ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Winter 2012

Lisa Borsi

Estimating the causal effect of switching to second-line antiretroviral HIV treatment using G-computation

Submission Date: March 2nd 2012

Advisers: Prof. Dr. Marloes Maathuis (ETH Zurich)
Dr. Markus Kalisch (ETH Zurich)
Dr. Thomas Gsponer (ISPM Bern)

I would like to thank Prof. Marloes Maathuis for giving me the opportunity to work on this interesting topic, as well as Markus Kalisch and Thomas Gsponer for their great advice and the many interesting discussions. It was a pleasure for me to deepen and apply my statistical knowledge in the field of biostatistics.

Abstract

Understanding causal effects between exposure and outcome is of great interest in many fields. In this work, the causal effect of switching to second-line antiretroviral treatment on death is estimated for a study population including HIV-infected patients experiencing immunological failure in Southern Africa (Zambia and Malawi). CD4 cell count is considered as a time-varying confounder of treatment switching and death, while it is itself affected by previous treatment. Given the impossibility to conduct a randomised experiment, we address the problem of time-varying confounding by G-computation. Under certain conditions, G-computation yields consistent estimates of the causal effect by simulating what would happen to the study population if treatment is set to a certain regime by intervention. In our analysis we compare intervention "always switch to second-line treatment" to intervention "always remain on first-line treatment". We find the resulting risk ratio to be 0.24 (95% CI 0.14-0.33), emphasizing that the risk of dying is smaller in the population that switched to second-line treatment than in the population that stayed on first-line treatment. Thus, we conclude that there is a beneficial causal effect of switching to second-line treatment among HIV-patients experiencing immunological failure.

Contents

Abstract iii

Contents v

List of Figures vii

List of Tables ix

1 Introduction 1
- 1.1 Brief Introduction to HIV and its Complications . . . 1
- 1.2 Problem Description . . . 2
- 1.3 Literature Survey . . . 3
- 1.4 Outline . . . 3

2 Causal Graphs 5
- 2.1 Basic Facts . . . 5
- 2.2 Statistical Association and Causality . . . 7
- 2.3 Conditional Independence and Interventions . . . 8
- 2.4 Control for Confounding . . . 10

3 The G-computation Procedure 15
- 3.1 Motivation . . . 15
- 3.2 Basic Idea of G-computation . . . 16
- 3.3 The Estimation Procedure . . . 17
  - 3.3.1 Modelling relationships between observed variables . . . 17
  - 3.3.2 Simulating a new dataset under intervention . . . 18
  - 3.3.3 Statistical inference . . . 18
  - 3.3.4 Some remarks . . . 18
- 3.4 Examples . . . 19
  - 3.4.1 Fixed time-point example . . . 19
  - 3.4.2 Two time-point example . . . 21
- 3.5 Drawbacks . . . 23
- 3.6 Reference to Counterfactuals . . . 23
- 3.7 An Alternative to G-computation . . . 24

4 Data Analysis 25
- 4.1 Data Description . . . 25
- 4.2 Imputation of Missing Values . . . 26
  - 4.2.1 Predictive mean matching . . . 27
  - 4.2.2 Combining results of multiple imputation . . . 27
  - 4.2.3 Alternative method for imputation of missing values . . . 28
- 4.3 G-computation . . . 28
  - 4.3.1 Models for log CD4 and death . . . 30

4.3.2 Simulation with intervention . . . 31
4.3.3 Estimates of the causal effect . . . 32
4.3.4 Statistical inference . . . 33

5 Results **35**

6 Discussion **39**

A **43**

B **44**
B.1 R-code corresponding to Section 3.4 . . . 44
B.1.1 Fixed time-point example (Section 3.4.1) . . . 44
B.1.2 Two time-point example (Section 3.4.2) . . . 47
B.2 R-output of G-computation corresponding to Section 3.4 . . . 50
B.2.1 Fixed time-point example (Section 3.4.1) . . . 50
B.2.2 Two time-point example (Section 3.4.2) . . . 50

C **51**
C.1 Notational equivalence . . . 51
C.2 Model for log CD4 . . . 51
C.3 R-Code corresponding to G-computation applied to data described in Chapter 4 . . . 53
C.4 R-Output corresponding to G-computation applied to data described in Chapter 4 . . . 61
C.5 Simulated datasets with intervention $\bar{1}$ and intervention $\bar{0}$ . . . 64

Bibliography **69**

List of Figures

1.1 Causal graph corresponding to the relations between CD4 cell count (C_t), treatment (A_t) and death (Y_t) over time $t = 0, 1$. 2

2.1 Example of a directed acyclic graph. 6
2.2 Graph illustrating d-separation. 7
2.3 Graphical illustration: statistical association vs. causality. 8
2.4 Graphical representation of blocked paths. 9
2.5 Causal DAG with nodes $A_1, \ldots, A_7$. 9
2.6 Causal graph modified by intervention on A_4. 10
2.7 Causal graph corresponding to the relations between CD4 cell count (C), treatment (A) and death (Y). 11

3.1 Causal graph corresponding to Example 3.1 and representing the relations between CD4 cell count (C_t), treatment (A_t) and death (Y) over time $t = 0, 1$. 15
3.2 Causal graph representing the relations between exposure A_t, time-varying confounder C_t and outcome Y_t over time $t = 0, \ldots, T$. 16
3.3 Causal graph corresponding to the relations between CD4 cell count (C), treatment (A) and death (Y). 19
3.4 Causal graph corresponding to Example 3.1 21

4.1 Causal graph representing the relations between treatment switching A_t, log CD4 cell count C_t and death Y_t over time $t = 0, \ldots, T$. 29
4.2 Average mean squared error for different linear regression models resulting from leave-one-out cross validation applied to the five imputed datasets. . . 31
4.3 AIC for different data and models. 32

List of Tables

2.1 Invented data for Example 2.2. 12

3.1 Invented data for Example 2.2 19
3.2 G-computation estimates for risk difference between intervention $A = 0$ and $A = 1$. 21
3.3 G-computation estimates for risk difference, risk ratio and odds ratio between intervention $\{A_0, A_1\} = \{0, 0\}$ and $\{A_0, A_1\} = \{1, 1\}$. 23

4.1 Distribution of death across baseline and treatment variables in the original dataset. 26

5.1 Results of G-computation with 50 bootstraps for log incidence rate. 36
5.2 Results of G-computation with 50 bootstraps for cumulative incidence in %. 36
5.3 Risk ratio obtained by combining results of G-computation with 50 bootstraps for log incidence rate. 36
5.4 Results obtained by combining G-computation results of the five imputed datasets. 37
5.5 Results of G-computation with 50 bootstrap for the dataset imputed by last observation carried forward. 37

6.1 Distribution of treatment switching across baseline variables in observed dataset. 41

C.1 Average mean squared error obtained from leave-one-out cross validation. . 52
C.2 Distribution of death across baseline variables in dataset 1. 64
C.3 Distribution of death across baseline variables in dataset 2. 65
C.4 Distribution of death across baseline variables in dataset 3. 65
C.5 Distribution of death across baseline variables in dataset 4. 66
C.6 Distribution of death across baseline variables in dataset 5. 66
C.7 Distribution of death across baseline variables in dataset 6. 67

Chapter 1

Introduction

The present work deals with an epidemiological problem setting provided by the Institute of Social and Preventive Medicine in Bern (referred to as ISPM) and IeDEA (International epidemiologic Databases to Evaluate AIDS). It consists of a longitudinal study of HIV-infected patients in Southern Africa (Zambia and Malawi). Based on this study we are seeking to clarify whether or not, after immunological failure, treatment switching is beneficial in terms of lifetime for HIV-patients in Southern Africa. To fully understand the subsequent problem description, we first give a short introduction to the Human Immunodeficiency Virus (HIV).

1.1 Brief Introduction to HIV and its Complications

The acronym HIV stands for Human Immunodeficiency Virus. As its name suggests, the virus weakens the immune system of its host, in particular by attacking a certain white blood cell type called CD4 cells. These cells play a major role in the immune system's response to viruses, bacteria and fungi.

HIV-infected persons receive an antiretroviral therapy aiming to maximally suppress the virus. However, the HI-virus is able to mutate and resist to antiretroviral drugs which may result in treatment failure. Patients suffering from treatment failure therefore have to be switched from their first-line treatment to a second-line treatment.

Of main interest among HIV-complications is the question when to switch treatment and hence, how to measure treatment failure. Though, it is not quite clear how to define the optimal switching point [1]. Switching too early may result in non-effectiveness of potential survival benefits due to the first-line treatment, whereas switching too late can inhibit the benefits of second-line treatment and expose the patient to additional death risk. A drawback, especially for less developed countries, is that second-line treatments bare higher costs [2].

According to the WHO's "2006 revision of Antiretroviral therapy for HIV infection in Adults and Adolescents: Recommendations for a public health approach" [1] there are three different failure types each of them (or a combination) being a possible measure for treatment failure:

- **Clinical failure** is based on disease progression and WHO staging (the WHO has

defined several disease stages for HIV infected patients [3]).

- **Virological failure** is based on measuring viral load (copies of the virus in a human's plasma). This procedure is not widely accessible and very cost-intense.
- **Immunological failure** is assessed by looking at the CD4 cell count.

Amongst others, the WHO suggests to decide on treatment switching by considering the CD4 cell count and underlines that the "CD4 cell count remains the strongest predictor of HIV-related complications" [1]. As mentioned in the same source, reasonable working definitions of immunological failure are:

1. CD4 count below 100 cells/mm^3 after six months of therapy;
2. a return to, or a fall below, the pre-therapy CD4 baseline after six months of therapy; or
3. a 50% decline from the on-treatment peak CD4 value (if known).

The dataset presented in this work contains only patients experiencing immunological failure. Nevertheless, as seen later, it is obvious that treatment switching was not (only) based on immunological failure.

1.2 Problem Description

In this work we are seeking to clarify whether or not, after immunological failure, treatment switching is beneficial in terms of lifetime for HIV-patients in Malawi and Zambia. The impossibility of having a randomised experiment allocating treatment switching at random to each patient (ethically unacceptable) renders the task more difficult and necessitates seeking for other solutions. In the absence of a randomised trial we have to deal with time varying confounding: at each point in time the CD4 cell count influences both treatment switching and death. On the other hand treatment switching has an influence on the next CD4 cell count and has an impact on death-risk. A graph corresponding to the described situation could be the graph presented in Figure 1.1.

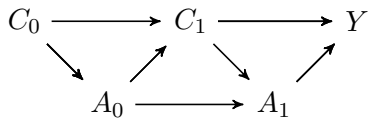

Figure 1.1: Causal graph corresponding to the relations between CD4 cell count (C_t), treatment (A_t) and death (Y_t) over time $t = 0, 1$.

Given this complexity, standard methods fail in estimating the causal effect of treatment switching on death. In the present work, G-computation was chosen as a method to yield unbiased estimates of the causal effect in the described problem setting.

1.3 Literature Survey

Amongst others G-computation has been implemented by Snowden, Rose and Mortimer in *Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique* [4]. They introduce a simple application of G-computation and compare it to standard regression. A more involved application of the method can be found in *Intervening on risk factors for coronary heart disease: an application of the parametric g-formula* by Taubman, Robins, Mittleman and Hernán [5] and in *Effects of Multiple Interventions* by Robins, Hernán and Siebert [6]. Daniel implemented G-computation in a Stata module entitled *GFORMULA: Stata module to implement the g-computation formula for estimating causal effects in the presence of time-varying confounding or mediation.* The paper *gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula* [7] by Daniel, De Stavola and Cousens describes how the Stata module can be applied to given datasets. In the present work, G-computation was implemented in the statistical software R following these literature examples with the goal to apply it to the data described in Chapter 4. In *Marginal Structural Models and Causal Inference* by Peter [8] and *The Causal Effect of Switching to Second-line ART in Programmes without Access to Routine Viral Load Monitoring* by Gsponer et al. [9], similar data were already analysed by using IPTW instead of G-computation.

1.4 Outline

Having defined the research question we will present the G-computation procedure as a possible solution to estimate the causal effect. In Chapter 2 causal graphs are introduced as a tool to detect identifiability of causal effects. Chapter 3 explains and illustrates the G-computation procedure. The dataset provided by ISPM Bern is introduced in Chapter 4 and the application of G-computation to the dataset is explained in detail. The results of this analysis are then presented in Chapter 5 and finally in Chapter 6 conclusions are drawn from the analysis.

Chapter 2

Causal Graphs

The aim of this chapter is to introduce a graphical tool to detect whether a given set of variables is sufficient to identify causal effects. First, basic definitions of graph theory are brought into the context of causality and some new definitions are stated, following Pearl [10], [11]. Interest lies in estimating the causal effect of some exposure variable on an outcome. In an epidemiological context, these could be for instance treatment and death. Bias arising due to confounding is essentially a consequence of the inability to conduct controlled randomised experiments. Therefore the content of this chapter is of main use in the setting of observational studies.

2.1 Basic Facts

The theory of Directed Acyclic Graphs (DAG) offers a way to analyse relationships between variables. Starting from a set of variables and a set of assumptions on causal influences, a graph can be constructed in the following way:

- Any variable is represented by a *node*.
- If a variable A directly influences a variable B, there is an *arrow* (a directed edge) going from A to B. The arrow represents a *direct effect* of A on B and A is said to be a *direct cause* of B.
- A variable A is a *cause* of another variable B, if there is a directed path of arrows leading from A to B. Then B is also said to be a *descendant* of A, or *affected* by A.

To complete the vocabulary concerning graphs, let us remind that,

- a *path* is a sequence of arrows regardless of their direction,
- a *directed path* from A to B is a path where all arrows point towards B,
- a *causal path* is a directed path,
- if B is directly affected by A then B is a *child* of A, and A is said to be a *parent* of B,

- a graph is *directed* if all edges are arrows,
- a graph is *acyclic* if no directed path in the graph forms a closed loop,
- a path that connects A to B is a *backdoor path* from A to B if it has an arrowhead pointing to A,
- B is called a *collider* on a path if the path enters and exits B with arrowheads pointing on B.

A directed acyclic graph is called a causal graph, reflecting the fact that all nodes are causes or effects (directed) and that no cause can cause itself (acyclic). Figure 2.1 illustrates some of the previously stated definitions. It is important to recognise that causal graphs are of qualitative nature only and do not include any assumption on the functional form of the relations and distributions among the variables.

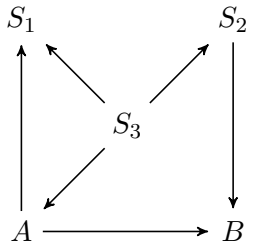

Figure 2.1: An example of a directed acyclic graph. S_1 is a collider on the path $A \rightarrow S_1 \leftarrow S_3 \rightarrow S_2 \rightarrow B$ from A to B. The path $A \leftarrow S_3 \rightarrow S_2 \rightarrow B$ is a *backdoor path* from A to B.

Definition 2.1. *Let A, B and S be three disjoint subsets of nodes in a DAG G, and let p be any path between a node in A and a node in B. S is said to block p if there is a node w on p satisfying one of the following two conditions:*

1. *w is a collider (along p) and neither w nor any of its descendants are in S, or,*
2. *w is not a collider (along p) and w is in S.*

Definition 2.2. *Let A, B and S be three disjoint subsets of nodes in a DAG G. S is said to d-separate A from B, if and only if S blocks every path from a node in A to a node in B.*

Illustration: Figure 2.2 shows a graph with two possible paths from A to B, let $S = \{S_2, S_3\}$. On the path $A \rightarrow S_1 \leftarrow S_3 \rightarrow S_2 \rightarrow B$, S_1 is the only collider, but S_1 is not in S. One could also argue that along the same path, S_2 as well as S_3 are no colliders and are in S. The path $A \leftarrow S_3 \rightarrow S_2 \rightarrow B$ has no colliders and S_2 and S_3 are in S. Hence all paths are blocked by nodes in S and S is said to d-separate A from B.

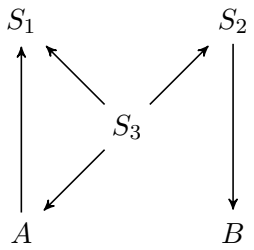

Figure 2.2: The set $S = \{S_2, S_3\}$ blocks every path from A to B and hence S is said to d-separate A from B.

Structural Equations

Let $A_1, \ldots, A_K$ be the nodes of a causal graph, then for each node A_i the value it takes a_i is determined by the values of its parents $pa(a_i)$ and given by a structural equation of the form

$$a_i = f_i(pa(a_i), u_i), \quad \text{for } i = 1, \ldots, K, \tag{2.1}$$

where u_i is a realisation of the error or disturbance U_i arising due to missing nodes in the graph. f_i is a deterministic function and if functions $f_1, \ldots, f_K$ given by (2.1) are autonomous, meaning that changes in f_i do not affect $f_1, \ldots, f_{i-1}, f_{i+1}, \ldots, f_K$, then they are called structural. As we will see later a causal graph is described by both, the structural equations and its joint probability.

2.2 Statistical Association and Causality

One of the main challenges in estimating causality between variables is to assume the correct model and to include all the necessary variables. The absence of a causal path between two variables is equivalent to the assumption that there is no causal effect of one variable on the other.

Example 2.1. As an illustration look at Figure 2.3 which shows the relationship between the number of sold ice-creams (I), the number of swimming pool drownings (S) and temperature (T) in a given city.
During summer when temperature is high, ice-cream consumption as well as swimming pool visits are at their peak. Furthermore, swimming pool drownings are more frequent when more people are exposed to that risk. Without the information on temperature one could be tempted to assume a causal effect of ice-cream consumption on drowning. Temperature T is said to be a *confounder* variable of both S and I. It is a common cause and there is no causal effect between ice-cream sales and swimming pool drownings, although they may be statistically correlated.

As illustrated in the previous example, an important issue in estimating causal effects is to assume the correct model in order to not draw wrong conclusions. Thus, for two variables all common causes have to be included in the model, this is equivalent to the

Figure 2.3: On the left: Incomplete graph only suggesting the statistical association between S and I. On the right: Graph representing the assumption made on causality, hence including the confounder T.

assumption that there is no unmeasured confounder. Assuming that the correct model was specified, to get an unbiased estimate of the causal effect, one may need to control for confounding variables. This issue will be discussed in Section 2.4. Note also that referring to structural equations, the no unmeasured confounder assumption will hold in a causal model given by (2.1) if the errors U_i are mutually independent (otherwise, they would represent unmeasured confounders).

2.3 Conditional Independence and Interventions

Nodes in a DAG can be seen as random variables taking values in a finite set. Let a be a realisation of the random variable A. If we assume that P is a probability measure and that we are in the discrete case, then by abuse of notation we write $\mathrm{P}(A = a) = \mathrm{P}(a)$. Conditional independence (definition 2.3) enables us to compute the joint distribution of a causal graph.

Definition 2.3. *Let A, B and S be some subsets in a DAG, P a joint probability. A and B are said to be conditionally independent given S if*

$$\mathrm{P}(a \,|\, b, s) = \mathrm{P}(a \,|\, s), \text{ whenever } \mathrm{P}(b, s) > 0. \tag{2.2}$$

Conditional independence may also be denoted by $A \perp\!\!\!\perp B \,|\, S$.

In a graph, the following situations can occur

- $A \longrightarrow B$: A and B are conditionally dependent given S ($A \not\perp\!\!\!\perp B \,|\, S$);
- $A \longrightarrow S \longleftarrow B$, A and B are independent $A \perp\!\!\!\perp B$, but given S, A and B are conditionally dependent ($A \not\perp\!\!\!\perp B \,|\, S$);
- $A \longrightarrow S \longrightarrow B$, A and B are dependent ($A \not\perp\!\!\!\perp B$), but given S become conditionally independent ($A \perp\!\!\!\perp B \,|\, S$).

Repeatedly using the equality $\mathrm{P}(X, Y) = \mathrm{P}(X \,|\, Y)\mathrm{P}(Y)$ (here, X and Y are events) for a set of variables $\{A_1, A_2, \ldots, A_K\}$ with realisations $a_1, \ldots, a_K$, results in a useful equation, namely

$$\mathrm{P}(a_1, \ldots, a_K) = \prod_{j=1}^{K} \mathrm{P}(a_j \,|\, a_1, \ldots, a_{j-1}), \tag{2.3}$$

where A_0 is the empty set, hence for $j = 1$ we have $\mathrm{P}(a_1 \,|\, a_0) = \mathrm{P}(a_1 \,|\, \emptyset) = \mathrm{P}(a_1)$.

For a causal DAG with nodes $A_1, \ldots, A_K$, some ordering among the variables can be assumed: for a given node A_j, the variables $A_1, \ldots A_{j-1}$ are the non-descendants of A_j reflecting the fact that A_j precedes its descendants. Let $pa(A_j)$ be the set of parents of A_j in a causal DAG. The joint distribution of the nodes $A_1, \ldots, A_K$ can then be factorised to

$$\mathrm{P}(a_1, \ldots, a_K) = \prod_{j=1}^{K} \mathrm{P}(a_j \,|\, pa(a_j)). \tag{2.4}$$

Equation (2.4) indicates that conditional on the parents of A_j, the non-descendants of A_j are independent of A_j (assuming the above mentioned ordering, i.e. that $A_1, \ldots A_{j-1}$ are all non-descendants).

There is a one-to-one correspondence between the set of conditional independencies, $A \perp\!\!\!\perp B \,|\, S$ in a DAG G implied by the recursive decomposition of equation (2.4) and the set of triples (A, S, B) in G that satisfy the d-separation criterion (definition 2.2) as long as $A \perp\!\!\!\perp B \,|\, S$ holds in all distributions compatible with G (see Appendix A) [11]. Figure 2.4 summarises which paths are blocked when conditioned on S (left hand side), while the right hand side presents a case which is blocked if not conditioned on S.

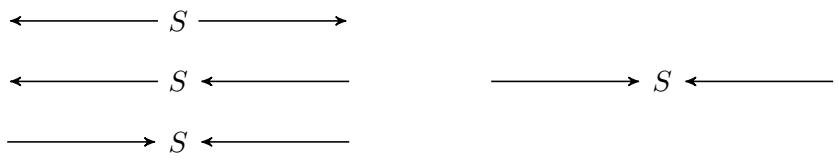

Figure 2.4: On the left: These paths are blocked if they are conditioned on S. On the right: This path is blocked if it is not conditioned on S.

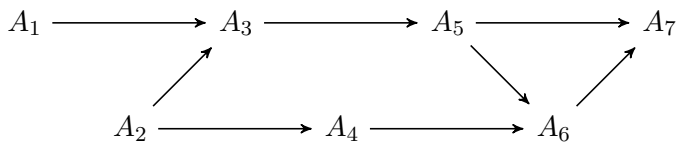

Figure 2.5: Causal DAG with nodes $A_1, \ldots, A_7$.

Illustration: The factorisation of the joint distribution of the nodes in the DAG corresponding to Figure 2.5 yields

$$\begin{aligned} \mathrm{P}(a_1, \ldots, a_7) \;=\; & \mathrm{P}(a_1)\mathrm{P}(a_2)\mathrm{P}(a_3 \,|\, a_1, a_2)\mathrm{P}(a_4 \,|\, a_2) \cdot \\ & \mathrm{P}(a_5 \,|\, a_3)\mathrm{P}(a_6 \,|\, a_4, a_5)\mathrm{P}(a_7 \,|\, a_5, a_6). \end{aligned} \tag{2.5}$$

Definition 2.4. *By intervening on a variable, the variable is forced to take a certain value.*

Given a causal DAG with nodes $A_1, \ldots, A_K$, referring to the notation of Pearl [11] intervening on A_i is denoted by $do(A_i = a_i)$ or $do(a_i)$. It follows that for A_i the conditional

probability then simplifies to

$$\mathrm{P}(a_i \,|\, pa(a_i)) = \begin{cases} 1, & \text{if } A_i = a_i \\ 0, & \text{if } A_i \neq a_i. \end{cases}$$

Thus, intervening on A_i transforms the original joint probability $\mathrm{P}(a_1, \ldots, a_K)$ into

$$\mathrm{P}(a_1, \ldots, a_K \,|\, do(a_i)) = \begin{cases} \prod_{j=1, j\neq i}^{K} \mathrm{P}(a_j \,|\, pa(a_j)), & \text{if } A_i = a_i \\ 0, & \text{if } A_i \neq a_i. \end{cases} \tag{2.6}$$

The intervention also modifies the corresponding causal graph by eliminating all the arrows entering that variable. Hence, the intervention only affects descendants of the designated variable. The concept of interventions leads to the definition of a causal effect given by Pearl [11]:

Definition 2.5. *Given two disjoint sets of variables, A and B, the causal effect of A on B, denoted by $P(b \,|\, do(a))$ is a function from A to the space of probability distributions on B.*

As we will see later, there are different measures of causal effects but all of them are based on the expression $\mathrm{P}(b \,|\, do(a))$.

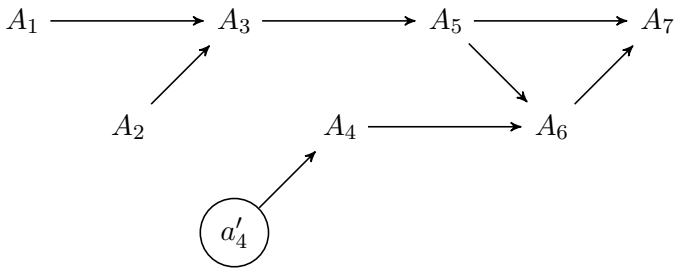

Figure 2.6: Causal graph modified by intervention on A_4.

Illustration: Getting back to the example presented in Figure 2.5, assume we intervene on A_4 and force it to equal a_4'. The resulting DAG (Figure 2.6) has no arrows between A_2 and A_4. Furthermore the joint distribution is modified:

$$\begin{aligned} \mathrm{P}(a_1, \ldots, a_7 \,|\, do(a_4')) \;=\; & \mathrm{P}(a_1)\mathrm{P}(a_2)\mathrm{P}(a_3 \,|\, a_1, a_2)\mathrm{P}(a_5 \,|\, a_3) \cdot \\ & \mathrm{P}(a_6 \,|\, a_4', a_5)\mathrm{P}(a_7 \,|\, a_5, a_6). \end{aligned} \tag{2.7}$$

2.4 Control for Confounding

Example 2.2. In the second example we assume a situation where for each subject of the study population, the CD4 cell count C, the treatment status A and information on death Y are collected. The study population includes only HIV-infected patients. A and Y are binary variables, A equals 1 if the patient is on second-line treatment and equals

0 if he is on first-line treatment. Y is 1 if the patient died and 0 if he is alive. It is also assumed that C precedes A and that A precedes Y. Figure 2.7 illustrates the situation in a causal graph for a given subject. To estimate the causal effect of treatment A on the outcome Y, we have to adjust for the confounder C.

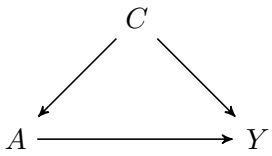

Figure 2.7: Causal graph corresponding to the relations between CD4 cell count (C), treatment (A) and death (Y).

The backdoor criterion. Let A, B and S be sets of variables in the same DAG. Suppose we are interested in the causal effect of A on B. There is a graphical criterion to test whether a set of variables S is sufficient for adjustment [10]. Adjusting for a set of confounders S means that the analysis can be restricted to strata of S within which there is no confounding of the treatment variable and the outcome.

Definition 2.6. *A set of variables S is said to satisfy the backdoor criterion relative to (A,B) if*

1. *no node in S is a descendant of A, and*
2. *S blocks every path between A and B which contains an arrow into A, hence every backdoor path from A to B.*

So adjusting for a set of variables satisfying the backdoor criterion yields an unbiased estimate of the causal effect of A on B.

Illustration: Getting back to the graph in Figure 2.2 and recalling that $S = \{S_2, S_3\}$ blocks the only present backdoor path between A and B implies that S satisfies the backdoor criterion.

Theorem 2.1. *If a set of variables S satisfies the backdoor criterion relative to (A,B), then the causal effect of A on B is identifiable and is given by the formula*

$$\mathrm{P}(B = b \,|\, do(a)) = \sum_s \mathrm{P}(B = b \,|\, A = a, S = s)\mathrm{P}(S = s). \tag{2.8}$$

Identifiability means that $\mathrm{P}(B = b \,|\, do(a))$ can be estimated consistently from an arbitrarily large sample randomly drawn from the joint distribution [10].

Assuming that a_0 and a_1 are two possible treatments and that Y is dichotomous, the causal effect of A on Y can be assessed by

- the odds ratio corresponding to $\frac{\mathrm{P}(Y=1 \,|\, do(a_1))}{1-\mathrm{P}(Y=1 \,|\, do(a_1))} \cdot \frac{1-\mathrm{P}(Y=1 \,|\, do(a_0))}{\mathrm{P}(Y=1 \,|\, do(a_0))}$;

- the risk ratio corresponding to $\mathrm{P}(Y = 1 \,|\, do(a_1))/\mathrm{P}(Y = 1 \,|\, do(a_0))$;
- the risk difference corresponding to $\mathrm{P}(Y = 1 \,|\, do(a_1)) - \mathrm{P}(Y = 1 \,|\, do(a_0))$.

If the odds ratio or the risk ratio differs from 1 then it is said that there is a causal effect of exposure A on the outcome Y. The same holds if the risk difference is different from 0.

Example 2.2 (continuation). Suppose we are in the setting of Example 2.2 with the corresponding DAG presented in Figure 2.7. Applying the backdoor criterion with $S = C$ shows that C satisfies the criterion relative to A and Y. The study population is divided into two strata according to their CD4 cell count: $C = 0$ if C is low, $C = 1$ if C is high. Recall that also A and Y are binary variables indicating treatment switching and death respectively. Table 2.1 presents some invented data to illustrate the calculation procedure. By Theorem 2.1 we have,

$$\begin{aligned} \mathrm{P}(Y = y \,|\, do(a)) &= \mathrm{P}(Y = y \,|\, A = a, C = 1)\mathrm{P}(C = 1) \\ &\quad + \mathrm{P}(Y = y \,|\, A = a, C = 0)\mathrm{P}(C = 0). \end{aligned} \tag{2.9}$$

As we have seen, due to confounding, the marginal association between A and Y is not causation, meaning that we cannot estimate the causal effect of A on Y without bias: $\mathrm{P}(Y \,|\, do(a)) \neq \mathrm{P}(Y \,|\, A = a)$. However within strata of C, association is causation, $\mathrm{P}(Y \,|\, do(a), C = c) = \mathrm{P}(Y \,|\, A = a, C = c)$, reflecting the fact that adjustment for C is sufficient to get a consistent estimate of the causal effect.

	C = 0		C = 1	
	A = 1	A = 0	A = 1	A = 0
Y = 1	1	5	7	1
Y = 0	11	3	15	3
TOTAL	12	8	22	4

Table 2.1: Invented data for Example 2.2, the total number of observed individuals equals 46.

By looking at Table 2.1 we can compute the causal effect of A on death Y:

$$\begin{aligned} \mathrm{P}(Y = 1 \,|\, do(A = 0)) &= \mathrm{P}(Y = 1 \,|\, A = 0, C = 1)\mathrm{P}(C = 1) \\ &\quad + \mathrm{P}(Y = 1 \,|\, A = 0, C = 0)\mathrm{P}(C = 0) \\ &= \frac{1}{4} \cdot \frac{26}{46} + \frac{5}{8} \cdot \frac{20}{46} \\ &= 0.413, \end{aligned} \tag{2.10}$$

and

$$\begin{aligned} \mathrm{P}(Y = 1 \,|\, do(A = 1)) &= \mathrm{P}(Y = 1 \,|\, A = 1, C = 1)\mathrm{P}(C = 1) \\ &\quad + \mathrm{P}(Y = 1 \,|\, A = 1, C = 0)\mathrm{P}(C = 0) \\ &= \frac{7}{22} \cdot \frac{26}{46} + \frac{1}{12} \cdot \frac{20}{46} \\ &= 0.216. \end{aligned} \tag{2.11}$$

Hence in this example the risk difference equals

$$\mathrm{P}(Y = 1 \,|\, do(A = 0)) - \mathrm{P}(Y = 1 \,|\, do(A = 1)) = 0.197. \tag{2.12}$$

So treatment $A = 0$ increases the probability of dying compared to treatment $A = 1$ and there is a causal effect of treatment on death. Note that the risk difference without adjusting for C gives a different result:

$$\begin{aligned} \mathrm{P}(Y = 1 \,|\, A = 0) - \mathrm{P}(Y = 1 \,|\, A = 1) &= \frac{\mathrm{P}(Y = 1, A = 0)}{\mathrm{P}(A = 0)} - \frac{\mathrm{P}(Y = 1, A = 1)}{\mathrm{P}(A = 1)} \\ &= \frac{6}{46} \cdot \frac{46}{12} - \frac{8}{46} \cdot \frac{46}{34} \\ &= 0.265, \end{aligned} \tag{2.13}$$

underlining the need for adjustment.

Chapter 3

The G-computation Procedure

In the previous chapter it was argued that given a DAG, a causal effect can be estimated without bias if there is no unmeasured confounder and if a set of variables is sufficient for adjustment, i.e. satisfying the backdoor criterion. The question is, what happens if we fail in finding a set of variables sufficient for adjustment? This could be the case in longitudinal observational studies, where for each subject data is collected over several time points. The mentioned problem setting can add another subtlety to the analysis: time-dependent confounder of treatment and outcome which is itself affected by previous treatment. In the present chapter G-computation is introduced as a method to cope with this type of problem in estimating causal effects.

3.1 Motivation

The motivation for using G-computation arises from the fact that for longitudinal observational studies standard methods to control for confounding as described in Chapter 2 may fail. To illustrate such a situation Example 3.1 is introduced.

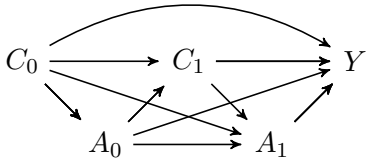

Figure 3.1: Causal graph corresponding to Example 3.1 and representing the relations between CD4 cell count (C_t), treatment (A_t) and death (Y) over time $t = 0, 1$.

Example 3.1. In this example we assume an observational study over two time points ($t = 0, 1$) represented by the causal graph in Figure 3.1. C_0, C_1 and A_0, A_1 represent CD4 cell count and treatment switching respectively at given time points. Y designates the information on death and is collected only at the end of follow-up. By looking at the corresponding causal graph we see that the set of time-varying confounders $S = \{C_0, C_1\}$ does not satisfy the backdoor criterion since the variables in S are also caused by treatment

variables contradicting condition 1 of the backdoor criterion (definition 2.6). To get an unbiased estimate of the causal effect of treatment on death some other method has to be applied.

As the previous example demonstrates, in presence of time-varying confounding there is an urge to use other estimation procedures than the standard methods explained in Chapter 2 to estimate the causal effect of exposure on outcome. As we will see, G-computation handles this type of setting.

3.2 Basic Idea of G-computation

The formulae introduced in Chapter 2 can be extended to the continuous case. Given the DAG in Figure 3.2 assumed to satisfy the no unmeasured confounder assumption, let A_t be the exposure variable, C_t the time-varying confounder and Y_t the outcome at time t ($t = 0, 1, \ldots, T$). It should be noticed here, that a DAG such as in Figure 3.2 could be interpreted also as non exhaustive, meaning that for legibility purpose not all arrows are included in the graph: whenever there is a causal path between two variables, there is also a direct arrow between them which is however not shown in the graph. This should be kept in mind when writing down the joint density.

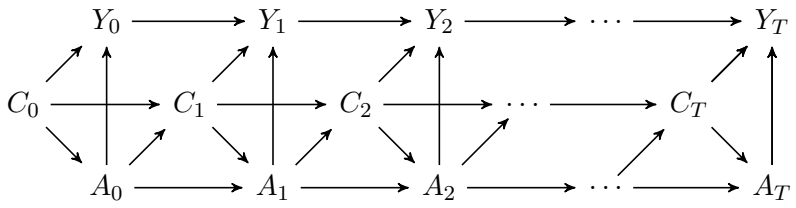

Figure 3.2: Causal graph representing the relations between exposure A_t, time-varying confounder C_t and outcome Y_t over time $t = 0, \ldots, T$.

Let $\bar{A}_t = \{A_0, A_1, \ldots, A_t\}$ denote the history of A_t up to time t. Under the no unmeasured confounder assumption, the joint density corresponding to the causal graph in Figure 3.2 is given by

$$\begin{aligned} f(\bar{c}_T, \bar{y}_T, \bar{a}_T) \quad = \quad & f(c_0) \cdots f(c_T \,|\, \bar{c}_{T-1}, \bar{a}_{T-1}) \cdot \\ & f(a_0 \,|\, c_0) \cdots f(a_T \,|\, \bar{c}_T, \bar{a}_{T-1}) \cdot \\ & f(y_0 \,|\, \bar{c}_0, \bar{a}_0) \cdots f(y_T \,|\, \bar{c}_T, \bar{a}_T, \bar{y}_{T-1}). \end{aligned} \tag{3.1}$$

similar to equation (2.4). Intervening on a set of variables $\bar{A}_T$ alters the joint density as shown in equation (2.6) and yields

$$\begin{aligned} f(\bar{c}_T, \bar{y}_T \,|\, do(\bar{A}_T = \bar{a}'_T)) \quad = \quad & f(c_0) \cdots f(c_T \,|\, \bar{c}_{T-1}, \bar{a}'_{T-1}) \cdot \\ & f(y_0 \,|\, \bar{c}_0, \bar{a}'_0) \cdots f(y_T \,|\, \bar{c}_T, \bar{a}'_T, \bar{y}_{T-1}). \end{aligned} \tag{3.2}$$

To obtain the causal effect of intervention $\bar{A}_T = \{a'_0, \ldots, a'_T\} = \bar{a}'_T$ on $\bar{Y}_T$, equation (3.2)

is integrated over all possible values of $\bar{c}_T$,

$$\begin{aligned} f(\bar{y}_T \,|\, do(\bar{A}_T = \bar{a}'_T)) &= \int f(\bar{c}_T, \bar{y}_T \,|\, do(\bar{A}_T = \bar{a}'_T)) d\mu(\bar{c}_T) \\ &= \int \prod_{t=0}^{T} f(c_t \,|\, \bar{c}_{t-1}, \bar{a}'_{t-1}) f(y_t \,|\, \bar{c}_t, \bar{a}'_t, \bar{y}_{t-1}) d\mu(\bar{c}_T), \end{aligned} \tag{3.3}$$

where $d\mu(\bar{c}_T)$ represents integration (or summation if $\bar{c}_T$ is discrete) over all possible $\bar{c}_T$ histories. This formula is referred to as the g-computation algorithm formula [12]. Note that the last expression in equation (3.3) is only based on observed data and not on the intervention. The goal of G-computation is now to estimate $f(c_t \,|\, \bar{c}_{t-1}, \bar{a}'_{t-1})$ and $f(y_t \,|\, \bar{c}_t, \bar{a}'_t, \bar{y}_{t-1})$ based on the observed data and to evaluate these estimates at the intervention $\bar{A}_T = \bar{a}'_T$ to obtain an expression for $f(\bar{y}_T \,|\, do(\bar{A}_T = \bar{a}'_T))$. The simple example presented in Section 3.4.1 allows to directly calculate an estimate of (3.3), but in more complex applications the direct calculation is computationally infeasible, so the result is approximated by Monte Carlo simulation. This is done by generating for a given intervention a simulated population in which the joint density of $Y_0, \ldots, Y_T$ and $C_0, \ldots, C_T$ is approximately equal to equation (3.3) [6].

3.3 The Estimation Procedure

Having presented the basic idea of G-computation and assuming the DAG in Figure 3.2, we will now explain the two main steps in detail, namely:

- modelling relationships between variables;
- using these models to simulate new datasets under various treatment interventions.

In the second step we simulate what happens to the subjects in the study if treatment is set to a certain value (i.e. we intervene on treatment). The new datasets enable us to estimate causal effects of different treatment interventions.

3.3.1 Modelling relationships between observed variables

In the first step of G-computation the goal is to model the relationships between the observed variables $C_1, \ldots, C_T, Y_0, \ldots, Y_T$ and their parents. We denote the conditional densities (continuous case) or distributions (discrete case) of C_t and Y_t by $f(c_t \,|\, pa(c_t))$ and $f(y_t \,|\, pa(y_t))$ respectively. Note that the conditional densities of $A_1, \ldots, A_T$ need not to be modelled since these variables are set to a certain value by intervention.

Starting at $t = 0$, a model has to be specified for $f(y_0 \,|\, c_0, a_0)$. If there are time-fixed confounders or baseline variables, they are included in C_0. Continuing at $t = 1$, we specify models for $f(c_1 \,|\, c_0, a_0)$ and $f(y_1 \,|\, \bar{c}_1, \bar{a}_1, y_0)$. Repeating this step for each $t \in [2, T]$ models for $f(c_t \,|\, \bar{c}_{t-1}, \bar{a}_{t-1})$ and $f(y_t \,|\, \bar{c}_t, \bar{a}_t, \bar{y}_{t-1})$ are obtained from observational data. If Y_t denotes death, then the model for $f(y_t \,|\, \bar{c}_t, \bar{a}_t)$ is fitted to the subset of data with $Y_{t-1} = 0$.

The models can be fitted separately at each time point (over all subjects) or pooled over all time points (and all subjects).

3.3.2 Simulating a new dataset under intervention

Having specified models for the relations between the variables in the DAG we are now able to simulate new datasets under various treatment interventions. We denote by $C_0^*, \ldots, C_T^*, Y_0^*, \ldots, Y_T^*$ the simulated variables. The treatment intervention being considered is $\bar{A}_T = \bar{a}'_T = \{a'_0, \ldots, a'_T\}$.

C_0 is not affected by any variable, so $C_0^* = c_0^* = c_0$ as in the observational data. Y_0^* is simulated from the previously fitted distribution $f(y_0 \mid c_0^*, a'_0)$. Then C_1^* is simulated from the distribution $f(c_1 \mid c_0^*, a'_0)$. Similarly we simulate Y_1^* from $f(y_1 \mid \bar{c}_1^*, \bar{a}'_1, y_0^*)$. This procedure is repeated until a complete new dataset is generated. In the case where the outcome is death, if for a given patient $Y_{t-1}^* = 1$ then no further values have to be simulated.

Note that having simulated the dataset under intervention we are able to compute an estimate of the causal effect, e.g. the incidence rate being equal to the number of incidences (i.e. $Y_t = 1$) divided by the total follow-up time (sum over all patients) in the study. The incidence rate of one intervention can then be compared to the incidence rate of another or of no intervention.

3.3.3 Statistical inference

To make inference on the causal effect estimate, the bootstrap is used. The bootstrap consists of generating B datasets from the original dataset by sampling with replacement from the study population. The B generated datasets are then used to compute mean estimate, standard error and confidence interval. In this work confidence intervals are based on the standard normal distribution. Note that depending on the size of B the bootstrap can be very time-consuming.

3.3.4 Some remarks

Dynamic regimes

In this section only static treatment regimes were used for intervention. Nevertheless, new datasets can also be simulated under dynamic treatment interventions. Such a treatment regime could be for instance “Switching treatment if the CD4 cell count is below a certain threshold” but it could also assign treatment switching with a certain probability to each patient.

Loss to follow-up

G-computation allows to handle the problem of loss to follow-up, i.e. when subjects drop out of the study before the end of follow-up. If dropping out can be assumed to occur at random, then it can be considered as an additional treatment trajectory and simulations of new datasets can be done under the intervention “getting treatment $\bar{a}'_T$

and never dropping out". The missing at random assumption is then implicit in the no unmeasured confounder assumption [7].

Competing risks

If there exists a competing risk to the outcome, it can be included in the model as an additional outcome. If so, the competing risk would also be a variable in the DAG and its relation to other variables should be modelled in step one of G-computation. In step two, the competing risk is then simulated in the same way as the outcome.

3.4 Examples

3.4.1 Fixed time-point example

Let us recall Example 2.2 from Section 2.4. Table 3.1 presents the corresponding contingency table, whereas the causal graph is shown in Figure 3.3.

	C = 0		C = 1	
	A = 1	A = 0	A = 1	A = 0
Y = 1	1	5	7	1
Y = 0	11	3	15	3
TOTAL	12	8	22	4

Table 3.1: Invented data for Example 2.2, the total number of observed individuals equals 46.

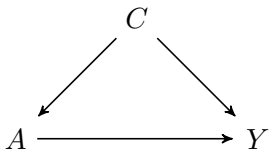

Figure 3.3: Causal graph corresponding to the relations between CD4 cell count (C), treatment (A) and death (Y).

As already described in the previous chapter, we are interested in estimating the causal effect of treatment (A) on death (Y). The joint distribution corresponding to the DAG in Figure 3.3 equals (by equation 2.4):

$$\mathrm{P}(y, a, c) \quad = \quad \mathrm{P}(y \,|\, a, c)\mathrm{P}(a \,|\, c)\mathrm{P}(c). \tag{3.4}$$

The causal effect of intervention $A = a$ on Y is given by

$$\mathrm{P}(y \,|\, do(a)) \quad = \quad \sum_{c} \mathrm{P}(y \,|\, a, c)\mathrm{P}(c). \tag{3.5}$$

We will now compute the risk difference $\mathrm{P}(Y = 1|do(A = 0)) - \mathrm{P}(Y = 1|do(A = 1))$ by using G-computation. Note that in this example, the formula in (3.3) corresponding to

equation (3.5) can be computed explicitly and hence the estimation procedure described in Section 3.3 is not necessary. First, we need to specify models for $\mathrm{P}(c)$ and $\mathrm{P}(y \mid a, c)$. We assume that C is a Bernoulli random variable with $\mathrm{P}(C = 1) = 26/46 = 0.565$ (see Table 3.1) and that $Y \mid A, C$ is also a Bernoulli random variable with

$$\log\left(\frac{\mathrm{P}(Y = 1 \mid A = a, C = c)}{1 - \mathrm{P}(Y = 1 \mid A = a, C = c)}\right) = \alpha + \beta a + \gamma c + \delta ac. \tag{3.6}$$

A logistic regression is fitted in R to estimate the parameters α, β, γ and δ which yields the (abbreviated) R-Output:

```
> fitY <- glm(Y~A*C,family="binomial",data=data)
> summary(fitY)

Call:
glm(formula = Y ~ A * C, family = "binomial", data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4006  -0.8752  -0.4172   0.9695   2.2293

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5108     0.7303   0.699   0.4843
A            -2.9087     1.2745  -2.282   0.0225 *
C            -1.6094     1.3663  -1.178   0.2388
A:C           3.2452     1.7796   1.824   0.0682 .

    Null deviance: 56.534  on 45  degrees of freedom
Residual deviance: 49.489  on 42  degrees of freedom
AIC: 57.489
```

Using the estimates $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ and $\hat{\delta}$ for the model of Y we can now compute the probabilities of $Y \mid A, C$:

$$\begin{aligned}
\mathrm{P}(Y = 1 \mid A = 1, C = 1) &= \frac{\exp(\hat{\alpha} + \hat{\beta} + \hat{\gamma} + \hat{\delta})}{(1 + \exp(\hat{\alpha} + \hat{\beta} + \hat{\gamma} + \hat{\delta}))} = 0.318, \\
\mathrm{P}(Y = 1 \mid A = 1, C = 0) &= \frac{\exp(\hat{\alpha} + \hat{\beta})}{(1 + \exp(\hat{\alpha} + \hat{\beta}))} = 0.083, \\
\mathrm{P}(Y = 1 \mid A = 0, C = 1) &= \frac{\exp(\hat{\alpha} + \hat{\gamma})}{(1 + \exp(\hat{\alpha} + \hat{\gamma}))} = 0.250, \\
\mathrm{P}(Y = 1 \mid A = 0, C = 0) &= \frac{\exp(\hat{\alpha})}{(1 + \exp(\hat{\alpha}))} = 0.625.
\end{aligned}$$

Yielding an estimate for the risk difference

$$\begin{aligned} \mathrm{P}(Y=1|do(A=0)) - \mathrm{P}(Y=1|do(A=1)) &= 0.25 \cdot 0.565 + 0.625 \cdot 0.435 \\ &\quad -(0.318 \cdot 0.565 + 0.083 \cdot 0.435) \\ &= 0.197, \end{aligned}$$

as found in (2.12). Due to the simplicity of this example we were able to compute equation (3.5) explicitly. Table 3.2 presents the results obtained by simulating new datasets under the two interventions by using the above fitted model. The dataset includes 46 persons and each simulation was bootstrapped 10000 times. For the R-code see Appendix B).

	BS average	BS std. error	normal based 95% CI
risk difference	0.1971	0.0999	$[0.0015, 0.3927]$

Table 3.2: G-computation estimates for risk difference between intervention $A = 0$ and $A = 1$.

3.4.2 Two time-point example

Returning to Example 3.1 from Section 3.1, we recall the causal graph presented in Figure 3.4. C_0 and C_1 correspond to the logarithm of the CD4 cell count and A_0, A_1 are indicators for treatment switching, i.e. $A_t = 1$ if patient was switched from first-line treatment to second-line treatment at time t. Note that once a patient is switched he remains on second-line treatment.

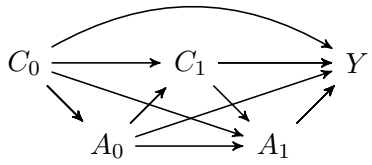

Figure 3.4: Causal graph corresponding to Example 3.1 and representing the relations between CD4 cell count (C_t), treatment (A_t) and death (Y) over time $t = 0, 1$.

To illustrate the G-computation procedure, a hypothetical dataset with 10000 patients is generated according to the following scheme:

- U is a normal random variable with mean 0 and variance 0.25;
- C_0 is a normal random variable with mean $5.5 + U$ and variance 0.4.
- A_0 is a Bernoulli random variable with

$$\mathrm{P}(A_0 = 1 \,|\, C_0 = c_0) = \frac{\exp(5 - c_0)}{1 + \exp(5 - c_0)};$$

- C_1 is generated from a normal distribution with mean $0.9C_0 + A_0 + 0.1U$ and variance 0.1;

- if $A_0 = 1$ then $A_1 = 1$ and otherwise, A_1 is generated from a Bernoulli distribution with

$$\mathrm{P}(A_1 = 1 \,|\, A_0 = a_0, C_0 = c_0, C_1 = c_1) = \frac{\exp(-7 + 0.1a_0 + c_1 + 0.5c_0)}{1 + \exp(-7 + 0.1a_0 + c_1 + 0.5c_0)};$$

- finally, Y is generated from a Bernoulli distribution with $p_{Y,1} = \mathrm{P}(Y = 1 \,|\, A_0 = a_0, A_1 = a_1, C_0 = c_0, C_1 = c_1, U = u)$,

$$p_{Y,1} = \frac{\exp(-16 + 1.2c_0 + 1.8c_1 - a_0 - a_1 - u)}{1 + \exp(-16 + 1.2c_0 + 1.8c_1 - a_0 - a_1 - u)}.$$

Starting from Figure 3.4, the corresponding joint density is given by

$$f(a_0, a_1, c_0, c_1, y) = f(a_0 \,|\, c_0) f(a_1 \,|\, a_0, c_0, c_1) f(c_0) f(c_1 \,|\, a_0, c_0) f(y \,|\, a_0, a_1, c_0, c_1), \quad (3.7)$$

and the causal effect of intervention $\bar{A}_1 = \{a_0', a_1'\} = \bar{a}_1'$ on Y yields

$$f(y \,|\, do(\bar{A}_1 = \bar{a}_1')) = \int\int f(c_0) f(c_1 \,|\, a_0', c_0) f(y \,|\, a_0', a_1', c_0, c_1) d\mu(c_1) d\mu(c_0), \quad (3.8)$$

if $a_0 = a_0'$ and $a_1 = a_1'$. Leading to

$$\mathrm{P}(y \,|\, do(\bar{A}_1 = \bar{a}_1')) = \int\int f(c_0) f(c_1 \,|\, a_0', c_0) \mathrm{P}(y \,|\, a_0', a_1', c_0, c_1) d\mu(c_1) d\mu(c_0). \quad (3.9)$$

In the first step of G-computation we have to specify models for the distribution of C_1 given C_0 and A_0 and for the distribution of Y given C_0, C_1, A_0, A_1. In this case, we fit a linear regression to C_1 with predictors C_0 and A_0 and a logistic regression to $Y \,|\, C_0, C_1, A_0, A_1$. New datasets sets are generated using the previously estimated model parameters and by considering a specific treatment intervention $\{A_0, A_1\} = \{a_0', a_1'\}$:

- take $C_0 = c_0$ from the original dataset,
- intervene on treatment, i.e. set treatment $\{A_0, A_1\}$ to $\{a_0', a_1'\}$,
- simulate C_1 given $C_0 = c_0, A_0 = a_0'$ by using the above fitted model,
- then simulate the outcome Y given $C_0 = c_0, C_1 = c_1, A_0 = a_0', A_1 = a_1'$, again by using the fitted model from the original dataset.

The simulation is done for each patient resulting in a dataset reflecting what would have been observed if for each patient treatment had been set to $\{a_0', a_1'\}$. Denote by p_0 the risk of death for intervention "never on second-line treatment"and by p_1 the risk of death for intervention "always on second-line treatment". The death risk in each simulated dataset corresponds to equation (3.9) with $y = 1$ and is estimated by computing the number of patients who died divided by the total number of patients, i.e. by simply computing the mean of the death variable in the new dataset. To compare the two treatment interventions, the odds ratio $(p_0(1-p_1)/((1-p_0)p_1))$, the risk ratio (p_0/p_1) and the risk difference $(p_0 - p_1)$ are calculated.

The analysis was conducted by generating 10000 bootstrap (BS) samples for each intervention yielding the results presented in Table 3.3 comparing interventions "always on second-line treatment" ($\{A_0, A_1\} = \{1, 1\}$) to "never on second-line treatment" ($\{A_0, A_1\} = \{0, 0\}$). There is only a small difference between the effect of no treatment and the effect of always treatment. Nevertheless the values indicate that treatment switching has a beneficial effect on survival. The R-Code for this example can be found in Appendix B.

	BS average	BS std. error	normal based 95% CI
risk difference	0.0394	0.0060	$[0.0275, 0.0512]$
risk ratio	1.1085	0.0176	$[1.0741, 1.1429]$
odds ratio	1.1818	0.0303	$[1.1225, 1.2411]$

Table 3.3: G-computation estimates for risk difference, risk ratio and odds ratio between intervention $\{A_0, A_1\} = \{0, 0\}$ and $\{A_0, A_1\} = \{1, 1\}$.

3.5 Drawbacks

In the absence of unmeasured confounders and model misspecification, G-computation yields a consistent estimate of the causal effect. It is worth noticing that the no unmeasured confounder assumption is untestable and will never be exactly true. To see whether there is model misspecification the causal effect of no intervention can be estimated by simulating the observed treatment regime and comparing the resulting estimate to the observed value (computed by using the observed dataset). If the estimate differs from the observed value, it is an indicator that there is model misspecification. Nevertheless, if the two values are similar, no such statement can be made. If it seems that important confounders are missing in the model, then the result obtained from G-computation is not expected to be similar to an experimental study.

Another drawback of G-computation in contrary to other methods, is the existence of the null-paradox in the presence of time-varying confounder. Suppose that $\mathrm{P}(Y \mid do(\bar{A} = \bar{a})) = \mathrm{P}(Y \mid do(\bar{A} = \bar{0}))$ for all possible interventions $\bar{a}$, where $\bar{0} = \{0, \ldots, 0\}$. This hypothesis is referred to as the sharp null hypothesis. Equation (3.3) would then be independent of $\bar{a}$. Suppose also, as will nearly always be the case, that pretreatment unmeasured health status is a causal determinant of both confounder and outcome. Then with probability going to 1, the estimator of the causal effect of treatment on outcome as in (3.3) will depend on $\bar{a}$, thus the sharp null hypothesis will be rejected for most choices of parametric models [13].

3.6 Reference to Counterfactuals

This work is based on the concept of interventions. In the literature however, G-computation is mostly related to the notion of counterfactuals [14], relying on Robins who first introduced G-computation in 1986 [12]. From a counterfactual point of view one asks what-if questions such as "what would have happened to patient i if he had instantly

switched treatment?" whereas from an intervention point of view one would ask "if patient i switches treatment now, will he have a longer live?". It is clear that in experimental studies interventions can be carried out explicitly, which is not the case for observational studies. In observational studies there is no mean to see "what happens if..." but we can only simulate "what would have happened if...". Keeping this in mind, we conclude that using intervention in the observational setting implicitly contains the counterfactual idea. So both concepts are the same for the issue discussed here and for reasons of clarity and simplicity we chose to rely only on interventions [11].

3.7 An Alternative to G-computation

An alternative to G-computation is provided by estimating causal effects using Inverse Probability of Treatment Weighted (IPTW). This method weighs each observation by its inverse probability of treatment, yielding an unconfounded pseudo-population enabling to consistently estimate causal effects by standard methods. An application of this method can be found in *Marginal Structural Models and Causal Inference* by Peter [8].

Chapter 4

Data Analysis

In this chapter we will explain how the data provided by the Institute of Social and Preventive Medicine in Bern (ISPM) and IeDEA (International epidemiologic Databases to Evaluate AIDS) are analysed using G-computation (introduced in Chapter 3). The goal is to clarify whether or not, after immunological failure, treatment switching is beneficial in terms of lifetime for HIV-patients in Southern Africa. Two treatment interventions, "always on first-line treatment and no loss to follow-up" and "always on second-line treatment and no loss to follow-up" are simulated using G-computation allowing to answer the above stated question.

4.1 Data Description

The dataset provided by ISPM and IeDEA consists of a longitudinal study of HIV-infected patients in Southern Africa (Zambia and Malawi) over the period of August 2004 until August 2009. Patients included in the study are all aged over 16 and experienced immunological failure as described in Chapter 1. In total there are 2648 patients participating, although there was one patient removed from the dataset (see next section).

For each patient, data is collected over several points in time, with quarters being the time unit. The total follow-up time is 17255 person-time, the maximal individual follow-up time equals 20 quarters ($T = 20$) and the average follow-up time is 6.5 quarters. A patient is considered to be lost to follow-up if he did not return to the clinic for twelve months and right censored if he is neither lost to follow-up nor did he die. Since we deal with an ongoing study, it is the closing date of August 2009 that causes right censoring. There are 178 patients who are lost to follow-up, 76 patients who died and 2393 patients who are right censored.

The dataset contains information on the following baseline variables (not evolving over time),

- age group with three categories: lower than 30, between 30 and 39, greater or equal to 40;
- CD4 group with four CD4 cell count categories: lower than 50, between 50 and 99, between 100 and 199, greater or equal to 200;

- gender, a binary variable with female = 1, male = 0;
- clinical stage, another binary variable with advanced stage = 1 and less advanced stage = 0. Advanced stage corresponds to stage III and IV whereas less advanced refers to stage I and II [9] [3].

In the analysis of the dataset, we use the logarithm of the CD4 cell count, the treatment variable (1 if switched to second-line, 0 if not) and information on death (1 if patient died, 0 if not). Other variables were created relying on existing ones which will be explained in Section 4.3.1. A total of 365 patients (13.8%) were switched to second-line treatment, four of them died. Table 4.1 outlines the distribution of death across baseline variables and treatment.

	total	**deaths**	**% deaths**
total	2647	76	2.87
male	1240	37	2.98
female	1407	39	2.77
less adv. stage	711	9	1.27
adv. stage	1936	67	3.46
age ≤ 30	1744	50	2.87
$30 <$ age ≤ 39	744	22	2.96
age ≥ 40	159	4	2.52
CD4 ≤ 50	357	18	5.04
$50 <$ CD4 ≤ 99	633	35	5.53
$100 \leq$ CD4 ≤ 199	840	16	1.90
$200 \leq$ CD4	817	7	0.86
switching to 2nd-line	365	4	1.10
no switching	2282	72	3.16

Table 4.1: Total number of patients and number of deaths in each category of baseline and treatment variables.

4.2 Imputation of Missing Values

Before analysing the observed data using G-computation, our attention turns to the 11049 missing CD4 cell counts in the dataset. 64% of all CD4 values (over time and subjects) are missing. These values are imputed using Multivariate Imputation by Chained Equations (MICE) [15] in R. Before imputation, the CD4 cell counts are transformed into their logarithms and imputation is done on the logarithm scale by predictive mean matching. Since there is one patient who has a CD4 baseline value equal to 0 (not compatible with the logarithm) this patient is omitted from the analysis, resulting in a study population of size 2647.

4.2.1 Predictive mean matching

Imputation is done in R using the mice-function in combination with predictive mean matching. The method is based on Rubin [16] p.166-168 and shall be explained in this section.

Let $Y = \{Y_1, \ldots, Y_n\}$ denote the variable containing the missing values, in our case the log CD4 counts. The model assumed for Y_i is $Y_i \sim N(X_i\beta, \sigma^2)$, where X_i are the q predictors of Y_i. We used all the variables in the dataset as predictors for log CD4. Let n_1 be the number of observed Y-values, n_0 the number of missing Y-values and obs and mis the sets of indices corresponding to the observed and missing values, respectively. Then β and σ can be estimated by

$$\hat{\beta}_1 = (\sum_{obs} X_i^t X_i)^{-1}(\sum_{obs} X_i^t Y_i), \tag{4.1}$$

and

$$\hat{\sigma}_1^2 = \frac{1}{n_1 - q}\sum_{obs}(Y_i - X_i\hat{\beta}_1)^2. \tag{4.2}$$

Estimation of the missing Y-values is done as follows:

1. Draw a $\chi^2_{n_1-q}$ random variable, say g, and let $\sigma_*^2 = \hat{\sigma}_1^2(n_1 - q)/g$.

2. Draw q independent $N(0,1)$ variables to create a q-component vector Z and let $\beta_* = \hat{\beta}_1 + \sigma_*(\sum_{obs} X_i^t X_i)^{1/2} Z$.

3. Calculate the n_0 predicted values in Y_{mis} as $Y_{i*} = X_i\beta_*$, $i \in mis$. Then, for each Y_i, $i \in mis$, find the value Y_i with $i \in obs$ closest to Y_{i*} and impute this value for Y_i.

To account for the uncertainty in the imputation, steps 1 to 3 are repeated m independent times. In our case, the procedure will be repeated $m = 5$ times resulting in five different data matrices differing by their imputed log CD4 cell counts. The analysis (G-computation) is done for all five datasets and their results are then combined following a scheme introduced by Rubin [16] and explained in the subsequent section.

4.2.2 Combining results of multiple imputation

Denote by $\hat{Q}_1, \ldots, \hat{Q}_5$ the estimates obtained in a given analysis and by $U_1, \ldots, U_5$ the corresponding variances of the estimates. In our case, $\hat{Q}$ is a measure of the causal effect of treatment switching on death. The overall estimate $\bar{Q}$ is then equal to

$$\bar{Q} = \frac{1}{5}\sum_{j=1}^{5} \hat{Q}_j. \tag{4.3}$$

To compute the overall standard error, we first have to calculate the within-imputation variance $\bar{U}$ and the between-imputation variance B:

$$\bar{U} = \frac{1}{5}\sum_{j=1}^{5} U_j, \tag{4.4}$$

$$B = \frac{1}{5-1}\sum_{j=1}^{5}(\hat{Q}_j - \bar{Q})^2. \tag{4.5}$$

The total variance is then equal to

$$T = \bar{U} + (1 + \frac{1}{5})B, \tag{4.6}$$

and hence the overall standard error is $\sqrt{T}$. Confidence intervals are obtained by computing $\bar{Q} \pm t_{\alpha,df}\sqrt{T}$, where t is a quantile of a Student's t-distribution with degrees of freedom equal to

$$df = (5-1)\left(1 + \frac{5\bar{U}}{(5+1)B}\right)^2. \tag{4.7}$$

To assess how strongly the estimate is influenced by missing data, Rubin suggests to compute the estimated rate of missing information:

$$\gamma = \frac{r + 2/(df+3)}{r+1}, \tag{4.8}$$

where

$$r = \frac{(1+1/5)B}{\bar{U}} \tag{4.9}$$

is the relative increase in variance due to nonresponse.

4.2.3 Alternative method for imputation of missing values

An additional dataset is analysed which was computed by imputation of missing values by last observation carried forward, meaning that if a CD4 value is missing it is imputed by the last observed value for that patient. By analysing this dataset we follow closely the work of Peters (Master's Thesis [8]) and Gsponer et al. [9] who analysed a very similar dataset using IPTW instead of G-computation. In the present work, however, we opted for another imputation method possibly yielding more accurate log CD4 values then those obtained by last observation carried forward especially in the light of 64% missing values.

4.3 G-computation

At the very beginning of our analysis we have to come up with a causal graph reflecting all the assumptions on causal influences in the data. The graph assumed in this analysis is shown in Figure 4.1, where

- BL denotes the baseline variables, namely baseline CD4 group, baseline age group, clinical stage and gender;
- C_0, A_0 and Y_0 are the baseline values for log CD4 cell count, treatment switching and death;
- C_t is the variable denoting the log CD4 cell count in quarter $(t-1, t]$ and if treatment was switched in this quarter then C_t corresponds to the last CD4 value measured before switching;
- A_t is a binary variable indicating whether or not treatment has been switched in $(t-1, t]$ ($A_t = 1$ if switched, $A_t = 0$ if not), once treatment is switched it remains switched, i.e. $A_j = 1$ whenever $j \geq t$ if $(t-1, t]$ is the switching quarter;
- Y_t denotes death and equals 1 if the patient died in quarter $(t-1, t]$. If so, values of other variables in that quarter correspond to measurements before death.

Note that such a graph can be drawn for each subject in the study.

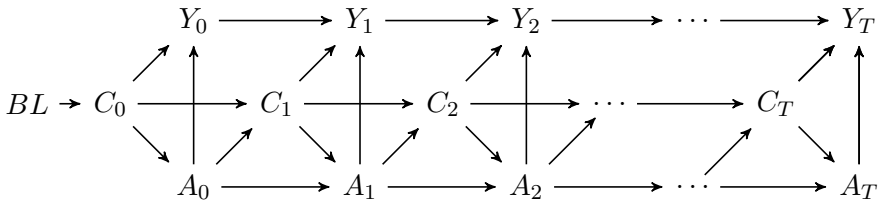

Figure 4.1: Causal graph representing the relations between treatment switching A_t, log CD4 cell count C_t and death Y_t over time $t = 0, \ldots, T$.

The joint density arising from the DAG in Figure 4.1 (assuming the absence of unmeasured confounders) is given by

$$\begin{aligned} f(\bar{c}_T, \bar{y}_T, \bar{a}_T) \quad = \quad & f(c_0) f(c_1 \,|\, c_0, a_0) \cdots f(c_T \,|\, \bar{c}_{T-1}, \bar{a}_{T-1}) \cdot \\ & f(a_0 \,|\, c_0) \cdots f(a_T \,|\, \bar{c}_T, \bar{a}_{T-1}) \cdot \\ & f(y_0 \,|\, c_0, a_0) \cdots f(y_T \,|\, \bar{c}_T, \bar{a}_T, y_{T-1}), \end{aligned} \tag{4.10}$$

and the causal effect of intervention $\bar{A}_t = \bar{a}'_T$ on death $\bar{Y}_T$ equals

$$f(\bar{y}_T \,|\, do(\bar{A}_T = \bar{a}'_T)) \quad = \quad \int \prod_{t=1}^{T} f(c_t \,|\, \bar{c}_{t-1}, \bar{a}'_{t-1}) f(y_t \,|\, \bar{c}_t, \bar{a}'_t, y_{t-1}) f(c_0) f(y_0 \,|\, c_0, a'_0) d\mu(\bar{c}_T), \tag{4.11}$$

where $\bar{y}_t = \{y_0, y_1, \ldots, y_t\}$, $\bar{a}_t = \{a_0, a_1, \ldots, a_t\}$ and $\bar{c}_t = \{c_0, c_1, \ldots, c_t\}$ and for notational purpose we assume that BL is included in C_0. The goal of the analysis is to compare the two interventions "always on first-line treatment and no loss to follow-up" and "always on second-line treatment and no loss to follow-up" and to assess if treatment switching has an effect in terms of lifetime. To estimate the causal effect we have to perform the two steps of G-computation: find models for the relationships in the DAG and simulate new datasets according to these models considering the above mentioned interventions.

4.3.1 Models for log CD4 and death

To estimate the causal effect of treatment on death given by equation (4.11), we have to find models for $C_1, \ldots, C_T, Y_1, \ldots, Y_T$ given their parents. The four baseline variables denoted by BL are included in each model. Relying on the existing variables, new variables are created, namely

- $Acum_t$, cumulative time on second-line treatment, and
- $nAcum_t$, cumulative time on first-line treatment.

We decided to fit models over all patients and all time points, assuming that for each point in time the same model holds. All the computations were done using the statistical software R.

Model for log CD4

Log CD4 (C_t) is assumed to depend linearly on its predictors, so a linear regression is fitted to log CD4 involving several predictor variables. We apply leave-one-out cross validation (on the patient level) and compute the mean squared error (MSE) for a total of 30 log CD4 models, all differing in their predictors. As possible predictors, lag one log CD4 (C_{t-1}), lag one cumulative time on first-line treatment ($nAcum_{t-1}$), lag one cumulative time on second-line treatment ($Acum_{t-1}$), time (t), a spline of time t ($ns(t, 5)$) and the baseline variables (BL) are used. We also allow for interaction between the baseline variables and the other predictors. This procedure is repeated for the five imputed datasets and the results are combined to yield an average MSE shown in Figure 4.2. The full results are provided in Appendix C.2. Model 28 has the smallest average MSE (0.31694312) and is thus chosen to be the best model for log CD4. It achieves not only the minimal average MSE over the five datasets but also corresponds to the minimal MSE for each imputed dataset. Consequently, the linear regression model used for G-computation is

$$C_t \sim C_{t-1} * BL + ns(t, 5) * BL + nAcum_{t-1} * BL, \tag{4.12}$$

where $*$ denotes the allowed interaction terms.

Model for death

Death (Y_t) is assumed to be a Bernoulli random variable and given the small number of death occurrence (76) compared to the size of the study population (2647), cross validation was considered to be infeasible. Instead, we seek for the best model using stepwise model selection based on the Akaike Information Criterion (AIC). The stepwise model selection is done using the R-function stepAIC and three different direction modes are used, backward, forward and both directions, allowing for more than one model to have minimal (and equal) AIC. The smallest model used in stepwise model selection includes only the baseline variables as predictors, whereas in addition to the baseline variables, the biggest model includes $C_t, C_{t-1}, nAcum_t, Acum_t$, time t and a spline of time, $ns(t, 5)$. Adding interaction terms to the model did not give valid results for stepwise selection, thus only first order terms were considered. Having performed stepwise model selection on the five imputed

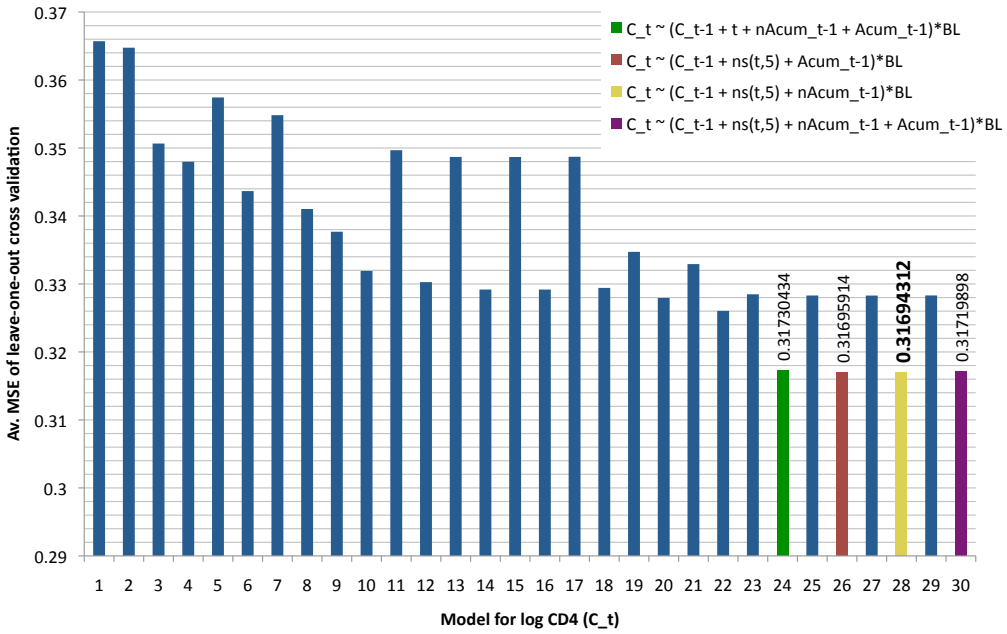

Figure 4.2: Average mean squared error for different linear regression models resulting from leave-one-out cross validation applied to the five imputed datasets.

datasets, we retain the best model as the one which performs best for the majority of the datasets. Figure 4.3 provides the models with the smallest AIC values for each of the five datasets, where

- Model 1 corresponds to the logistic regression of $Y_t \sim C_t + Acum_t + t + BL$,
- Model 2 corresponds to the logistic regression of $Y_t \sim C_t + Acum_t + nAcum_t + BL$,
- Model 3 corresponds to the logistic regression of $Y_t \sim C_{t-1} + C_t + Acum_t + nAcum_t + BL$, and
- Model 4 corresponds to the logistic regression of $Y_t \sim C_{t-1} + C_t + Acum_t + t + BL$.

Figure 4.3 reveals that Model 1 (in green) occurs three times, so it was chosen to be the best logistic model for death. Finally, a logistic regression was fitted to

$$Y_t \sim C_t + Acum_t + t + BL. \tag{4.13}$$

4.3.2 Simulation with intervention

For the simulation of new datasets we only consider two treatment interventions. The first one is "always on first-line treatment and never dropping out", $\bar{A}_T = \{0, \ldots, 0\}$ and the second one is "always on second-line treatment and never dropping out" $\bar{A}_T = \{0, 1, \ldots, 1\}$. Note that in both interventions $A_0 = 0$ reflecting the fact that $t = 0$ corresponds to the

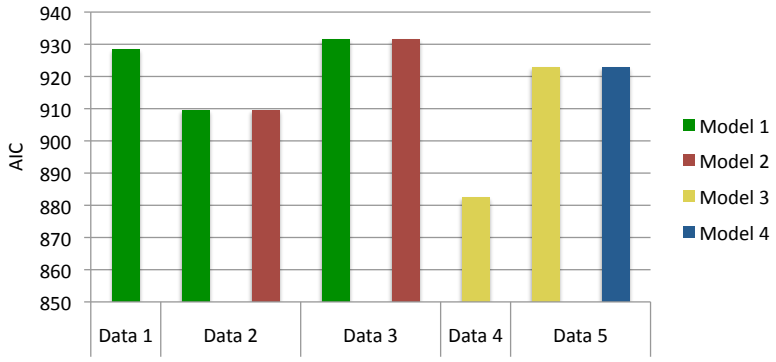

Figure 4.3: AIC for different data and models.

lag one variables of $t = 1$ and that patients entering the study are still on first-line treatment. We will denote the first treatment regime by intervention $\bar{0}$ and the second by intervention $\bar{1}$. Note that the new datasets will then be free of loss to follow-up. Given these interventions, two datasets are simulated.

Denote by T the maximal observed follow-up time. For each patient, C_0, Y_0 and the baseline variables are taken from the original dataset. It turns out that due to the structure of the original dataset, at $t = 1$ no values have to be simulated since it always holds that $C_1 = C_0$ and $Y_1 = 0$. Then for time-points $t = 2, \ldots, T$, C_t and Y_t are simulated by predicting from models (4.12) and (4.13). Note that C_t depends only on variables simulated at time $t-1$ and baseline variables. Having simulated C_t and knowing the value of A_t by intervention, we possess all the information needed to compute Y_t. Each subject is simulated as long as $Y_t \neq 1$, if $Y_t = 1$ then the subject died and there is no need for further simulation. Note that if for a given patient death does not occur his follow-up time is $T = 20$ and hence the new generated datasets are not only free of loss to follow-up but also free of right censoring. The R-code for G-computation can be found in Appendix C.3.

4.3.3 Estimates of the causal effect

After having simulated the two datasets by intervention, an estimate of the causal effect can be obtained by computing the

- log incidence rate = log (number of deaths / total follow-up time), and the
- cumulative incidence = number of deaths / total number of subjects,

as it was done in [7]. To combine information on the two interventions, we compute the risk ratio:

- risk ratio = exp(log incidence rate of intervention $\bar{1}-$ log incidence rate of intervention $\bar{0}$).

4.3.4 Statistical inference

To make statistical inference on the log incidence rate as well as on the cumulative incidence the bootstrap is used. It turned out to be very time-consuming and so only 50 bootstrap samples were generated for each intervention. The computation time for 50 bootstraps and two interventions was approximately 25 hours. The bootstrap consists of sampling with replacement from the study population yielding a bootstrap sample which is then analysed using G-computation. Hence, in each bootstrap the log incidence rate and the cumulative incidence are computed and the overall estimate is chosen to be the mean resulting from the 50 bootstraps. Furthermore, the standard errors and normal based 95% confidence intervals are produced.

As already described in Section 4.2, G-computation and bootstrap are repeated for the five datasets obtained from multiple imputation and the results are then combined according to Rubin [16], yielding standard errors and confidence intervals for the average log incidence rate and the average cumulative incidence.

Chapter 5

Results

In this chapter we present the results obtained by applying G-computation to the dataset introduced in Chapter 4. Note that the corresponding R-code as well as the crude R-output can be found in Appendix C.3 and Appendix C.4.

In a first step, we apply G-computation to the five imputed datasets (Section 4.2) with the two interventions "always on first-line treatment and no loss to follow-up" (intervention $\bar{0}$) and "always on second-line treatment and no loss to follow-up" (intervention $\bar{1}$) to assess if treatment switching has an effect in terms of lifetime. As mentioned in the previous chapter, we give estimates of the log incidence rate (log IR), the cumulative incidence and the risk ratio. Nevertheless, to present the results we always refer to the cumulative incidence as percentage (inc. %) representing the % of patients who died during follow-up. This also holds for the corresponding standard errors (std. error) and confidence intervals (CI). The results of this first analysis are provided in the first column of Table 5.1, Table 5.2 and Table 5.3 (log IR, inc. % and risk ratio). The distribution of death across the baseline variables of the simulated datasets can be found in Appendix C.5.

In a second step, the bootstrap is applied to the datasets yielding average estimates (av. log IR, av. inc. %, av. risk ratio), standard errors and normal based 95% confidence intervals (Tables 5.1, 5.2 and 5.3). Combining the results for the bootstrap average estimates by following the scheme introduced in Section 4.2.2 for intervention $\bar{1}$ and intervention $\bar{0}$ yields overall estimates of the log incidence rate, the % inc. and the risk ratio, as well as standard errors and 95% confidence intervals. These results are provided in Table 5.4. The rate of missing information (γ) as defined in equation (4.8) is also included.

Finally, we apply G-computation to the dataset imputed by last observation carried forward as explained in Section 4.2.3 resulting in the values presented in Table 5.5. Again, the distribution of death across baseline variables under the two interventions for this dataset can be found in Appendix C.5.

intervention $\bar{1}$	log IR	av. log IR	std. error	95% CI
dataset 1	-7.3707	-7.4805	0.2094	$[-7.8908, -7.0702]$
dataset 2	-7.4013	-7.4752	0.1859	$[-7.8396, -7.1108]$
dataset 3	-7.6914	-7.4666	0.1761	$[-7.8117, -7.1216]$
dataset 4	-7.9269	-7.6450	0.2168	$[-8.0699, -7.2200]$
dataset 5	-7.6107	-7.4683	0.1587	$[-7.7794, -7.1572]$
intervention $\bar{0}$	log IR	av. log IR	std. error	95% CI
dataset 1	-5.9549	-6.0400	0.0721	$[-6.1813, -5.8986]$
dataset 2	-6.1225	-6.0653	0.0960	$[-6.2535, -5.8772]$
dataset 3	-5.9867	-6.0378	0.0852	$[-6.2048, -5.8707]$
dataset 4	-6.0609	-6.0910	0.1011	$[-6.2892, -5.8928]$
dataset 5	-6.1138	-6.0250	0.0915	$[-6.2044, -5.8456]$

Table 5.1: Results of G-computation with 50 bootstraps for log incidence rate.

intervention $\bar{1}$	inc. %	av. inc. %	std. error	95% CI
dataset 1	1.2467	1.1175	0.2065	$[0.7127, 1.5223]$
dataset 2	1.2089	1.1190	0.1943	$[0.7382, 1.4998]$
dataset 3	0.9067	1.1258	0.1929	$[0.7476, 1.5040]$
dataset 4	0.7178	0.9513	0.1968	$[0.5656, 1.3369]$
dataset 5	0.9822	1.1220	0.1876	$[0.7543, 1.4898]$
intervention $\bar{0}$	inc. %	av. inc. %	std. error	95% CI
dataset 1	5.0246	4.5531	0.3224	$[3.9212, 5.1850]$
dataset 2	4.2690	4.4458	0.4073	$[3.6474, 5.2442]$
dataset 3	4.8734	4.5561	0.3682	$[3.8345, 5.2777]$
dataset 4	4.5334	4.3461	0.4332	$[3.4969, 5.1952]$
dataset 5	4.3068	4.6158	0.3921	$[3.8472, 5.3844]$

Table 5.2: Results of G-computation with 50 bootstraps for cumulative incidence in %.

	risk ratio	av. risk ratio	std. error	95% CI
dataset 1	0.2427	0.2418	0.0451	$[0.1534, 0.3303]$
dataset 2	0.2784	0.2490	0.0479	$[0.1550, 0.3430]$
dataset 3	0.1818	0.2439	0.0463	$[0.1533, 0.3346]$
dataset 4	0.1547	0.2166	0.0487	$[0.1212, 0.3121]$
dataset 5	0.2238	0.2401	0.0457	$[0.1504, 0.3298]$

Table 5.3: Risk ratio, average risk ratio, standard error and 95% CI obtained by combining the results of G-computation with 50 bootstraps for log incidence rate.

log IR	ov. est.	ov. se	95% CI	γ
intervention $\bar{1}$	-7.5071	0.2085	$[-7.9192, -7.0950]$	0.1759
intervention $\bar{0}$	-6.0518	0.0943	$[-6.2371, -5.8665]$	0.0975
inc. %	ov. est.	ov. se	95% CI	γ
intervention $\bar{1}$	1.0871	0.2127	$[0.6673, 1.5070]$	0.1630
intervention $\bar{0}$	4.5034	0.4039	$[3.7060, 5.3007]$	0.0877
	ov. est.	ov. se	95% CI	γ
risk ratio	0.2383	0.0488	$[0.1425, 0.3340]$	0.0826

γ = rate of missing information, se = std. error

Table 5.4: Overall estimates, standard errors and 95% CI for log incidence rate, cumulative incidence and risk ratio obtained by combining the results of the five imputed datasets (dataset 1 to dataset 5).

	log IR	av. log IR	std. error	95% CI
intervention $\bar{1}$	-7.5005	-7.5562	0.1826	$[-7.9140, -7.1984]$
intervention $\bar{0}$	-5.8593	-5.9927	0.0891	$[-6.1673, -5.8182]$
	inc. %	av. inc. %	std. error	95% CI
intervention $\bar{1}$	1.0956	1.0374	0.1875	$[0.6700, 1.4048]$
intervention $\bar{0}$	5.5157	4.7881	0.3798	$[4.0436, 5.5326]$
	risk ratio	av. risk ratio	std. error	95% CI
	0.1937	0.2137	0.0438	$[0.1279, 0.2995]$

Table 5.5: Sixth Dataset: Results of G-computation with one single simulation and 50 bootstraps for additional dataset imputed by last observation carried forward.

Chapter 6

Discussion

We performed G-computation to assess whether there is a causal effect of switching to second-line treatment among HIV-infected patients experiencing immunological failure. Table 5.4 reveals that more patients die when imposing intervention $\bar{0}$ rather than intervention $\bar{1}$. This holds for the log incidence rate, -6.05 (95% CI $[-6.24, -5.87]$) against -7.51 (95% CI $[-7.92, -7.10]$) , as well as the cumulative incidence (%), 4.50 (95% CI $[3.71, 5.30]$) against 1.09 (95% CI $[0.67, 1.51]$). Recall that the log incidence rate is defined as the log of (number of deaths)/(total follow-up time) and so a small value means that less patients died and/or that patients lived longer resulting in an increase of the total follow-up time. The cumulative incidence (%) equals (number of deaths)/(total number of subjects)$\cdot$100 and decreases if the number of deaths decreases. Additionally the obtained risk ratio 0.24 (95% CI $[0.14, 0.33]$) indicates that the risk of dying is smaller in the population that switched to second-line treatment than in the population that stayed on first-line treatment. Given these results we can conclude that treatment switching has a beneficial effect on lifetime.

Computing the log incidence rate and the cumulative incidence (%) of the original dataset yields -5.4251 and 2.8712, respectively. The cumulative incidence value lies between the results obtained by G-computation, reflecting the fact that the observed treatment regime is situated somewhere between intervention $\bar{0}$ and $\bar{1}$. A substantial difference between the original and the simulated datasets is that in the former there is loss to follow-up which is ignored by G-computation. Furthermore, opposite to the right censoring in the original dataset, patients in the new dataset are simulated until time $T = 20$ except for death occurring. These two facts explain why the log incidence rate in the observed dataset is bigger than the results obtained by G-computation.

Comparing the G-computation estimates to those obtained by applying another estimation procedure to the same dataset should give valuable information on the correctness of our procedure. A very similar dataset was analysed in Peter's Master's Thesis [8] and in *The Causal Effect of Switching to Second-line ART in Programmes without Access to Routine Viral Load Monitoring* by Gsponer et al. [9] using IPTW (Inverse Probability of Treatment Weighted). IPTW consists of weighing each observation with a specific weight so that the time-varying confounding is neutralised and a standard survival analysis can be

performed. Gsponer et al. mention that "mortality was lower among patients who switched compared to patients remaining on failing first-line antiretroviral treatment: hazard ratio 0.25 (95% CI 0.09-0.72)" [9]. The hazard ratio of switching to second-line treatment in Peter's work differs slightly and equals 0.32. The risk ratio in our analysis is 0.24 (95% CI $[0.14, 0.33]$) for the five datasets imputed by MICE and 0.21 (95% CI $[0.13, 0.30]$) for the dataset imputed by last observation carried forward which corresponds also to the imputation method used in [8] and [9]. The results are very close to those found by Gsponer et al. and Peters. Due to the fact that the null-paradox as stated in Section 3.5 does not hold for IPTW and given the fact that the results for both procedures are very similar we can conclude that the sharp null hypothesis was not falsely rejected. Recall that the paradox reveals that the estimated causal effect depends on treatment although in reality this is not true.

In Chapter 3 we argued that under the assumption of no unmeasured confounder and the absence of model misspecification, G-computation yields a consistent estimate of the causal effect of switching to second-line treatment. An indication for model misspecification could be gained by modelling treatment in the observed dataset. Finding a model for treatment (as done for log CD4 and death) would permit to perform G-computation with the observed treatment intervention, i.e. at each time-point simulating all the variables (C_t, A_t, Y_t), except for the baseline variables and C_0. The resulting log incidence rate and cumulative incidence should then be close to the observed values. If this is not the case, it is an indication for model misspecification. Note that the reverse does not hold, i.e. similar results do not exclude model misspecification. In our example, however, this comparison is problematic since the observed intervention does not account for loss to follow-up nor for right censoring resulting in a simulated dataset that differs substantially from the observed one.

Assuming an appropriate causal graph is crucial for the analysis particularly in the light of the no unmeasured confounder assumption. This assumption is not testable and in observational studies it will never be exactly true. The longitudinal study we considered in this work includes 13.8 % patients who are switched to second-line treatment, which is a rather small percentage. As already mentioned in the first Chapter, second-line treatment is more cost-intense which could be a reason for the low switching rate. Furthermore, there are clinics able to measure viral loading yielding additional information on treatment failure (see also Chapter 1). If viral loading measurements influence the decision on switching, this variable should be included in the analysis. Table 6.1 presents the distribution of treatment switching across the baseline variables. Patients with small CD4 baseline counts are more likely to switch treatment: 30.53% (CD4 baseline < 50) against 5.14% (CD4 baseline ≥ 200).

	% switched
total	13.79
male	13.87
female	13.72
less adv. stage	14.07
adv. stage	13.69
age ≤ 30	13.76
$30 <$ age ≤ 39	14.52
age ≥ 40	10.69
CD4 ≤ 50	30.53
$50 <$ CD4 ≤ 99	17.54
$100 \leq$ CD4 ≤ 199	12.26
$200 \leq$ CD4	5.14

Table 6.1: Distribution of treatment switching across baseline variables in observed dataset.

Possible improvements

Computation time. The code for G-computation implemented in R turned out to be very time-intense. This is due to the fact that for each subject each time-point is simulated separately. Note that in our example a simulated data matrix includes more than 51000 rows (over time and subjects). Due to time constraints, we were unable to perform more than 50 bootstraps although originally we considered 1000 bootstraps.

Loss to follow-up, a competing risk? Right censoring occurs due to the closure date of the database whereas loss to follow-up occurs when a patient does not return to the clinic for twelve consecutive months. It is not clear to what extend loss to follow-up is caused by death, making it a competing risk for the outcome. Efforts are made to improve information on loss to follow-up in the database, enabling to reveal whether a patient is lost to follow-up due to death or if it is due to some other reason.

Missing CD4 values. The presence of 64% missing CD4 values (over time and subjects) is of great concern when analysing the dataset. A question arising in this context is why so many CD4 cell counts are missing and efforts should be made to encounter this problem. We imputed the missing values using Multivariate Imputation by Chained Equations (MICE) in R in combination with predictive mean matching. Another dataset was obtained using imputation by last observation carried forward and both imputation methods yielded similar results. We want to notice that we could also have used linear regression in combination with MICE, baring the disadvantage that it may result in negative CD4 values, although, when using log CD4 this should not be problematic. The reason why we did not use linear regression is that by then, we had already based our

search regarding models for log CD4 and death on the datasets imputed by predictive mean matching.

Include model fitting in the bootstrap. To perform G-computation we estimated models for log CD4 and death. The models were fitted to the original dataset yielding estimates for their parameters, say $\hat{\theta}$. In the bootstrap we then sampled with replacement from the observed study population to obtain a new population to which G-computation was applied using the model parameters $\hat{\theta}$ from the original dataset. Instead, the bootstrap should also fit the models for log CD4 and death based on the bootstrap sample and use them for simulation under intervention.

Simulation with other interventions. G-computation gives the possibility to simulate new datasets under various interventions. In this work we only considered two of them yielding the necessary information to estimate the causal effect. One could turn to interventions with different switching points such as "stay on first-line treatment for 3 quarters then switch and no loss to follow-up". G-computation also allows for dynamic interventions, not considered in this work but possibly interesting for future studies. Such an intervention could be, for instance, "switch to second-line treatment if CD4 cell count is less than a certain value".

Concluding words

Our aim was to estimate the causal effect of switching to second-line treatment among HIV-patients in Southern Africa. This was successfully done, especially in the light of confirming results obtained by IPTW ([8],[9]). The risk ratio 0.24 (95% CI 0.14-0.33) can be retained as the main indicator for a beneficial effect of treatment switching.

Appendix A

d-separation and conditional independence

We provide some definitions and theorems which may complete the view on the one-to-one correspondence between the set of conditional independencies $A \perp\!\!\!\perp B \mid S$ implied by the recursive decomposition of equation (2.4) in a DAG and the set of triples (A, S, B) that satisfy the d-separation criterion (definition 2.2). For further reading see first chapter of Pearl [11].

Definition A.1. *If a probability function P admits the factorisation of 2.4 relative to a DAG G, we say that G represents P, that G and P are compatible, or that P is Markov relative to G.*

Theorem A.1. *For any three disjoint subsets of nodes (X, Y, Z) in a DAG G and for all probability functions P, we have:*

1. *Z d-separates X and $Y \Longrightarrow X \perp\!\!\!\perp Y \mid Z$ whenever G and P are compatible; and*
2. *if $X \perp\!\!\!\perp Y \mid Z$ holds in all distributions compatible with G, if follows that Z d-separtes X.*

Theorem A.2. *A necessary and sufficient condition for a probability distribution P to be Markov relative to a DAG G is that, conditional on its parents in G, each variable be independent of all its predecessors in some ordering of the variables that agrees with the arrows of G.*

Theorem A.3. *A necessary and sufficient condition for a probability distribution P to be Markov relative to a DAG G is that every variable be independent of all its non-descendants (in G), conditional on its parents.*

Appendix B

B.1 R-code corresponding to Section 3.4

B.1.1 Fixed time-point example (Section 3.4.1)

```
##############################################
# Chapter 3: Fixed time-point example
##############################################

#------------------------------
# fit logistic regression for Y
#------------------------------

fitY <- glm(Y~A*C,family="binomial",data=data)
summary(fitY)

#------------------------------
# compute causal effect explicitly
#------------------------------
coef1 <- fitY$coefficients

# p1 = probability of Y = 1 | do(A=1)
p1 <- exp(t(coef1)%*%c(1,1,1,1))/(1+exp(t(coef1)%*%c(1,1,1,1)))*26/46
+ exp(t(coef1)%*%c(1,1,0,0))/(1+exp(t(coef1)%*%c(1,1,0,0)))*20/46

# p0 = probability of Y = 1 | do(A=0)
p0 <- exp(t(coef1)%*%c(1,0,1,0))/(1+exp(t(coef1)%*%c(1,0,1,0)))*26/46
+ exp(t(coef1)%*%c(1,0,0,0))/(1+exp(t(coef1)%*%c(1,0,0,0)))*20/46

# risk difference
p0-p1

#------------------------------
# G-COMPUTATION
#------------------------------
```

```
# function sim simulates new dataset under intervention
# dat = dataset
# fit.Y = fit for the outcome
# int = intervention on treatment

sim <- function(dat,fit.Y,int){
Y.new <- rep(NA,46)
A <- int
for (i in 1:46){
new <- data.frame(A=A[i],C=dat[i,]$C)
pY <- predict(fit.Y,newdata=new,type="response")
Y.new[i] <- rbinom(1,1,pY)
}
return(data.frame(dat$C,A,Y.new))
}

# perform bootstrap
# B = number of bootstrap samples
# fit.Y = fit for the outcome
# int1 = intervention 1
# int0 = intervention 2

BS <- function(B,sim,dat,fit.Y,int1,int0){
est <- matrix(NA,B,2)
n <- nrow(dat)
for (j in 1:B){
tmp <- sample(n,replace=TRUE)
datB <- dat[tmp,]
datB1 <- sim(datB,fit.Y,int1)
datB0 <- sim(datB,fit.Y,int0)
est[j,1] <- sum(datB0$Y)/46
est[j,2] <- sum(datB1$Y)/46
}
risk <- est[,1] - est[,2]
est.risk <- mean(risk)
std.err <- sd(risk)
z <- est.risk/std.err
pval <- pnorm(z,0,1)
confint <- cbind(est.risk-1.96*std.err,est.risk+1.96*std.err)
result <- data.frame(est.risk,std.err,z,pval,confint)
names(result) <- c("G-computation estimate of risk difference P0-P1",
"bootstrap std. error","z","P>|z|","normal-based","95% conf.int.")
return(result)
}
```

```
#------------------------------
# Apply G-computation
#------------------------------
int0 <- rep(0,46)
int1 <- rep(1,46)

# perform G-computation: get estimate of risk difference
# and confidence interval
# 10000 bootstrap samples
set.seed(81)
riskdiff <- BS(10000,sim,data,fitY,int1,int0)
```

B.1.2 Two time-point example (Section 3.4.2)

```
############################################################
# Chapter 3: Two time-point example
############################################################

# n = number of individuals
# 5 variables: C0, C1, A0, A1, Y
# C0 = baseline CD4 cell count
# C1 = CD4 cell count at time 1
# A0 = switching at time 0
# A1 = switching at time 1 (=1 if switched, 0 if not)
# Y = 1 if person has died between time 0 and time 1

sim <- function(n){
C0 <- C1 <- A0 <- A1 <- Y <- rep(NA,n)

for (j in 1:n){
U <- rnorm(1,0,0.5)
C0[j] <- rnorm(1,5.5+U,0.2)
A0[j] <- rbinom(1,1,exp(5-C0[j])/(1+exp(5-C0[j])))
C1[j] <- rnorm(1,0.9*C0[j]+A0[j]+0.1*U,sqrt(0.01))

if (A0[j]==1) A1[j] <- 1
else A1[j] <- rbinom(1,1,exp(-7 + 0.1*A0[j]+C1[j]+0.5*C0[j])
/(1+exp(-7 + 0.1*A0[j]+C1[j]+0.5*C0[j])))

Y[j] <- rbinom(1,1,exp(-16 + 1.2*C0[j] + 1.8*C1[j]- 1*A0[j]
- 1*A1[j] - U)/(1+exp(-16 + 1.2*C0[j] + 1.8*C1[j]-
1*A0[j] - 1*A1[j] - U)))
}
data <- cbind(C0,C1,A0,A1,Y)
data <- data.frame(data)
names(data) <- c("C0","C1","A0","A1","Y")
return(data)
}

#------------------------------
# create original dataset and
# fit models for C1 and Y
#------------------------------

# generate a dataset with n subjects:
n <- 10000
set.seed(21)
```

```
data <- sim(n)

# estimate the model parameters:
# have to fit a model for C1 (by linear regression)
# and one for Y (by logistic regression)

fitC1 <- lm(C1~ C0 + A0, data=data)

fitY <- glm(Y~ C0 + A0 + C1 + A1, family="binomial", data=data)

#------------------------------
# functions needed for G-computation
#------------------------------

# create generates new dataset by intervention on treatment

create <- function(data,f.C1,f.Y,int){
sig <- summary(f.C1)$sigma
n <- nrow(data)
A0_new <- int
new_dataC1 <- data.frame(C0=data$C0,A0=A0_new)
C1_est <- predict(f.C1,new_dataC1)
C1_new <- C1_est + rnorm(n,0,sig)
A1_new <- int
new_dataY <- data.frame(C0=data$C0,A0=A0_new,C1=C1_new,A1=A1_new)
pY_new <- predict(f.Y,new_dataY,type="response")
Y_new <- rbinom(n,1,pY_new)

data_new <- data.frame(data$C0,C1_new,A0_new,A1_new,Y_new)
return(data_new)
}

# BSest computes bootstrap estimates for risk difference
# risk ratio and odds ratio

BSest <- function(B,dat,fct.gen,int,fit.Y,fit.C){
risk.diff <- risk.ratio <- odds.ratio <- rep(NA,B)

for (j in 1:B){

tmp <- sample(n,replace=TRUE)
datB <- dat[tmp,]

dat0 <- fct.gen(datB,fit.C,fit.Y,int[,1])
dat1 <- fct.gen(datB,fit.C,fit.Y,int[,2])
```

```
p1 <- mean(dat1$Y)
p0 <- mean(dat0$Y)
risk.diff[j] <- p0-p1
risk.ratio[j] <- p0/p1
odds.ratio[j] <- p0/(1-p0)*(1-p1)/p1
print(j)
}
return(list(risk.diff,risk.ratio,odds.ratio))
}

# BSci computes confidence intervals etc.

BSci <- function(BS.est){
p <- length(BS.est)
est <- std.err <- z <- pval <- rep(NA,p)
confint <- matrix(NA,p,2)
for (i in 1:p){
est[i] <- mean(BS.est[[i]])
std.err[i] <- sd(BS.est[[i]])
z[i] <- est[i]/std.err[i]
pval[i] <- pnorm(z[i],0,1)
confint[i,] <- c(est[i]-1.96*std.err[i],est[i]+1.96*std.err[i])
}
intname <- rbind("risk diff (p0-p1)","risk ratio (p0/p1)",
"odds ratio (p0*(1-p1)/((1-p0)*p1)")
result <- data.frame(intname,est,std.err,z,pval,confint)
names(result) <- c("measure","av. G-computation estimate of measure",
"Bootstrap Std. Err.","z","P>|z|","normal-based","95% conf. int.")
return(result)
}

#------------------------------
# Apply G-computation
#------------------------------

int0 <- rep(0,n)
int1 <- rep(1,n)
int <- cbind(int0,int1)

set.seed(81)
est <- BSest(10000,data,create,int,fitY,fitC1)
est.ci <- BSci(est)
```

B.2 R-output of G-computation corresponding to Section 3.4

B.2.1 Fixed time-point example (Section 3.4.1)

R-output corresponding to R-code in Appendix B.1.1.

```
> set.seed(81)
> riskdiff <- BS(10000,sim,data,fitY,int1,int0)
> riskdiff
  G-computation estimate of risk difference P0-P1
1                                        0.1970957
  bootstrap std. error       z     P>|z| normal-based
1           0.09981954 1.97452 0.9758387  0.001449348
  95% conf.int.
1      0.392742
```

B.2.2 Two time-point example (Section 3.4.2)

R-output corresponding to R-code in Appendix B.1.2.

```
> set.seed(81)
> est <- BSest(10000,data,create,int,fitY,fitC1)
> est.ci <- BSci(est)
> est.ci
                            measure av. G-computation estimate of measure
1                 risk diff (p0-p1)                              0.03936222
2                risk ratio (p0/p1)                              1.10849343
3 odds ratio (p0*(1-p1)/((1-p0)*p1)                              1.18178695
  Bootstrap Std. Err.         z P>|z| normal-based 95% conf. int.
1         0.006039668  6.517282     1   0.02752447     0.05119997
2         0.017549300 63.164539     1   1.07409680     1.14289006
3         0.030266361 39.046218     1   1.12246488     1.24110901
```

Appendix C

This appendix relates to Chapter 4 and 5 providing notational remarks, results from model seeking, R-code and R-output from G-computation.

C.1 Notational equivalence

Equivalence of variable notations between the R-code, Table C.1 and Chapter 4:

- t = qrt
- C_t = lcd4
- C_{t-1} = lcd4lag1
- A_t = sl_yn
- Y_t = death
- $Acum_t$ = acum
- $Acum_{t-1}$ = acumlag1
- $nAcum_t$ = anticum
- $nAcum_{t-1}$ = anticumlag1

C.2 Model for log CD4

Table C.1 provides the full results of the leave-one-out cross validation for models 1 to 30, also shown in Figure 4.2. The average mean squared error (av. MSE) corresponds to the average over the five imputed datasets. For more details we refer to Section 4.3.1.

model	predictors for lcd4	av. MSE
1	lcd4lag1 + acumlag1 + BL	0.3657061
2	(lcd4lag1 + acumlag1)*BL	0.36475214
3	lcd4lag1 + anticumlag1 + BL	0.35065262
4	(lcd4lag1 + anticumlag1)*BL	0.34797538
5	qrt + acumlag1 + BL	0.3574325
6	(qrt + acumlag1)*BL	0.3436616
7	qrt + anticumlag1 + BL	0.35482512
8	(qrt + anticumlag1)*BL	0.34102508
9	lcd4lag1 + anticumlag1 + acumlag1 + BL	0.33768246
10	(lcd4lag1 + anticumlag1 + acumlag1)*BL	0.33191436
11	qrt + anticumlag1 + acumlag1 + BL	0.34967444
12	(qrt + anticumlag1 + acumlag1)*BL	0.33027226
13	ns(qrt,5) + acumlag1 + BL	0.3486849
14	(ns(qrt,5) + acumlag1)*BL	0.32918106
15	ns(qrt,5) + anticumlag1 + BL	0.34867466
16	(ns(qrt,5) + anticumlag1)*BL	0.32917686
17	ns(qrt,5) + anticumlag1 + acumlag1 + BL	0.3487092
18	(ns(qrt,5) + anticumlag1 + acumlag1)*BL	0.32941998
19	lcd4lag1 + qrt + acumlag1 + BL	0.33472186
20	(lcd4lag1 + qrt + acumlag1)*BL	0.3279395
21	lcd4lag1 + qrt + anticumlag1 + BL	0.33292074
22	(lcd4lag1 + qrt + anticumlag1)*BL	0.32603942
23	lcd4lag1 + qrt + anticumlag1 + acumlag1 + BL	0.32848062
24	(lcd4lag1 + qrt + anticumlag1 + acumlag1)*BL	0.31730434
25	lcd4lag1 + ns(qrt,5) + acumlag1 + BL	0.32830256
26	(lcd4lag1 + ns(qrt,5) + acumlag1)*BL	0.31695914
27	lcd4lag1 + ns(qrt,5) + anticumlag1 + BL	0.32829548
28	(lcd4lag1 + ns(qrt,5) + anticumlag1)*BL	0.31694312
29	lcd4lag1 + ns(qrt,5) + anticumlag1 + acumlag1 + BL	0.32830944
30	(lcd4lag1 + ns(qrt,5) + anticumlag1 + acumlag1)*BL	0.31719898

Table C.1: Average mean squared error obtained from leave-one-out cross validation.

C.3 R-Code corresponding to G-computation applied to data described in Chapter 4

In the following R-Code the functions used for G-computation are defined.

```
#----------------------------------------------------------------
# compute the LOG INCIDENCE RATE
#----------------------------------------------------------------

# IR computes the log incidence rate of the dataset = data
# logIR corresponds to (number of events occured)/(person time)

IR <- function(data){
log(sum(data$death,na.rm=TRUE)/sum(data$qrt[!duplicated(data$id,fromLast=TRUE)]))
}

#----------------------------------------------------------------
# compute the CUMULATIVE INCIDENCE
#----------------------------------------------------------------

# cum computes the percentage of patients dying during the study

cum <- function(data){
sum(data$death,na.rm=TRUE)/max(data$id)
}

#----------------------------------------------------------------
# simulate datasets with TREATMENT INTERVENTION
#----------------------------------------------------------------

# create computes new datasets for given treatment intervention
# dat = original dataset
# int = intervention
# fit.Y = fitted model for Y (=death)
# fit.C = fitted model for C (=cd4 or similar)

create <- function(dat,int,fit.Y,fit.C){
n <- max(dat$id)
k <- max(dat$qrt)
# at qrt=1 lcd4=lcd4lag1
C0 <- dat[dat$qrt==1,]$lcd4
# vector of baseline variables
bas_age_gp <- dat$bas_age_gp[!duplicated(dat$id)]
adv_stage <- dat$adv_stage[!duplicated(dat$id)]
gender <- dat$gender[!duplicated(dat$id)]
```

```
bas_cd4_gp <- dat$bas_cd4_gp[!duplicated(dat$id)]
dat <- matrix(NA, n*(k+1), 14)
sig <- summary(fit.C)$sigma
counter <- 0

for (i in 1:n){
counter <- counter + 1
# for qrt=1
C <- C0[i]
A <- int[1]
Anticum <- 1-A
Acum <- A
# vec has components id,qrt,death,lcd4,lcd4lag1,
# sl_yn,acum,anticum,acumlag1,anticumlag1, gender,
# bas_age_gp, adv_stage, bas_cd4_gp
vec <- c(i,1,0,C,C,A,Acum,Anticum,0,0,gender[i],bas_age_gp[i],
adv_stage[i],bas_cd4_gp[i])
dat[counter,] <- vec

j = 2
Y = 0

# for qrt=2 to k
while (Y==0 && j <= k){
counter <- counter + 1
Clag <- C
Alag <- A
Anticumlag <- Anticum
A <- int[j]
Acumlag <- Acum
Acum <- Acum + A
Anticum <- 1 - A + Anticum

new.dataC <- data.frame(lcd4lag1=Clag,anticumlag1=Anticumlag,qrt=j,
bas_age_gp=bas_age_gp[i],adv_stage=adv_stage[i],bas_cd4_gp=bas_cd4_gp[i],
gender=gender[i])
Cpred <- predict(fit.C,new.dataC)
C <- Cpred + rnorm(1,0,sig)

new.dataY <- data.frame(lcd4=C,acum=Acum,qrt=j,bas_age_gp=bas_age_gp[i],
adv_stage=adv_stage[i],bas_cd4_gp=bas_cd4_gp[i],gender=gender[i])
pY <- predict(fit.Y,new.dataY,type="response")
Y <- rbinom(1,1,pY)
v <- c(i,j,Y,C,Clag,A,Acum,Anticum,Acumlag,Anticumlag,gender[i],
bas_age_gp[i],adv_stage[i],bas_cd4_gp[i])
```

```
dat[counter,] <- v
j=j+1
}
}
dat <- dat[1:counter,]
# data has 14 columns: id,qrt,death,lcd4,lcd4lag1,
# sl_yn,acum,anticum,acumlag1,anticumlag1, gender
# bas_age_gp, adv_stage, bas_cd4_gp
data <- data.frame(dat,row.names=c(1:nrow(dat)))
names(data) <- c("id","qrt","death","lcd4","lcd4lag1","sl_yn","acum",
"anticum","acumlag1","anticumlag1","gender","bas_age_gp","adv_stage",
"bas_cd4_gp")

return(data)
}

#-----------------------------------------------------------------
# BOOTSTRAP
#-----------------------------------------------------------------

# BSest generates estimates for the bootstrap samples
# B = number of bootstrap samples
# dat = original dataset
# fct.gen = function used to create datasets with intervention
# n = number of subjects in the dataset
# k = max. value of time
# int = intervention matrix with one column per intervention
# fit.Y = fitted model for Y (=death)
# fit.C = fitted model for C (=cd4)
# the return of BSest is a list with matrices, first column
# corresponding to logIR and second column to cumulative incidence

BSest <- function(B,dat,fct.gen,int,fit.Y,fit.C){
n <- max(dat$id)
k <- max(dat$qrt)
res1 <- res0 <- matrix(NA,B,2)

for (j in 1:B){
id <- dat$id
tmp <- sample(n,replace=TRUE)
ind <- NULL
idNew <- NULL
for (i in 1:n){
                  chosenRows <- which(id==tmp[i])
                  ind <- c(ind,chosenRows)
```

```
                  idNew <- c(idNew, rep(i, length(chosenRows)))
}

# datB is the bootstrap sample
datB <- dat[ind,]
# idNew is valid id for bootstrap sample
datB$id <- idNew

# compute IR and cum of bootstrap sample datB
# for intervention "always 1"
create1 <- fct.gen(datB,int[,1],fit.Y,fit.C)
res1[j,1] <- IR(create1)
res1[j,2] <- cum(create1)
# for intervention "always 0"
create0 <- fct.gen(datB,int[,2],fit.Y,fit.C)
res0[j,1] <- IR(create0)
res0[j,2] <- cum(create0)
print(j)
}
res <- list(res1,res0)
return(res)
}

#-----------------------------------------------------------------
# compute the bootstrap confidence intervals etc.
#-----------------------------------------------------------------
# output computes confidence intervals and standard error of
# bootstrap samples logIR and cum
# BS.est is an object of class BSest, i.e. list of matrices,
# list has length corresponding to number of interventions
# each matrix corresponding to an intervention on each
# bootstrap sample (B = number of rows) and containing IR (1st
# column) and cum (2nd column) estimates (2 = number of columns)
# p is the number of interventions used in BS.est
# p = length(BS.est)

output <- function(BS.est){
p <- length(BS.est)
avIR <- stderrIR <- zIR <- pvalIR <- rep(NA,p)
avCum <- stderrCum <- zCum <- pvalCum <- rep(NA,p)
confintIR <- confintCum <- matrix(NA,p,2)
for (i in 1:p){
mat <- BS.est[[i]]
# logIR is in first column
avIR[i] <- mean(mat[,1])
```

```
stderrIR[i] <- sd(mat[,1])
zIR[i] <- avIR[i]/stderrIR[i]
pvalIR[i] <- pnorm(zIR[i],0,1)
confintIR[i,] <- c(avIR[i]-1.96*stderrIR[i],avIR[i]+1.96*stderrIR[i])
# cumulative incidence is in second column
avCum[i] <- mean(mat[,2])
stderrCum[i] <- sd(mat[,2])
zCum[i] <- avCum[i]/stderrCum[i]
pvalCum[i] <- pnorm(zCum[i],0,1)
confintCum[i,] <- c(avCum[i]-1.96*stderrCum[i],avCum[i]+1.96*stderrCum[i])
}
intname <- rbind("always 1","always 0")
resultIR <- data.frame(intname,avIR,stderrIR,zIR,pvalIR,confintIR)
names(resultIR) <- c("Intervention","G-computation estimate of av. log IR",
"Bootstrap Std. Err.","z","P>|z|","normal-based","95% conf. int.")
resultCum <- data.frame(intname,avCum,stderrCum,zCum,pvalCum,confintCum)
names(resultCum) <- c("Intervention","G-computation estimate of cum.
incidence",
"Bootstrap Std. Err.","z","P>|z|","normal-based","95% conf. int.")
return(list(resultIR,resultCum))
}

#---------------------------------------------------------------
# Combine results for Multiply Imputed Data, i.e. five datasets.
#---------------------------------------------------------------

# combMICE combines results for multiple imputed data according to
# Rubin
# Input is an object of class output

combMICE <- function(out.1,out.2,out.3,out.4,out.5){
# m=number of imputed datasets
m <- 5
# nint=number of interventions
nint <- nrow(out.1[[1]])
# store the estimates for IR int1, int0
# and cum int1, int0
IR <- cum <- matrix(NA,nint,m)

IR[,1] <- out.1[[1]][,2]
IR[,2] <- out.2[[1]][,2]
IR[,3] <- out.3[[1]][,2]
IR[,4] <- out.4[[1]][,2]
IR[,5] <- out.5[[1]][,2]
```

```
cum[,1] <- out.1[[2]][,2]
cum[,2] <- out.2[[2]][,2]
cum[,3] <- out.3[[2]][,2]
cum[,4] <- out.4[[2]][,2]
cum[,5] <- out.5[[2]][,2]

# store the standard errors for each estimate
std.errIR <- std.errCum <- matrix(NA,nint,m)
std.errIR[,1] <- out.1[[1]][,3]
std.errIR[,2] <- out.2[[1]][,3]
std.errIR[,3] <- out.3[[1]][,3]
std.errIR[,4] <- out.4[[1]][,3]
std.errIR[,5] <- out.5[[1]][,3]

std.errCum[,1] <- out.1[[2]][,3]
std.errCum[,2] <- out.2[[2]][,3]
std.errCum[,3] <- out.3[[2]][,3]
std.errCum[,4] <- out.4[[2]][,3]
std.errCum[,5] <- out.5[[2]][,3]

est.IR <- est.cum <- W.IR <- W.cum <- B.IR <- B.cum <- rep(NA,nint)
var.IR <- var.cum <- df.IR <- df.cum <- rate.IR <- rate.cum <- rep(NA,nint)
CI.IR <- CI.cum <- matrix(NA,nint,2)
for (i in 1:nint){
# compute overall estimate, i.e. mean for IR and cum over m datasets
est.IR[i] <- mean(IR[i,])
est.cum[i] <- mean(cum[i,])
# within-imputation variance
W.IR[i] <- mean(std.errIR[i,]^2)
W.cum[i] <- mean(std.errCum[i,]^2)
# between imputation variance
B.IR[i] <- 1/(m-1)*sum((IR[i,]-est.IR[i])^2)
B.cum[i] <- 1/(m-1)*sum((cum[i,]-est.cum[i])^2)
# total variance
var.IR[i] <- W.IR[i] + (1+1/m)*B.IR[i]
var.cum[i] <- W.cum[i] + (1+1/m)*B.cum[i]
# degrees of freedom for student's t distribution
df.IR[i] <- (m-1)*(1 + m*W.IR[i]/((m+1)*B.IR[i]))^2
df.cum[i] <- (m-1)*(1 + m*W.cum[i]/((m+1)*B.cum[i]))^2
# confidence intervals
qIR <- qt(0.975,df.IR[i])
qcum <- qt(0.975,df.cum[i])
CI.IR[i,] <- c(est.IR[i]-qIR*sqrt(var.IR[i]),est.IR[i]+qIR*sqrt(var.IR[i]))
CI.cum[i,] <- c(est.cum[i]-qcum*sqrt(var.cum[i]),
est.cum[i]+qcum*sqrt(var.cum[i]))
```

```
# relative increase in variance due to non-response
rIR <- (1+1/m)*B.IR[i]/W.IR[i]
rcum <- (1+1/m)*B.cum[i]/W.cum[i]
# estimated rate of missing information
rate.IR[i] <- (rIR+2/(df.IR[i]+3))/(rIR+1)
rate.cum[i] <- (rcum+2/(df.cum[i]+3))/(rcum+1)
}

result.IR <- data.frame(est.IR,W.IR,B.IR,sqrt(var.IR),CI.IR,rate.IR)
result.cum <- data.frame(est.cum,W.cum,B.cum,sqrt(var.cum),CI.cum,rate.cum)

names(result.IR) <- c("Overall estimate of log incidence rate",
"Within-imputation variance","Between imputation variance","Overall std. error",
"confidence interval","based on student t","rate of missing information")
names(result.cum) <- c("Overall estimate of cumulative incidence",
"Within-imputation variance","Between imputation variance","Overall std. error",
"confidence interval","based on student t","rate of missing information")

return(list(result.IR,result.cum))
}

#-----------------------------------------------------------------
# Code for G-computation with intervention=int (always 1, always 0)
# on 6 datasets, 5 generated by MICE-imputation, the 6th by
# last observation carried forward
#-----------------------------------------------------------------
# Gcomp applies G-computation on 5 imputed datasets and on
# dataset 6 imputed by last observation carried forward

Gcomp <- function(datl1,datl2,datl3,datl4,datl5,datl6,create,BSest,
output,B,int,cum,IR,combMICE, seed = 199){
# for each dataset fit the models for lcd4
fitL1 <- lm(lcd4~(lcd4lag1+ns(qrt,5)+anticumlag1)*(bas_age_gp+adv_stage
+bas_cd4_gp+gender),data=datl1)
fitL2 <- lm(lcd4~(lcd4lag1+ns(qrt,5)+anticumlag1)*(bas_age_gp+adv_stage
+bas_cd4_gp+gender),data=datl2)
fitL3 <- lm(lcd4~(lcd4lag1+ns(qrt,5)+anticumlag1)*(bas_age_gp+adv_stage
+bas_cd4_gp+gender),data=datl3)
fitL4 <- lm(lcd4~(lcd4lag1+ns(qrt,5)+anticumlag1)*(bas_age_gp+adv_stage
+bas_cd4_gp+gender),data=datl4)
fitL5 <- lm(lcd4~(lcd4lag1+ns(qrt,5)+anticumlag1)*(bas_age_gp+adv_stage
+bas_cd4_gp+gender),data=datl5)
fitL6 <- lm(lcd4~(lcd4lag1+ns(qrt,5)+anticumlag1)*(bas_age_gp+adv_stage
+bas_cd4_gp+gender),data=datl6)
# for each dataset fit the models for death
```

```
fitY1 <- glm(death~lcd4+acum+qrt+bas_age_gp+adv_stage+gender
+bas_cd4_gp,family="binomial",data=datl1)
fitY2 <- glm(death~lcd4+acum+qrt+bas_age_gp+adv_stage+gender
+bas_cd4_gp,family="binomial",data=datl2)
fitY3 <- glm(death~lcd4+acum+qrt+bas_age_gp+adv_stage+gender
+bas_cd4_gp,family="binomial",data=datl3)
fitY4 <- glm(death~lcd4+acum+qrt+bas_age_gp+adv_stage+gender
+bas_cd4_gp,family="binomial",data=datl4)
fitY5 <- glm(death~lcd4+acum+qrt+bas_age_gp+adv_stage+gender
+bas_cd4_gp,family="binomial",data=datl5)
fitY6 <- glm(death~lcd4+acum+qrt+bas_age_gp+adv_stage+gender
+bas_cd4_gp,family="binomial",data=datl6)
# for each dataset create new datasets for each intervention
# and B bootstrap samples
set.seed(seed)
BS.est1 <- BSest(B,datl1,create,int,fitY1,fitL1)
print("datl1---BSest done")
set.seed(seed)
BS.est2 <- BSest(B,datl2,create,int,fitY2,fitL2)
print("datl2---BSest done")
set.seed(seed)
BS.est3 <- BSest(B,datl3,create,int,fitY3,fitL3)
print("datl3---BSest done")
set.seed(seed)
BS.est4 <- BSest(B,datl4,create,int,fitY4,fitL4)
print("datl4---BSest done")
set.seed(seed)
BS.est5 <- BSest(B,datl5,create,int,fitY5,fitL5)
print("datl5---BSest done")
set.seed(seed)
BS.est6 <- BSest(B,datl6,create,int,fitY6,fitL6)
print("datl6---BSest done")
# for each dataset generate output as in stata
out1 <- output(BS.est1)
out2 <- output(BS.est2)
out3 <- output(BS.est3)
out4 <- output(BS.est4)
out5 <- output(BS.est5)
out6 <- output(BS.est6)
# for datasets 1 to 5 combine the results to one single output
out.MICE <- combMICE(out1,out2,out3,out4,out5)

return(list(out1,out2,out3,out4,out5,out6,out.MICE,BS.est1,BS.est2,BS.est3,
BS.est4,BS.est5,BS.est6))
}
```

C.4 R-Output corresponding to G-computation applied to data described in Chapter 4

```
> # create bootstrap estimates of log incidence rate and cumulative incidence
> int <- cbind(rep(1,20),rep(0,20))
>
> # Parameter
> BB <- 50
> # run G-computation
> testG <- Gcomp(datl1,datl2,datl3,datl4,datl5,datl6,create,BSest,output,BB,
int,cum,IR,combMICE)
>
> # testG[[1]] = output G-computation for dataset 1 (datl1)
> # testG[[2]] = output G-computation for dataset 2 (datl2)
> # testG[[3]] = output G-computation for dataset 3 (datl3)
> # testG[[4]] = output G-computation for dataset 4 (datl4)
> # testG[[5]] = output G-computation for dataset 5 (datl5)
> # testG[[6]] = output G-computation for dataset 6 (datl6)
> # testG[[7]] = combined MICE output for datl1-datl5
>
>
> testG[[1]]
[[1]]
  Intervention G-computation estimate of av. log IR Bootstrap Std. Err.
1     always 1                            -7.480504          0.20935756
2     always 0                            -6.039966          0.07212937
          z        P>|z| normal-based 95% conf. int.
1 -35.73075 6.58444e-280    -7.890844      -7.070163
2 -83.73796  0.00000e+00    -6.181340      -5.898593

[[2]]
  Intervention G-computation estimate of cum. incidence Bootstrap Std. Err.
1     always 1                              0.01117491          0.00206512
2     always 0                              0.04553079          0.00322396
          z P>|z| normal-based 95% conf. int.
1  5.411267     1   0.00712728     0.01522255
2 14.122628     1   0.03921183     0.05184975

> testG[[2]]
[[1]]
  Intervention G-computation estimate of av. log IR Bootstrap Std. Err.
1     always 1                            -7.475187          0.18591828
2     always 0                            -6.065342          0.09601562
          z P>|z| normal-based 95% conf. int.
1 -40.20684     0    -7.839587      -7.110787
```

```
2 -63.17036     0    -6.253533      -5.877151

[[2]]
  Intervention G-computation estimate of cum. incidence Bootstrap Std. Err.
1     always 1                                0.01119003         0.001942735
2     always 0                                0.04445788         0.004073489
          z P>|z| normal-based 95% conf. int.
1  5.759933     1  0.007382265     0.01499779
2 10.913955     1  0.036473838     0.05244192

> testG[[3]]
[[1]]
  Intervention G-computation estimate of av. log IR Bootstrap Std. Err.
1     always 1                             -7.466646          0.17605447
2     always 0                             -6.037789          0.08523415
          z P>|z| normal-based 95% conf. int.
1 -42.41100     0    -7.811713      -7.121579
2 -70.83767     0    -6.204848      -5.870730

[[2]]
  Intervention G-computation estimate of cum. incidence Bootstrap Std. Err.
1     always 1                                0.01125803         0.001929361
2     always 0                                0.04556101         0.003681809
          z P>|z| normal-based 95% conf. int.
1  5.835108     1  0.007476481     0.01503958
2 12.374628     1  0.038344668     0.05277736

> testG[[4]]
[[1]]
  Intervention G-computation estimate of av. log IR Bootstrap Std. Err.
1     always 1                             -7.644993           0.2168143
2     always 0                             -6.091001           0.1011308
          z         P>|z| normal-based 95% conf. int.
1 -35.26056 1.181824e-272    -8.069949      -7.220037
2 -60.22895  0.000000e+00    -6.289217      -5.892784

[[2]]
  Intervention G-computation estimate of cum. incidence Bootstrap Std. Err.
1     always 1                                0.009512656         0.001967762
2     always 0                                0.043460521         0.004332346
          z     P>|z| normal-based 95% conf. int.
1  4.834251 0.9999993  0.005655842     0.01336947
2 10.031638 1.0000000  0.034969124     0.05195192

> testG[[5]]
```

```
[[1]]
  Intervention G-computation estimate of av. log IR Bootstrap Std. Err.
1     always 1                            -7.468289           0.1587379
2     always 0                            -6.025016           0.0915159
          z P>|z| normal-based 95% conf. int.
1 -47.04794     0    -7.779416      -7.157163
2 -65.83573     0    -6.204387      -5.845645

[[2]]
  Intervention G-computation estimate of cum. incidence Bootstrap Std. Err.
1     always 1                                0.01122025         0.001876166
2     always 0                                0.04615791         0.003921385
          z P>|z| normal-based 95% conf. int.
1  5.980413     1  0.007542963     0.01489754
2 11.770819     1  0.038472000     0.05384383

> testG[[6]]
[[1]]
  Intervention G-computation estimate of av. log IR Bootstrap Std. Err.
1     always 1                            -7.556188          0.18256598
2     always 0                            -5.992745          0.08906894
          z P>|z| normal-based 95% conf. int.
1 -41.38881     0    -7.914017      -7.198359
2 -67.28210     0    -6.167320      -5.818170

[[2]]
  Intervention G-computation estimate of cum. incidence Bootstrap Std. Err.
1     always 1                                0.01037401         0.001874551
2     always 0                                0.04788062         0.003798476
          z P>|z| normal-based 95% conf. int.
1  5.534129     1  0.006699888     0.01404813
2 12.605219     1  0.040435607     0.05532563

> testG[[7]]
[[1]]
  Overall estimate of log incidence rate Within-imputation variance
1                             -7.507124                 0.03631950
2                             -6.051823                 0.00805782
  Between imputation variance Overall std. error confidence interval
1               0.0059707682         0.20852919           -7.919218
2               0.0006934452         0.09428655           -6.237112
  based on student t rate of missing information
1          -7.095029                  0.17588146
2          -5.866533                  0.09754873
```

```
[[2]]
  Overall estimate of cumulative incidence Within-imputation variance
1                               0.01087117                3.830692e-06
2                               0.04503362                1.493788e-05
  Between imputation variance Overall std. error confidence interval
1                5.777517e-07         0.002126968           0.00667256
2                1.148870e-06         0.004039372           0.03709946
  based on student t rate of missing information
1         0.01506979                   0.16302106
2         0.05296779                   0.08774426
```

C.5 Simulated datasets with intervention $\bar{1}$ and intervention $\bar{0}$

G-computation is applied twice to the six datasets (for explanations see Section 4.2) with intervention $\bar{1}$ and $\bar{0}$. For each simulated dataset we provide the distribution of death across the baseline variables.

		intervention $\bar{1}$		**intervention $\bar{0}$**	
	total	deaths	% deaths	deaths	% deaths
total	2647	33	1.25	133	5.03
male	1240	18	1.45	59	4.76
female	1407	15	1.07	74	5.26
less adv. stage	711	4	0.56	10	1.41
adv. stage	1936	29	1.50	123	6.35
age ≤ 30	1744	23	1.32	96	5.51
$30 <$ age ≤ 39	744	9	1.21	30	4.03
age ≥ 40	159	1	0.63	7	4.40
CD4 ≤ 50	357	9	2.52	16	1.96
$50 <$ CD4 ≤ 99	633	13	2.05	56	8.85
$100 \leq$ CD4 ≤ 199	840	5	0.60	25	2.98
$200 \leq$ CD4	817	6	0.73	16	10.08

Table C.2: Total number of patients and number of deaths in each category of baseline variables in simulated dataset 1 with intervention $\bar{0}$ and intervention $\bar{1}$.

		intervention $\bar{1}$		**intervention $\bar{0}$**	
	total	deaths	% deaths	deaths	% deaths
total	2647	32	1.21	133	4.27
male	1240	16	1.29	63	5.08
female	1407	16	1.14	50	3.55
less adv. stage	711	2	0.28	8	1.13
adv. stage	1936	30	1.55	105	5.42
age ≤ 30	1744	24	1.38	68	3.90
$30 <$ age ≤ 39	744	7	0.94	36	4.84
age ≥ 40	159	1	0.63	9	5.66
CD4 ≤ 50	357	6	1.68	19	5.32
$50 <$ CD4 ≤ 99	633	14	2.21	50	7.90
$100 \leq$ CD4 ≤ 199	840	4	0.48	36	4.29
$200 \leq$ CD4	817	8	0.98	8	0.98

Table C.3: Total number of patients and number of deaths in each category of baseline variables in simulated dataset 2 with intervention $\bar{0}$ and intervention $\bar{1}$.

		intervention $\bar{1}$		**intervention $\bar{0}$**	
	total	deaths	% deaths	deaths	% deaths
total	2647	24	0.91	129	4.87
male	1240	15	1.21	61	4.92
female	1407	9	0.64	68	4.83
less adv. stage	711	1	0.14	12	1.69
adv. stage	1936	23	1.19	117	6.04
age ≤ 30	1744	18	1.03	83	4.76
$30 <$ age ≤ 39	744	5	0.67	37	4.97
age ≥ 40	159	1	0.63	9	5.66
CD4 ≤ 50	357	4	1.12	24	6.72
$50 <$ CD4 ≤ 99	633	15	2.37	61	9.64
$100 \leq$ CD4 ≤ 199	840	3	0.36	32	3.81
$200 \leq$ CD4	817	2	0.25	12	1.47

Table C.4: Total number of patients and number of deaths in each category of baseline variables in simulated dataset 3 with intervention $\bar{0}$ and intervention $\bar{1}$.

		intervention $\bar{1}$		**intervention $\bar{0}$**	
	total	deaths	% deaths	deaths	% deaths
total	2647	19	0.72	120	4.54
male	1240	11	0.89	62	5.00
female	1407	8	0.57	58	4.12
less adv. stage	711	3	0.42	16	2.25
adv. stage	1936	16	0.83	104	5.37
age ≤ 30	1744	14	0.80	78	4.47
$30 <$ age ≤ 39	744	4	0.54	38	5.11
age ≥ 40	159	1	0.63	4	2.52
CD4 ≤ 50	357	3	0.84	24	6.72
$50 <$ CD4 ≤ 99	633	10	1.58	53	8.37
$100 \leq$ CD4 ≤ 199	840	4	0.48	33	3.93
$200 \leq$ CD4	817	2	0.25	10	1.22

Table C.5: Total number of patients and number of deaths in each category of baseline variables in simulated dataset 4 with intervention $\bar{0}$ and intervention $\bar{1}$.

		intervention $\bar{1}$		**intervention $\bar{0}$**	
	total	deaths	% deaths	deaths	% deaths
total	2647	26	0.98	114	4.31
male	1240	12	0.97	59	4.76
female	1407	14	1.00	55	3.91
less adv. stage	711	2	0.28	15	2.11
adv. stage	1936	24	1.24	99	5.11
age ≤ 30	1744	15	0.86	76	4.36
$30 <$ age ≤ 39	744	9	1.21	35	4.70
age ≥ 40	159	2	1.26	3	1.89
CD4 ≤ 50	357	5	1.40	33	9.24
$50 <$ CD4 ≤ 99	633	14	2.21	48	7.58
$100 \leq$ CD4 ≤ 199	840	4	0.48	25	2.98
$200 \leq$ CD4	817	3	0.37	8	0.98

Table C.6: Total number of patients and number of deaths in each category of baseline variables in simulated dataset 5 with intervention $\bar{0}$ and intervention $\bar{1}$.

		intervention $\bar{1}$		**intervention $\bar{0}$**	
	total	deaths	% deaths	deaths	% deaths
total	2647	29	1.10	146	5.52
male	1240	16	1.29	72	5.81
female	1407	13	0.92	74	5.26
less adv. stage	711	1	0.14	16	2.25
adv. stage	1936	28	1.45	130	6.72
age ≤ 30	1744	24	1.38	92	5.28
$30 <$ age ≤ 39	744	5	0.67	47	6.32
age ≥ 40	159	0	0	7	4.40
CD4 ≤ 50	357	13	3.64	30	8.40
$50 <$ CD4 ≤ 99	633	13	2.05	70	11.06
$100 \leq$ CD4 ≤ 199	840	2	0.24	26	3.10
$200 \leq$ CD4	817	1	0.12	20	2.45

Table C.7: Total number of patients and number of deaths in each category of baseline variables in simulated dataset 6 with intervention $\bar{0}$ and intervention $\bar{1}$.

Bibliography

[1] World Health Organization. Antiretroviral Therapy for HIV Infection in Adults and Adolescents: Recommendations for a public health approach. - 2006 revision. Technical report, World Health Organization, 2006.

[2] World Health Organization. HIV DRUG RESISTANCE. Fact sheet, World Health Organization, April 2011.

[3] World Health Organization. Interim WHO Clinical Staging of HIV/AIDS and HIV/AIDS case definitions for surveillance. Technical report, World Health Organization, 2005.

[4] J. M. Snowden, S. Rose, K. M. Mortimer. Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique. *American Journal of Epidemiology*, 173(7):731–738, 2011.

[5] S. L. Taubman, J. M. Robins, M. A. Mittleman, M. A. Hernán. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, 38(6):1599–1611, 2009.

[6] J. Robins, M. Hernán, U. Siebert. *Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors*, chapter 18: Effects of Multiple Interventions. World Health Organization Geneva, 2004.

[7] R. M. Daniel, B. L. De Stavola, S.N. Cousens. gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. *Stata Journal*, 11(4):479–517, 2011.

[8] K. Peter. Marginal Structural Models and Causal Inference. Master's thesis, ETH Zurich, Seminar for Statistics, 2010.

[9] T. Gsponer et al. The Causal Effect of Switching to Second-line ART in Programmes without Access to Routine Viral Load Monitoring. *AIDS*, 26(1):57–65, 2012.

[10] J. Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–710, 1995.

[11] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[12] J. Robins. A new approach to causal inference in mortality studies with a sustained exposure period–Application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

[13] J. M. Robins, S. Greenland, F. C. Hu. Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome: Rejoinder. *Journal of the American Statistical Association*, 94:708–712, 1999.

[14] J. M. Robins, M. A. Hernán. *Longitudinal Data Analysis*, chapter 23: Estimation of the causal effects of time-varying exposures. 2009.

[15] S. van Buuren, K. Groothuis-Oudshoorn. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 45(3), 2011.

[16] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, 1987.