgeUB","Incidence","Remission",
+                      "Case_Fatality","Mortality","Prevalence","RR_Mortality")
> DM_table
AgeLB AgeUB Incidence Remission Case_Fatality Mortality Prevalence RR_Mortality
[1,] 0 4 0.0000 0 0.0075 0.0000 0.0000 23.1455
[2,] 5 9 0.0000 0 0.0044 0.0000 0.0000 17.7234
[3,] 10 14 0.0000 0 0.0130 0.0000 0.0000 107.2864
[4,] 15 19 0.0000 0 0.0063 0.0000 0.0000 11.4346
...
```

The complete table is shown in Table 5.1.

Table 5.1: An Exhaustive Epidemiology of Breast Cancer

AgeLB	AgeUB	Incidence	Rem.	Case Fat.	Mort.	Prev.	RR Mort.
0	4	0.0000	0	0.0075	0.0000	0.0000	23.1455
5	9	0.0000	0	0.0044	0.0000	0.0000	17.7234
10	14	0.0000	0	0.0130	0.0000	0.0000	107.2864
15	19	0.0000	0	0.0063	0.0000	0.0000	11.4346
20	24	0.0000	0	0.0017	0.0000	0.0000	7.8994
25	29	0.0001	0	0.0149	0.0000	0.0001	5.8430
30	34	0.0003	0	0.1246	0.0001	0.0006	301.6283
35	39	0.0006	0	0.0666	0.0001	0.0018	86.5755
40	44	0.0012	0	0.0512	0.0003	0.0048	52.6450
45	49	0.0019	0	0.0393	0.0004	0.0101	23.1919
50	54	0.0027	0	0.0305	0.0006	0.0183	12.4041
55	59	0.0029	0	0.0268	0.0008	0.0285	7.5060
60	64	0.0033	0	0.0248	0.0010	0.0387	4.7940
65	69	0.0038	0	0.0232	0.0012	0.0499	3.2314
70	74	0.0034	0	0.0210	0.0013	0.0615	2.2517
75	79	0.0036	0	0.0218	0.0015	0.0707	1.7499
80	84	0.0038	0	0.0261	0.0020	0.0796	1.4939
85	89	0.0036	0	0.0329	0.0027	0.0861	1.3438
90	94	0.0030	0	0.0446	0.0036	0.0875	1.2624
95	100	0.0019	0	0.0742	0.0053	0.0786	1.2873

Additionally, a graph of Table 5.1 could also be generated, albeit this is only a weaker version of Figure 5.4, as it is more summarized. Using the following codes, Figure 5.5 is produced:

Moreover, as mentioned in the previous chapter, users of DisModR may opt to use the original data. This means that individuals belonging to a certain age group have the

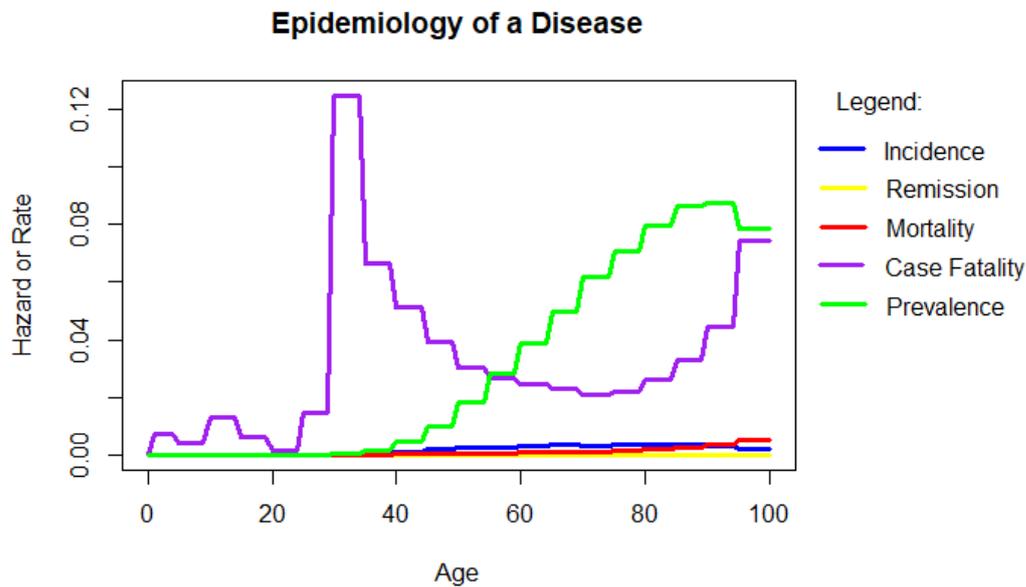


Figure 5.5: Output variables generated by DisMod R (summarized into age groups)

same values for the input variables. In this case, a 30 year old woman have the same incidence, remission, and mortality data with a female aged 34 years old. This can be done by using the following codes:

```
> DM3=DisMod(Data=bc, AgeLB='AgeLB', AgeUB = 'AgeUB', Incidence=
+           'Incidence', Remission='Remission',Mortality = 'Mortality',
+           Total_Mortality = Totmor$Females, Cubic_spline=FALSE)
```

The difference between the objects DM and DM3 is that cubic splines were not employed in the latter. The two objects both used the default optimization options. In order to see the results, Figure 5.6 is generated using the usual call function:

```
> DM3$Plot_Final_estimates
```

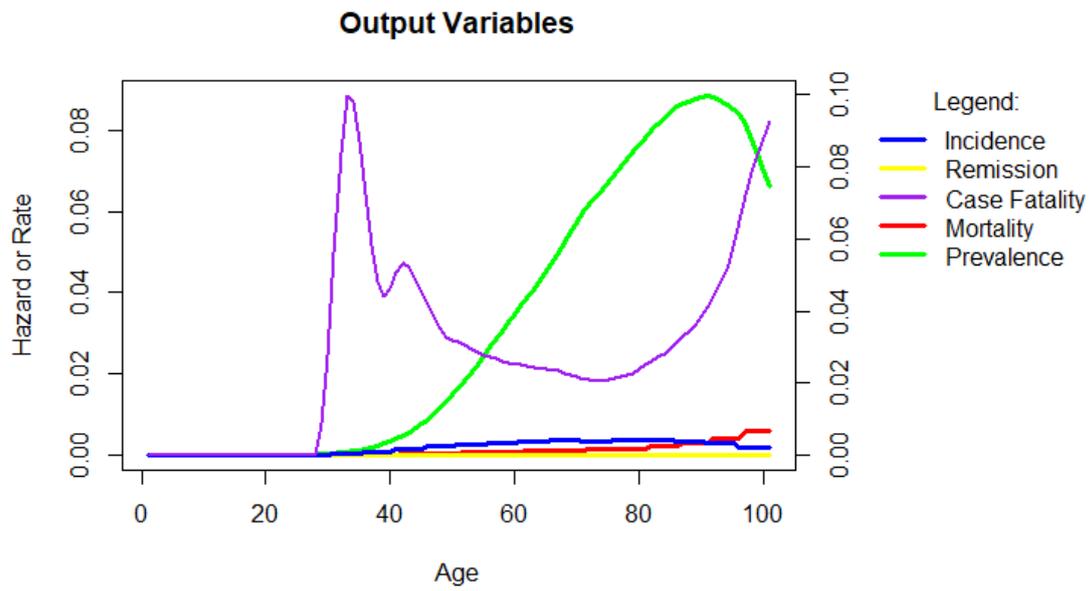


Figure 5.6: Output variables generated with DisModR (no interpolation)

Chapter 6

Discussion and Some Ideas for Further Research

6.1 Discussion

Although primary data are always preferable, it cannot be denied that such data are not always available or reliable. The focus of this thesis was then to create an R package that could be used as an incidence-prevalence-mortality model, which accounts for data problems in disease epidemiology. Because of the popularity and usefulness of the R software, this is an attractive tool for researchers, who could use this to study the epidemiology of a disease.

In the earlier chapters, the advantage of such a software package and model is highlighted. The main objective of the study was to provide estimates for variables that are not present, but otherwise describe the epidemiology of a disease. For instance, in the breast cancer data set, the data available are only on incidence, remission, and mortality. Using the model discussed in this study, it is now possible to identify other components of the epidemiology of breast cancer. In fact, in the example, case fatality was back calculated in `DisModR` and prevalence was computed. This is very useful especially when data are not readily available. This package offers a way to have disease epidemiology estimates at a much cheaper cost.

Initially, the compartmental model (and consequently, the R package) only accepts three inputs: incidence, remission, and mortality. In order to have a wider array of inputs, the `fminsearch()` function from the `neldermead` package is employed. Currently, five inputs can be accepted by the package: incidence, remission, case fatality, mortality, and prevalence. This is advantageous since it is not always the case that all three original inputs are available. With this option, it is now possible to estimate disease epidemiology with say, incidence, remission, and mortality, as in the breast cancer data set.

Moreover, `DisModR` also offers users a tool to remove discontinuities by way of cubic smoothing splines. The assumption that two adjacent ages have similar values for the variables in the epidemiology of a disease is reasonable. Likewise, not removing such discontinuity may change the results in a considerable way. In such cases of this kind, the researcher is expected to make a decision on which result to follow.

Another application that the `DisModR` brings is in the way of checking data consistency. Since the parameters are embedded in a logical relationship, one can exploit this property in order to study datasets that the researchers feels is unreliable. For example, data on prevalence may be compared to the prevalence output in this package to see if it is underestimated or not. Of course, a degree of caution must be exercised in this application. Barendregt et al. (2003) warns against real or deceptive inconsistencies. "Real inconsistency may be due to the combination of measurements from different sources, or to measurement and sampling error. Deceptive inconsistency is due to past trends in disease epidemiology." (Barendregt et al., 2003)

`DisModR` offers the researcher the chance to run the compartmental model in the R environment, which is especially advantageous for a variety of reasons. When one wants to edit and create graphs, the environment is easy to use. Likewise, `DisModR` avoids the burdensome nature of `DisMod II` and offers a simplistic process in its operation. The uploading of data in `DisModR` is simple and fast. When one wants to employ cubic splines for all parameters, `DisModR` is efficient.

Finally, it is a given that the results in `DisModR` is not the evidence, but only part of it. The output must be combined with expert knowledge to arrive at a reasonable solution. When a result does not make sense for a certain variable, this must alert the researcher that there is a potential problem. As a matter of fact, options in the package give the researcher the allowance to modify the process. In the end, the user must decide whether the output is acceptable or not.

6.2 Further Research

The development of this package, as well as its applications, does not end with this study. There is still much work that needs to be done. This project acts as a prelude for a wide area of further research.

One such aspect that needs to be addressed immediately is the incorporation of Bayesian techniques. `DisMod MR` shows that a negative binomial distribution works well as a prior distribution. However, it only considers the three original input variables: incidence, remission, and case fatality. Indeed, one can start from this to develop a model that allows the Bayesian measure of uncertainty. It should be noted, however, that the conception of this package is based more on `DisMod II`, and may have varying principles with that of `DisMod MR`. This may be a possible roadblock.

Although there is already a wider range of input variables that the package accommodates, it can still accommodate more. Such components like duration and survival are some possibilities that can be included as input variables. A problem foreseen with including disease duration among the inputs is that it is affected by mortality from all other causes. Because of this, computation time may be affected. In fact, `DisMod II` only allows duration data which are less than one year (Barendregt et al., 2003), and which is not realistic for many diseases. On the other hand, the survival variable may pose another type of problem. The current assumption is that the input variables are not dependent on time since onset of the disease, but on age (Barendregt and Ott, 2005). This may not be as realistic for survival.

Lastly, how the components of a disease epidemiology change with respect to time is a concept that needs to be addressed. Even in the latest iteration of DisMod (DisMod MR), this has been an issue that needs to be further developed. As for DisMod II, incorporation of time trends is present. However, this is not as efficient as it is and clearly needs to be improved. Issues on computational efficiency is a foreseeable challenge in this aspect.

References

- Ageno, W., Dentali, F., Pomero, F., Fenoglio, L., Squizzato, A., Pagani, G., ..., and Bonzini, M. (2017). Incidence rates and case fatality rates of portal vein thrombosis and Budd-Chiari Syndrome. *Thrombosis and Haemostasis*, 117(4), 794-800.
- Barendregt, J. J., and Ott, A. (2005). Consistency of epidemiologic estimates. *European journal of epidemiology*, 20(10), 827-832.
- Barendregt, J. J., Van Oortmarsen, G. J., Van Hout, B. A., Van Den Bosch, J. M., and Bonneux, L. (1998). Coping with multiple morbidity in a life table. *Mathematical Population Studies*, 7(1), 29-49.
- Barendregt, J. J., Van Oortmarsen, G. J., Vos, T., and Murray, C. J. (2003). A generic model for the assessment of disease epidemiology: the computational basis of DisMod II. *Population health metrics*, 1(1), 4.
- Bihorel, S., and Baudin, M. (2015). neldermead: R port of the Scilab neldermead module. R package version 1.0-10. URL <https://CRAN.R-project.org/package=neldermead>
- Char, B. W., Geddes, K. O., Gonnet, G. H., Leong, B. L., Monagan, M. B., and Watt, S. M. (2012). First leaves: a tutorial introduction to Maple V. *New York: Springer*.
- Flaxman, A. D., Vos, T., and In Murray, C. J. L. (2015). An integrative metaregression framework for descriptive epidemiology.
- Gilsanz, P., Winning, A., Koenen, K. C., Roberts, A. L., Sumner, J. A., Chen, Q., ..., and Kubzansky, L. D. (2017). Post-traumatic stress disorder symptom duration and remission in relation to cardiovascular disease risk among a large cohort of women. *Psychological medicine*, 47(8), 1370-1378.
- Harari, S., Caminati, A., Madotto, F., Conti, S., and Cesana, G. (2017). Epidemiology, survival, incidence and prevalence of idiopathic pulmonary fibrosis in the USA and Canada. *European Respiratory Journal*, 49(1), 1601504.
- Kruijshaar, M. E., Barendregt, J. J., and Hoeymans, N. (2002). The use of models in the estimation of disease epidemiology. *Bulletin of the World Health Organization*, 80(8),

622-628.

McLean, S., Barbour, V., Wild, S., Simpson, C., and Sheikh, A. (2015). Models for estimating projections for disease prevalence and burden: a systematic review focusing on chronic obstructive pulmonary disease. *Journal of health services research and policy*, 20(4), 246-253.

Pollock, D. S. G. (1993). Smoothing with cubic splines.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>

Raheel, S., Shbeeb, I., Crowson, C. S., and Matteson, E. L. (2017). Epidemiology of Polymyalgia Rheumatica 2000-2014 and Examination of Incidence and Survival Trends Over 45 Years: A Population-Based Study. *Arthritis care & research*, 69(8), 1282-1285.

Sözmen, K., Unal, B., Capewell, S., Critchley, J., and O'Flaherty, M. (2015). Estimating diabetes prevalence in Turkey in 2025 with and without possible interventions to reduce obesity and smoking prevalence, using a modelling approach. *International journal of public health*, 60(1), 13-21.

Wang, C., Cui, Y., Li, Y., Liu, X., and Han, J. (2015). A systematic review of hereditary spherocytosis reported in Chinese biomedical journals from 1978 to 2013 and estimation of the prevalence of the disease using a disease model. *Intractable and rare diseases research*, 4(2), 76-81.

World Health Organization (2005). Chronic diseases and their common risk factors. URL www.who.int/chp

European Commission (2014). Chronic Diseases: The health challenge of our times. URL https://ec.europa.eu/health/sites/health/files/major_chronic_diseases/docs/2014_chronic_diseases_informationsheet_en.pdf

Yang, L. P., Liang, S. Y., Wang, X. J., Li, X. J., Wu, Y. L., and Ma, W. (2015). Burden of disease measured by disability-adjusted life years and a disease forecasting time series model of scrub typhus in Laiwu, China. *PLoS neglected tropical diseases*, 9(1), e3420.