MASTER'S DEGREE THESIS

Interuniversity Master in Statistics and Operations Research

Title: Missing Data in Clinical trials.

Author: Marcella Marinelli

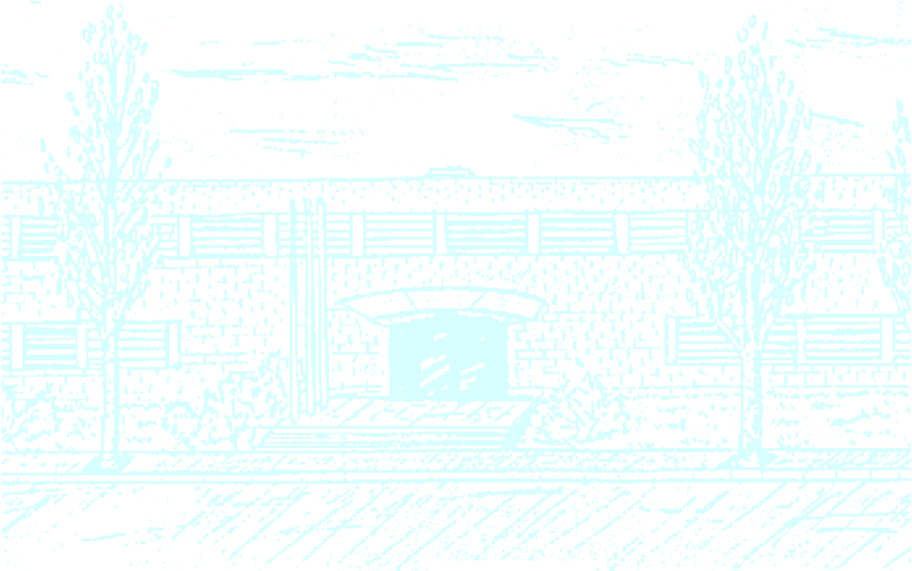
Advisor: Nuria Perez Alvarez

Co-advisor: Guadalupe Gómez Melis

Department: Statistics and Operation Research

University: Universitat Politècnica de Catalunya and Universitat de Barcelona

Academic year: 2010/2011



Facultat de Matemàtiques i Estadística
UNIVERSITAT POLITÈCNICA DE CATALUNYA

UNIVERSITAT DE BARCELONA

Missing Data in Clinical Trials

Master Thesis in
Interuniversity Master in Statistics and Operations Research

Universitat Politecnica de Catalunya

Author:
Marcella Marinelli

Main Advisor:
NURIA PEREZ ALVAREZ
Department of Statistics and Operations Research (UPC)
Fundacio' de Lluita contra la SIDA

Co-Advisor:
GUADALUPE GOMEZ MELIS
Department of Statistics and Operations Research (UPC)
Universitat Politecnica de Catalunya

June 13, 2011

Contents

Chapter 1

Introduction

1.1 Motivation

It is common for longitudinal clinical trials to face problems with missing data that occur when patients do not complete the study or lose some visit. In HIV clinical trials, blood plasma HIV-1 RNA concentration (”viral load”) and CD4 are markers of the clinical evolution of HIV infected patients and are used as primary and secondary endpoints. In fact clinicians are interested in evaluating the percentage of patients with HIV-1 RNA less than copies/mL at week 48 (in general lower values than 50 are unmeasurable and they are refered as undetectable viral load) and the virological changes in CD4 at time 48 weeks for determining the effectiveness of the antiretroviral drugs compared in the trial. Problems arise in the presence of missing data for these variables. Although it has been shown that missing data are a source of bias and may led to different results when comparing treatments effectiveness, the recommended method from the European and American guidelines for clinical trials is to exclude them or to consider an intention to treat analysis. In general, whenever the percentage of missing values is lower than 5%, the effect on bias is negligible.

There is a need to plan an appropriate statistical analysis of missing data. Based on it, the main objectives of this thesis are:

- Review the different typologies of missing data in clinical trials in HIV.
- Explore the most important methods used to handle missing data, describe the advantages and disadvanteges, apply some of these methods to the Lluta Fundacio’ study and implement in the R environment one of the new

approaches.

- Generate some guidelines that should be followed by researchers working in clinical trials (in particular in the Lluita Fundacio') who have to analyse data sets with missing data and give some recommendation when it would be better to use a method instead of another according to the estimates and confidence intervals.

In order to achieve these objectives:

1. I have introduced some useful concepts used in the HIV literature, described the different HIV stages, the HIV diffusion pattern in developed countries and overviewed the antiretroviral treatments at the moment in use.

2. I have defined the causes of dropouts and the sources of missings data and provided a review of the state of the art in the development of the strategies used to handle missing data in clinical trials in HIV from later '90 up to now.

3. I have analized data and compared the different techniques according to some of the statistical values obtained.

4. I have given some conclusions and planned the future steps of the analysis.

1.2 Definitions: HIV-1 RNA viral load and CD4

In a randomized control trial in HIV, measurement of blood plasma HIV-1 RNA concentration ("viral load") and CD4 are markers of the clinical evolution of HIV infected patients. We summarize them briefly:

Viral load is a measure of the severity of an infection such as HIV, cytomegalovirus, hepatitis B and hepatitis C viruses. In HIV the units of interest are copies of the virus in a milliliter (mL) of blood. Changes in viral load are usually reported as a log change (in powers of 10). For example, a three log increase in viral load (3 Log10) is an increase of 103 or 1000 times the previously reported level, while a drop from 500,000 to 500 copies would be a three-log-drop (also 3 Log10).

CD4 cells are a type of lymphocyte (white blood cell) and they are an important part of the immune system. CD4 cells are sometimes called T-cells. There are two main types of CD4 cells: T-4 cells, also called CD4+, are "helper" cells. They lead the attack against infections. T-8 cells (CD8+) are ″suppressor″ cells that complete the immune response and CD8 cells can also be ″killer″ cells that kill cancer cells and cells infected with a virus. When someone is infected with HIV but has not started treatment, the number of CD4 cells they have goes down. This is a sign that the immune system is being weakened. The lower the CD4 cell count, the more likely the person will get sick.

1.3 HIV

Human immunodeficiency virus (HIV) is a lentivirus (a member of the retrovirus family) that causes acquired immunodeficiency syndrome (AIDS), a condition in humans in which the immune system begins to fail, leading to life-threatening opportunistic infections. Infection with HIV occurs by the transfer of blood, semen, vaginal fluid, pre-ejaculate, or breast milk. Viral load (the HIV RNA level) and the CD4 counts are used as markers to asses the patient's health status, as predictive factors for mortality and also their levels are fundamental to determine the treatment efficacy.

In absence of treatment, CD4 counts tend to decrease while the HIV RNA level tends to increase. We may distinguish different stages:

- Primary infection (also known as acute infection): refers to a early stage of HIV. In this stage patients experiment a high viral load RNA levels of > 100.000 copies/m (corresponding to the first maximum in the red tendency in the figure 1.1). This is often accompanied by a dramatic drop in CD4 count. In general, clinicians miss to diagnose HIV in its first stage as the HIV antibody becomes positive just 25 days after infection (this is called HIV seroconversion, converting from HIV negative to HIV positive by blood testing).

- Chronic infection: Chronic HIV infection refers to the period following seroconversion, lasting until the development of symptomatic immune failure and AIDS. This period could variate between weeks and years. This phase is characterized by deteriorating immune function with declining CD4-count. The 75-80% of all patients maintain in the chronical stage during a median period of 7-10 years. The remaining patients, corresponding to 10-15% experience a rapid progression of the disease, reaching AIDS in 2-3 years. Few people (5-10%), have no progression for long time, usually

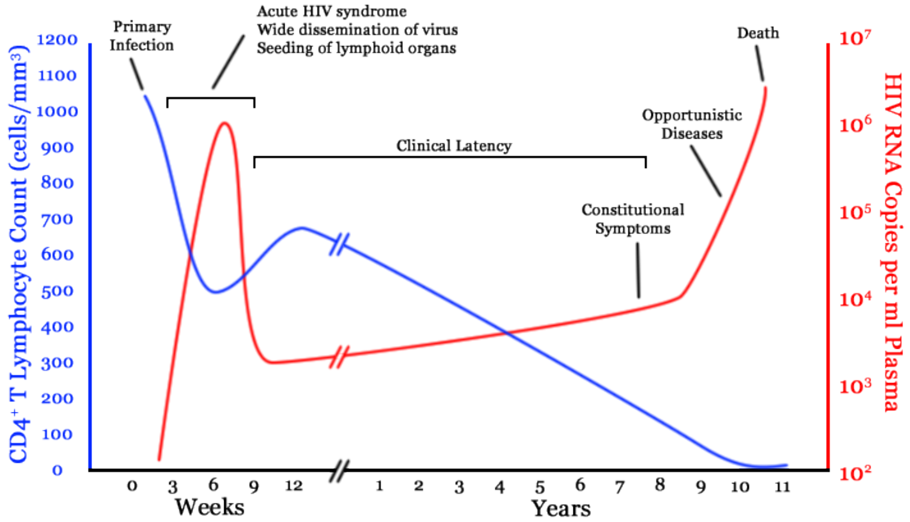


Figure 1.1: *HIV progression*

around 10 years. In some rare case, known in the literature as HIV controllers or elite suppressors, patients do not present the virus in the blood stream, this means that they maintain the load viral below the limit of detection (< 50 HIV RNA -copies/ml) for extended periods. The disease progression depends on the magnitude of load viral peak, load viral decline, viral diversity and so on during the acute stage.

In the figure 1.1 is represented the normal disease progression for the major part of infected people.

1.4 Antiretroviral Treatments

Prior to 1987, no antiretroviral drugs were available and treatment consisted of treating complications from the immunodeficiency. In the last twenty years, we passed from a single drug (AZT) to a dual-drug therapy and, now, to highly active antiretroviral therapy (HAART), consisting in the drug somministration of a series of drugs, usually three or four:

- Nucleoside Reverse Transcriptase Inhibitor (NRTI): NRTIs is analogues of the nucleotides. When they are incorporated into the viral DNA their slightly different structure provokes the blockage of DNA synthesis and therefore the termination of DNA chain assembling. All NRTIs and NtRTIs are classified as competitive substrate inhibitors.

- Protease inhibitor (PI): The HIV protease is an essential enzyme for virus replication. Its activity is to cut the precursors of the virus proteins during the HIV life-cycle. Protease inhibitors are molecules which are able to bind the enzyme and blocking its activity and their use results in protease inactivity and uninfectious viruses.

  or

- non-nucleotide reverse transcriptase inhibitor (NNRTI): the NNRTIs block the enzyme reverse transcriptase, this is achieved by binding at a different site on the enzyme, inhibiting the reverse transcriptase activity. NNRTIs are classified as non-competitive inhibitors of reverse transcriptase.

When antiretroviral drugs were introduced, most clinicians agreed that HIV positive patients with low CD4 counts should be treated. So patients undergo treatments when the CD4 count reaches a low point, around 350 cells per microlite or plasma HIV ribonucleic acid (RNA) levels of > 55.000 copies/mL. For the limit cases (cell counts between 350 and 500) the HAART therapy is recommended in the 50% of cases. Otherwise treatment should be highly reccomended in case of hepatitis C and B coinfection requiring therapy, HIV-associated nephropathy or other specific organ deficiency, age higher than 50, pregnancy or malignancy. No consensus formed as to whether to treat patients with high CD4 counts (> 500 cells/mm3). The 50% view initiating therapy at this stage as optional. In this case clinicians have to evaluate the potential benefits and risks of initiating a therapy as the improvement in the quality of life.

The initial goal of therapy in primary stage, are evaluated through plasma HIV RNA levels, which are expected to indicate a 1.0 log10 decrease at 2-8 weeks and to suppress the HIV viral load to undetectable levels (< 50 copies/mL) at 4-6 months after treatment initiation. The current guidelines for the clinical management and treatment of HIV-infected adults in Europe, consider antiretroviral therapy, the best therapy to treat infected individual at the moment. In 2003 and 2009 the WHO and the United States Department of Health and Human Services Use, established a series of criteria to consider starting HAART based on the stage of the infection (determined by the number of CD4 counts). In particular the recommended goal for patients starting an antiretroviral drugs is the achievement of a viral load (HIV RNA level)< 50 copies/ml within 16 or 24 weeks and the maintenance of such level. Even among people who respond well to HAART, the treatment does not get rid of HIV. The virus continues to reproduce but at a slower pace. The goal of HAART therapy is to improve the patient survival and the quality of life.

There are at the moment 28 drugs in common use for HIV treatment. Treatment success needs strict lifelong drug adherence. Although the widely used drugs are generally well tolerated, most have some short-term toxic effects and all have the potential for both known and unknown long-term toxic effects.

1.5 HIV origin and diffusion around the world

The origin of AIDS in men is relative new and the first cases recognised as AIDS occured in USA in the early 1980s. In March 1981 there was an evidence between an homosexual comunity of New York of a more aggressive form of infection and cancers that seem to be resistent to any treatment available at the moment. Later, clinicians showed a strong connection between these cases. The discovery of the Human Immunodeficiency Virus (HIV), was made soon after and was considered the source of AIDS. Since the 1981 the number of AIDS cases increased dramatically around the world. At the moment there are two types of HIV: HIV-1 and HIV-2, that could be transmitted sexually, through blood, and from mother to progeny at birth. HIV-1 is the most prevalent form while the HIV-2 is less easily transmitted and is highly concentrated in West Africa.

A virus similar to HIV have been found in cats, sheeps, horses and cattles but the most important for investigation of the origin of HIV is the Simian Immunodeficiency Virus (SIV) that affects monkeys, which is believed to be at least 32,000 years old. In 1999 a group of researchers from the University of Alabama found a SIV's virus in chimpanzees almost identical to HIV-1 and claimed that chimpanzees could have in some way brought the HIV between men. A recent theory note as the ″hunter theory″ affirms that the SIV was transferred to humans as a result of chimps being eaten or through their blood enter in contact with hunters. Another theory note as ″colonialism theory″ concentrated the attention on the colonial countries. In these areas slaves were forced to work in really bad conditions. In such conditions sick chimpanzees with SIV could have been an extra source of food for the workers.

While the origin of AIDS is still controversial and under discussion, its diffusion is no longer isolated. UNAIDS, the Joint United Nations Programme on HIV/AIDS (is an innovative partnership that leads and inspires the world in achieving universal access to HIV prevention, treatment, care and support) estimates that the absolute number of people living with HIV rose from around 8 million in 1990 to 33 million by the end of 2009 (Figure 1.2)

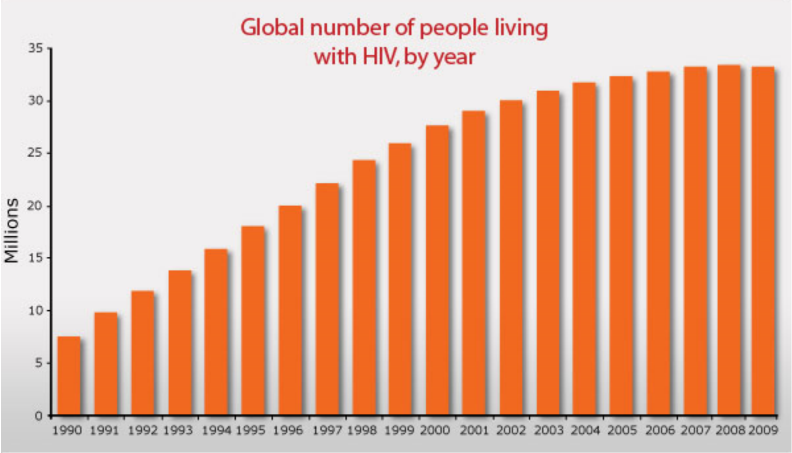


Figure 1.2: *HIV trend*

The increase in the number of persons living with HIV is related with the increase in people receiving antiretroviral therapy and with the reduction in the number of AIDS-related deaths. So it is not surprising that the number of new cases in HIV has steadily declined (Figure 1.3)

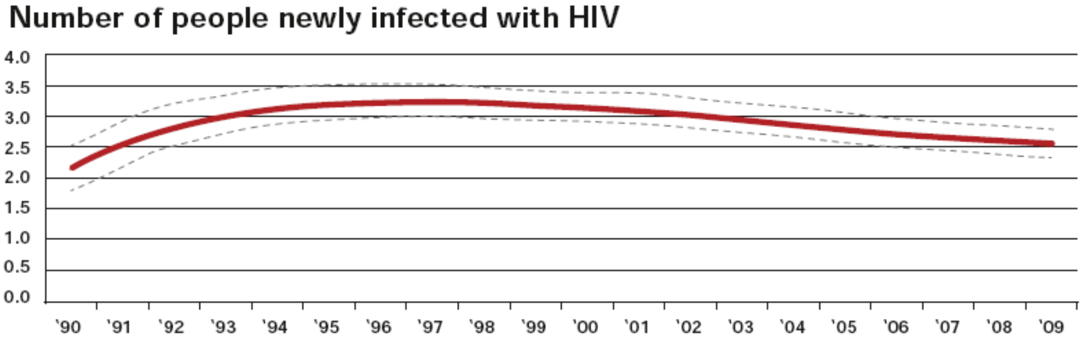
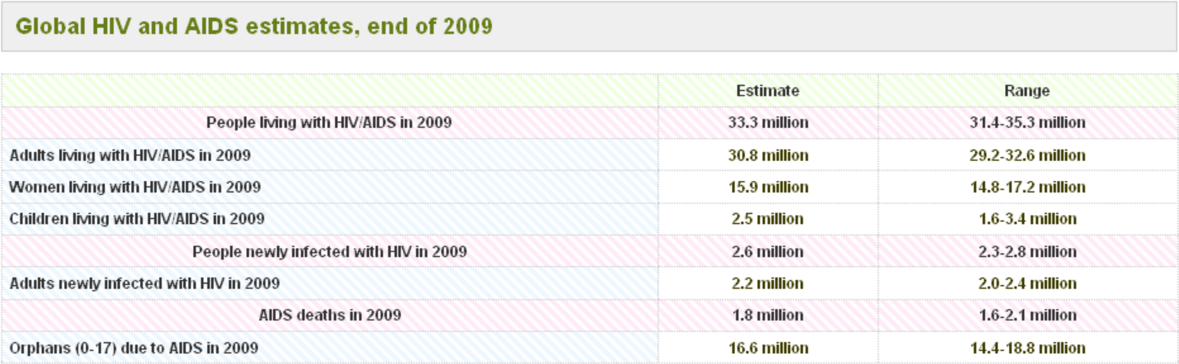


Figure 1.3: *Global HIV estimates*

Figure 1.4 summarizes the estimated number of people infected with HIV or diagnosed with AIDS at the end of 2009 distinguishing between adults, women and children. The number of people newly infected with HIV was 2.6 millions and the number of deaths equal to 1.8 millions for 2009. The number of orphans between 0 and 1 year due to AIDS was estimated to 16.6 millions.

Figure 1.5, summarizes the number of adults and children living with HIV/AIDS, the adult and children newly infected, the number of deaths and the adult prevalences at the end of 2009 around the world. 22.5 millions of adults and children(corresponding to almost the 68 percent of all people living with HIV) live in sub-Saharan Africa, the region carries the greatest burden of the epidemic. Epidemics in Asia have remained relatively stable and are still largerly concentrated

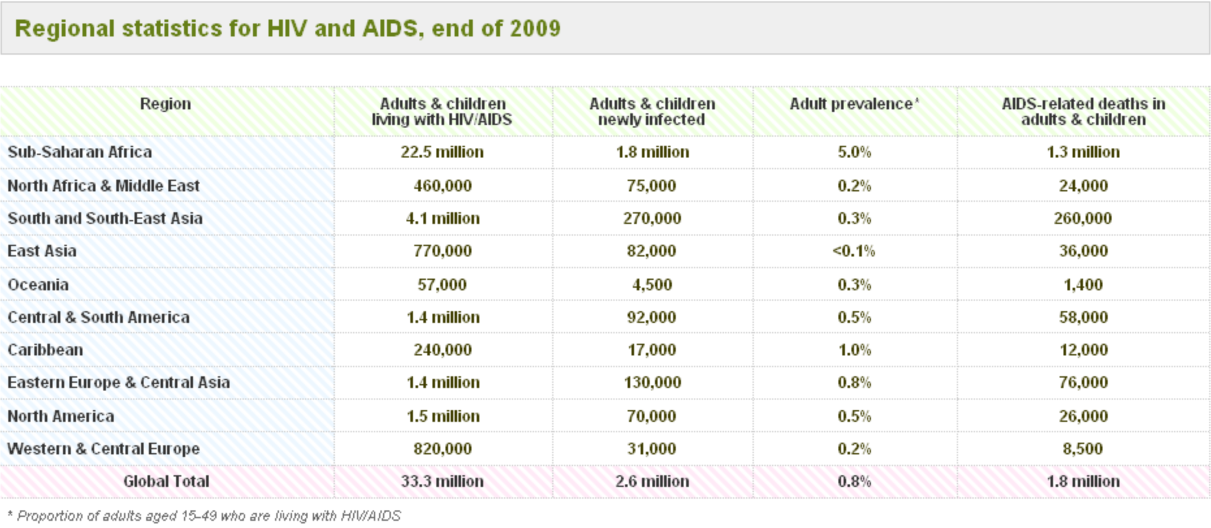


Global HIV and AIDS estimates, end of 2009

	Estimate	Range
People living with HIV/AIDS in 2009	33.3 million	31.4-35.3 million
Adults living with HIV/AIDS in 2009	30.8 million	29.2-32.6 million
Women living with HIV/AIDS in 2009	15.9 million	14.8-17.2 million
Children living with HIV/AIDS in 2009	2.5 million	1.6-3.4 million
People newly infected with HIV in 2009	2.6 million	2.3-2.8 million
Adults newly infected with HIV in 2009	2.2 million	2.0-2.4 million
AIDS deaths in 2009	1.8 million	1.6-2.1 million
Orphans (0-17) due to AIDS in 2009	16.6 million	14.4-18.8 million

Figure 1.4: *Global HIV estimates*

among high risk groups. Conversely, the number of people living with HIV in Eastern Europe and Central Asia has almost tripled since 2000.



Regional statistics for HIV and AIDS, end of 2009

Region	Adults & children living with HIV/AIDS	Adults & children newly infected	Adult prevalence*	AIDS-related deaths in adults & children
Sub-Saharan Africa	22.5 million	1.8 million	5.0%	1.3 million
North Africa & Middle East	460,000	75,000	0.2%	24,000
South and South-East Asia	4.1 million	270,000	0.3%	260,000
East Asia	770,000	82,000	<0.1%	36,000
Oceania	57,000	4,500	0.3%	1,400
Central & South America	1.4 million	92,000	0.5%	58,000
Caribbean	240,000	17,000	1.0%	12,000
Eastern Europe & Central Asia	1.4 million	130,000	0.8%	76,000
North America	1.5 million	70,000	0.5%	26,000
Western & Central Europe	820,000	31,000	0.2%	8,500
Global Total	33.3 million	2.6 million	0.8%	1.8 million

* Proportion of adults aged 15-49 who are living with HIV/AIDS

Figure 1.5: *HIV trend around the world*

The number of deaths for HIV reduced drastically in particular in sub-Saharian regions after the introduction of antiretroviral therapy since 2003. The future step for reducing the number of death of about 10 millions consists of guarantee the access to life-safe medicines.

1.6 Clinical Trial

Many areas of HIV/AIDS research involve clinical trials. A clinical trial is a research study involving patients which may be conducted by universities, hospitals, pharmaceutical agencies and others. Patients are randomly assigned to two or three different treatments for evaluating the effects of new or existing drugs. The main reason for using a randomized clinical trial is to avoid bias in the allocation of the patients to each treatment. Each patient has the same probability of receiving any of the intervention under study. So neither the participant nor the investigator will know in advance the allocation assignment.

A trial is said to be blind whether the subjects involved in the study as clinicians, statistician, person responsible for the treatment administration and patients do not know the treatment allocation. In particular we may distinguish:

- single blind: when just the patient does not know the treatment allocation
- double blind: when patient and also the researchers do not know which treatment is being given to any given subject.
- triple blind: when in the experiment, neither the subject nor the person administering the treatment nor the person evaluating the response to treatment knows which treatment any particular subject is receiving.

Clinical trials could have different trials design:

- Superiority trial: the Superiority trial has the aim of showing that one treatment (the new one) is superior (better) to another (the older one).

  $\mu_N - \mu_R > \varepsilon$

where μ_N is the average effect of the new treatment and μ_R the average effect of the old treatment and ε represents a small value higher than zero. In mathematical terms this means that the average effect of the new treatment is higher than the one of the old treatment for a small value equal to ε.

- Equivalent: In equivalent trials, a new treatment is equivalent to another one whether the difference in its effects not reach a certain value:

  $|\mu_N - \mu_R| < \varepsilon$

  or

  $\varepsilon_1 < \mu_N - \mu_R < \varepsilon_2$

  where ε_1 and $< \varepsilon_2$ are values higher than zero with $\varepsilon_2 > \varepsilon_1$

- Non inferiority: the new treatment is at least as good as the old Treatment. In mathematical terms:

  $\mu_N - \mu_R \geq \varepsilon$

Noninferiority trials are intended to show that the effect of a new treatment is not worse than that of an active drug by more than a tollerance margin denoted by ε.

Due to increased HIV treatment options and ethical concerns (this means that is no longer possible to design a clinical study with placebo versus drug treatment), superiority randomized controlled may be changed for non-inferiority designs.

Before starting a randomized control trial, investigators need to consider a number of design features:

- the number of subjects and duration of follow-up.
- whether the trial will evaluate efficacy or effectiveness or safety. Efficacy trials (explanatory trials) determine whether an intervention has any promise of being as good as or better than existing treatments under ideal circumstances. Effectiveness trials (pragmatic trials) permit to evaluate if treatments can work under real-life conditions. Safety trials permit to collect information about adverse drug reactions and adverse effects of the new drug.
- the phase of the trials:

  i) Phase I: Patients are treated with placebo and a new drug but are uninfected (at low risk of HIV) as the issue of the trial is to evaluate the safety of the new drug. Phase I trials usually last 12-18 months. Permits to define the drug dosis.

  ii) Phase II: is performed on hundreds of participants (200-300) and is designed to assess how the drug works and to better characterize the safety of the treatment. It usually lasts two years.

  iii) Phase III: is performed on thousands of high-risk participants and is designed to assess if the treatment works in preventing HIV infection. Phase III trials can last 3-5 years

  iv) Phase IV: Post marketing studies to better understand the drug's risks, benefits and optimal uses.

There are two preferred approaches to the analysis of most clinical trials :

- Intention to treat analysis: as pointed out by Fisher *et al.* (1990) "a clinical trial includes all randomized patients in the groups to which they were randomly assigned, regardless of their adherence with the entry criteria, regardless of the treatment they actually received, and regardless of subsequent withdrawal from treatment or deviation from the protocol”.

- protocol analysis or on treatment analysis: can only be restricted to the participants who complete the entire clinical trial in the terms of the eligibility, interventions, and outcome assessment.
- available case analysis : analysis are done with the available information of the patient that entered in the clinical trial and followed the trial guidelines.

Chapter 2

Missing data

2.1 Missing Causes

As observed in Chapter 1, a randomized control trial has the advantage to produce unbiased experiments. Patients are followed during a follow up time that could be enough longer for experiensing some dropout. A useful classification of dropouts was given by Rubin (1976) and Little and Rubin (1998):

- Completely at random dropouts (CRD): the drop-out and measurements of CD4 and/or viral load are independent. Example are: a patient moving to another city for non-health reasons, a different disease from the examined outcome, an uncooperative patient, protocolol violation.
- Informative dropout (ID): the dropout may lead to unmeasured information on CD4 and/or viral load and on the final effect of the administrated drugs. Example are: adverse events, death, unpleasant study procedures, lack of improvement and/or early recovery.

In the literature we may refer to two types of missing data patterns:

- intermittent missings (non-monotone missings): in which an observed sequence has missings in one or more points of time and then observations again. Example are: missing visits for practical or administrative reasons, measurement equipment failures.
- monotone missings (dropouts or withdrawals): in which an observed sequence has missings from a time point to the end of the follow-up. Example are: loss to follow-up, lack of efficacy or adverse reaction to the treatment.

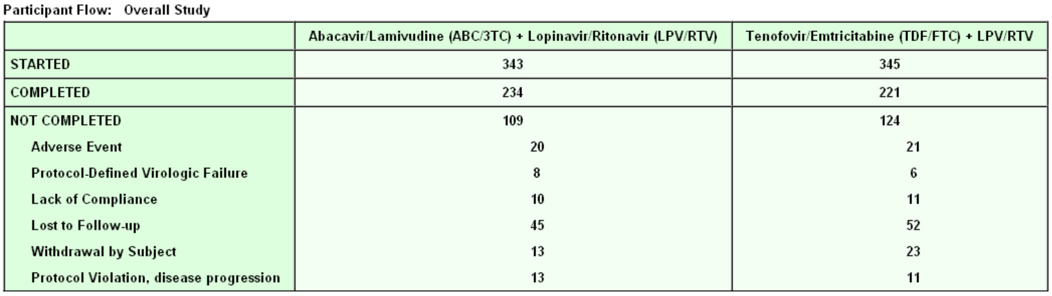
2.2 Missing Causes in clinical trials in HIV

In order to know which are the most common missing data in HIV, we conducted a search of the clinical trials in HIV in the ClinicalTrials.gov registry. ClinicalTrials.gov gives information about a trial's purpose, who may participate, locations, and the centers involved in the study in the United States and around the world. We performed a structured search based parametres: HIV and clinical trials. We found 4804 studies according to these criteria. I reduced the search result to randomized double blind control trials of fase III or IV and I obtained 1219 studies. I considered some examples with published results showing the participants flow for evaluating the common missing data in HIV:

Example 1: A Multicenter randomized double blind control trial with the purpose to evaluate safety and efficacy. The combination of protease inhibitor with two NRTIs has resulted in dramatic decreases in HIV-1-related morbidity and mortality and is currently considered a standard of care regimen for initial treatment of HIV-1-infected patients (Smith, 2009). The study was conducted to evaluate the efficacy and safety of TDF/FTC with ABC/3TC that were both combined with two NRTIs: LPV/RTV.

- emtricitabine/tenofovir + Lopinavir/Ritonavir (TDF/FTC + LPV/RTV)
- abacavir/lamivudine + Lopinavir/Ritonavir (ABC/3TC + LPV/RTV)

Patients are recluted from 76 study sites in the US and 2 study sites in Puerto Rico between 26 July 2005 and 16 June 2006. RNA > 1000 copies/mL at screening was asseted as an inclusion criteria to partecipate in the trial. The primary efficacy endpoint of the trial was to evaluate the percentage of participants with HIV-1 RNA < 50 copies/mL at week 48. The safety endpoint was asset to evaluate the proportion of patients experiensing adverse events over 96 weeks. The study randomized 343 patients in the abacavir/lamivudine + Lopinavir/Ritonavir (ABC/3TC + LPV/RTV) group and 345 in the emtricitabine/tenofovir + Lopinavir/Ritonavir (TDF/FTC + LPV/RTV) group. Patients are followed at baseline (1 day), 2, 6, 12, 18, 24, 32, 40, 48, 60, 72, 84 and 96 weeks or withdrawall. At each visit CD4 and HIV-1 RNA were collected. The flow of participants are shown in the following figure:



Participant Flow: Overall Study

	Abacavir/Lamivudine (ABC/3TC) + Lopinavir/Ritonavir (LPV/RTV)	Tenofovir/Emtricitabine (TDF/FTC) + LPV/RTV
STARTED	343	345
COMPLETED	234	221
NOT COMPLETED	109	124
Adverse Event	20	21
Protocol-Defined Virologic Failure	8	6
Lack of Compliance	10	11
Lost to Follow-up	45	52
Withdrawal by Subject	13	23
Protocol Violation, disease progression	13	11

Figure 2.1: PARTICIPANTS FLOW

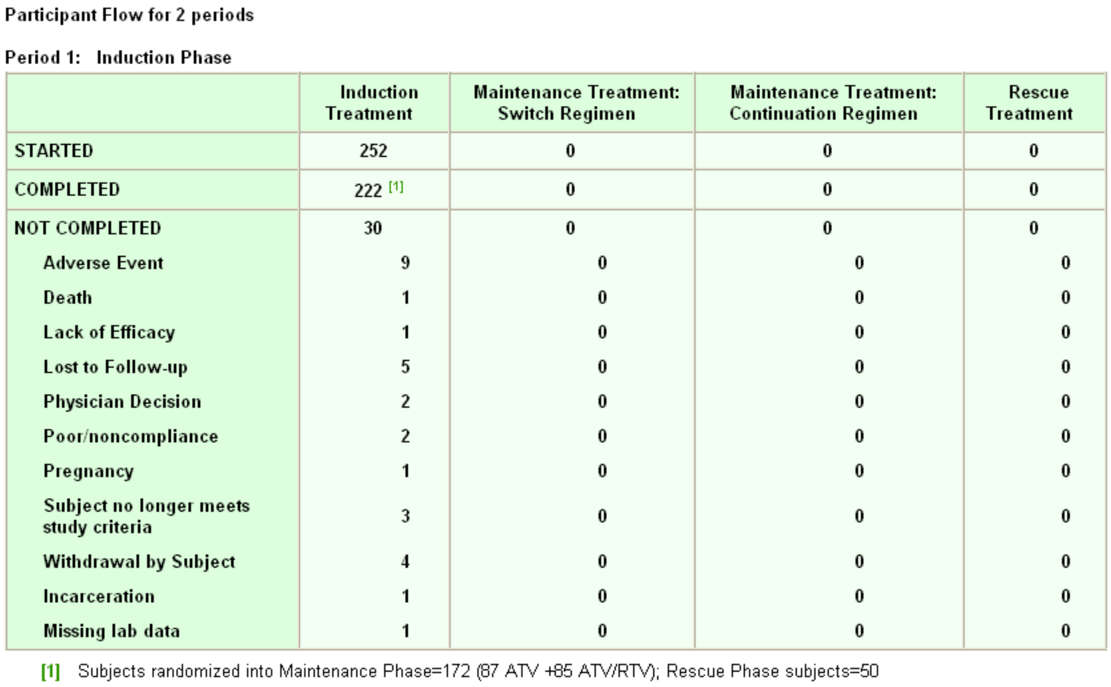
As shown in figure 2.1, just 234 participants in the first group and 221 participants in the second group completed the trial. As summarized in the not completed part of the figure, there are different causes for having missing observations: loss during follow-up, lack of compliance, protocol defined virological failure, protocol violation, adverse event. Some of them may be thought to be independent with the primary endpoint as for example withdrawal by subject. Adverse events may or may not hide an informative nature for the primary enpoint: in some cases patients withdraw because of an adverse reaction to the therapy but they reach an HIV-1 RNA below the undetectable limit. So we may talk of a failure in tolerability but not in the efficicay of the treatment itself. Virologic failure was defined in the study as either failure to achieve HIV-1 RNA below 200 c/ml or confirmed rebound to 200 c/ml after reduction to below 50 c/ml by week 24. After week 24, virologic failure was defined as a confirmed HIV-1 RNA rebound to 200 c/ml. In both cases virological failure is an informative missing example. Loss to follow up could be also dependent to the efficacy endpoints as a patient could decide to abandon a study because of its perception of lack of results in reducing their HIV-1 RNA levels.

Example 2: A multicenter randomized double blind control trial assessing the efficacy of a Treatment maintenance phase with unboosted vs. boosted reyataz after an induction phase with reyataz and ritonavir in treatment naive HIV patients. The trial is divided into two phases:

- Induction Phase: from 26 to 33 weeks. Patients are treated with Atazanavir + Ritonavir + 2 NRTIs.
- Maintenance Phase: from the end of the induction phase. Patients are treated with:
  - Atazanavir + 2 NRTIs
  - Atazanavir + Ritonavir + 2 NRTIs

The primary outcome of the study was to evaluate the percentage of participants with HIV-1 RNA < 50 copies/mL (c/mL) through week 48 of the maintenance phase. Inclusion criteria were: treatment naive HIV-1 infected subjects (< 10 days of treatment with any ARV), subjects who have an HIV-1 RNA level = 5000 c/mL at screening. and subjects who have a CD4 count = 50 cells/mm3.

The flow of participants for the two phases are shown in the following figures:



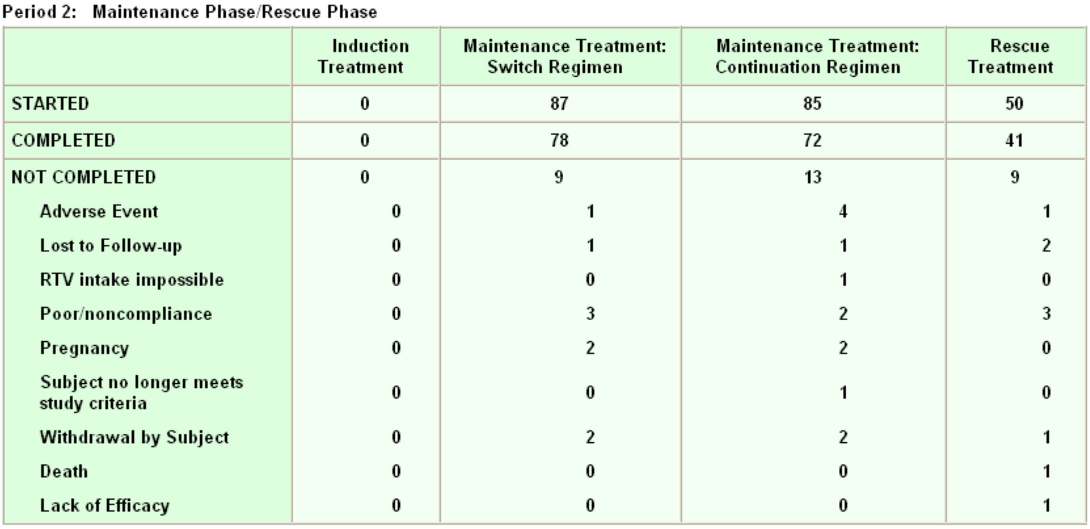
Participant Flow for 2 periods

Period 1: Induction Phase

	Induction Treatment	Maintenance Treatment: Switch Regimen	Maintenance Treatment: Continuation Regimen	Rescue Treatment
STARTED	252	0	0	0
COMPLETED	222 [1]	0	0	0
NOT COMPLETED	30	0	0	0
Adverse Event	9	0	0	0
Death	1	0	0	0
Lack of Efficacy	1	0	0	0
Lost to Follow-up	5	0	0	0
Physician Decision	2	0	0	0
Poor/noncompliance	2	0	0	0
Pregnancy	1	0	0	0
Subject no longer meets study criteria	3	0	0	0
Withdrawal by Subject	4	0	0	0
Incarceration	1	0	0	0
Missing lab data	1	0	0	0

[1] Subjects randomized into Maintenance Phase=172 (87 ATV +85 ATV/RTV); Rescue Phase subjects=50

Figure 2.2: *PARTICIPANTS FLOW*



Period 2: Maintenance Phase/Rescue Phase

	Induction Treatment	Maintenance Treatment: Switch Regimen	Maintenance Treatment: Continuation Regimen	Rescue Treatment
STARTED	0	87	85	50
COMPLETED	0	78	72	41
NOT COMPLETED	0	9	13	9
Adverse Event	0	1	4	1
Lost to Follow-up	0	1	1	2
RTV intake impossible	0	0	1	0
Poor/noncompliance	0	3	2	3
Pregnancy	0	2	2	0
Subject no longer meets study criteria	0	0	1	0
Withdrawal by Subject	0	2	2	1
Death	0	0	0	1
Lack of Efficacy	0	0	0	1

Figure 2.3: *PARTICIPANTS FLOW*

In the previous example, we may observe a low percentage in the number of missing observations. There are different causes for dropouts. As shown in the example 1, the nature of the missing causes is important to determine whether the missing information may be informative or not. In the case of incarceration, laboratory problems, pregnancy and death (when for other causes different from the objective of the study), we may state that missing data are independent from the primary endpoints of the study. For the other causes, we should make the same considerations presented in the example 1.

2.3 Brief overview of missing data mechanisms

Rubin (1976) and Little and Rubin (2002) have given a review of the different parameter modeling frameworks that can be used for modeling the following data density:

$$f(Y_i, R_i \mid X_i, W_i, \theta, \psi) \tag{2.1}$$

where X_i, W_i are vectors of covariates for the measurements and for the missings and θ and ψ are the corresponding parameter vectors. Let consider a sample of N individuals identified by the indicator i with i=1.....N and a set of measures collected over time j for each unit i, Y_{ij} $(j = 1.....n_i)$. So for each subject i, we will have a vector $Y_i = (Y_{i1}.....Y_{in_i})$. We define R_{ij} as the dummy variable representing missing value for the individuals i at time j. So in case of longitudinal studies, for each individuals the dummy variable will be identified with the vector $R_i = (R_{i1}.....R_{in_i)}$ and will assume a value equal to:

0 if the outcome is observed at time j

1 if the outcome is missing at time j

The parameterization of the joint distribution of R and Y, can be put into the selection moded, the fixed pattern-mixture models or shared parameters models:

Selection models:

$$f(Y_i, R_i \mid X_i, W_i, \theta, \psi) = f(Y_i \mid X_i, \theta) f(R_i \mid Y_i, W_i, \psi) \tag{2.2}$$

Y_i is divided into two parts Y_o and Y_m. In this case, we will have a model for the observed mechanism Y_i and a model for the missing mechanism R_i.

Pattern mixture models:

$$f(Y_i, R_i \mid X_i, W_i, \theta, \psi) = f(Y_i \mid R_i, X_i, \theta) f(R_i \mid W_i, \psi) \tag{2.3}$$

a distribution probability for the missing data patterns and a different response model Y_i for each pattern of missing data R_i

Shared parameters models:

$$f(Y_i, R_i \mid X_i, W_i, \theta, \psi) = f(Y_i \mid X_i, R_i, \theta, b_i) f(R_i \mid W_i, \psi, b_i) \tag{2.4}$$

In case of selection patterns, we may distinguish:

- Missing completely at random (MCAR)

  $$f(R_i \mid Y_o, Y_m, W_i, \psi) = f(R_i \mid W_i, \psi) \tag{2.5}$$

  the probability density function or the probability mass function of being missing does not depend of observed Y_o or missing observation Y_m.

- Missing at random (MAR):

  $$f(R_i \mid Y_o, Y_m, W_i, \psi) = f(R_i \mid Y_o, W_i, \psi) \tag{2.6}$$

  the probability density function or the probability mass function of being missing does depend of observed Y_o. MAR means that a participants probabilities of response may be related only to his or her own set of observed items, a set that may change from one participant to another.

- Missing not at random (MNAR):

  $$f(R_i \mid Y_o, Y_m, W_i, \psi) = f(R_i \mid Y_o, Y_m, W_i, \psi) \tag{2.7}$$

  the probability density function or the probability mass function of being missing does depend of observed Y_o and missing Y_m parts .

MAR is considered ignorable non reponse and MNAR non ignorable. In the section 2.6, we will review some techiniques used to treat missing data that are based on: deleting the missing observations (as in the case of complete case analysis or available case analysis), to fill the missing observations with some assumption (as in the case of hot deck imputation and LOCF). In both case, for being valid and not biased, the MCAR assumption have to be hold.

Let consider a probability density function (or probability mass function in the case of discrete distribution) that depends on X, Y, W, ψ and Θ:

Y $\rightarrow\ f(Y, R|\mathrm{X}, \mathrm{W}, \Theta, \psi)$

Where Θ and ψ are vector of parameters, X and W are vectors of covariants and Y is the outcome. The likelihood function for the all subjects is:

$\theta \rightarrow\ f(Y, R|\mathrm{X}, \mathrm{W}, \Theta, \psi)$

That could be written as:

$$\text{Ł}(\Theta, \psi|\mathrm{X}, \mathrm{W}, \mathrm{Y}, \mathrm{R}) \quad = f(Y, R|\mathrm{X}, \mathrm{W}, \Theta, \psi)$$

In other words, when $f(Y, R|\mathrm{X}, \mathrm{W}, \Theta, \psi)$ is viewed as a function of Y and R with Θ and ψ fixed, it is a probability density function, and when viewed as a function of Θ and ψ with Y and R fixed, it is a likelihood function. We have described a likelihood function in case of values Y observed. Under missing data for the outcome, the likelihood function will be tranformed in the following way:

$$L(\Theta, \psi|\mathrm{X}, \mathrm{W}, \mathrm{Y}_o, \mathrm{R}) \quad = f(Y_o, R|\mathrm{X}, \mathrm{W}, \Theta, \psi)$$

In this case we cannot evaluate the conditional distribution because of the dependency of missing values. Instead we will have:

$$f(Y_o, R \mid \theta, \psi) = \int f(Y, R \mid \theta, \psi) dY_m \tag{2.8}$$

Substituing the joint distribution for the selection model:

$$\int f(Y_o, Y_m \mid X, \theta) f(R \mid Y_o, Y_m, W, \psi) dY_m \tag{2.9}$$

So under the MCAR assumption, we will have:

$$\int f(Y_o, Y_m \mid X, \theta) f(R \mid Y_o, Y_m, W, \psi) dY_m = f(Y_o \mid X, \theta) f(R \mid W \psi) \tag{2.10}$$

as the missing part does not depend of Y_o and Y_m

Under the MAR assumption, the missing value depends on the observed part, in this case:

$$\int f(Y_o, Y_m \mid X, \theta) f(R \mid Y_o, Y_m, W, \psi) dY_m = f(Y_o \mid X, \theta) f(R \mid Y_o, W \psi) \tag{2.11}$$

In the case of MCAR and MAR the separability property is satisfied and the parameter estimation could be based on the maximum likelihood (using the EM algoritm) or a bayesian approach (as we will see for the imputation approach). Under the MNAR assumption no simplification of the joint distribution is possible and this approach is usually not examined fot the lack of statistical software.

2.4 Missing Data approaches in HIV/AIDS clinical trials

In a recent paper Wood *et al.* (2004), review the missing data treatment for the clinical trials published in the most important medical journals. The authors identifies 71 trials of which 63 (89%) reported having some missing outcome data, with 13 of these having more than 20% of patients with missing outcomes. The study suggests that the major part of the trials just report missing data or treated them employing some traditional strategy as for example complete case analysis, available case analysis and last observation carried forward.

2.4.1 Later '90-2000

In clinical trials in HIV, complete case analysis and last observation carried forward were also considered the most used statistical techniques up to the later '90 in case of "attrition bias", that occurs when data are collected over two or more points in time and some participants drop out of the study prematurely:

- Complete case analysis or listwise deletion: the strategy consists at excluding the missing data for each follow up time. For clinical trials this means, for example, to graphically display the change in HIV RNA level from baseline to 48 weeks and reporting the summary statistics as means, medians, ranges and others, just considering the effective number of patients in each follow up time and omitting those cases with missing data. This approach could drastically reduce the sample size and generate biases, in particular when the number of missing data in the treatment groups is unbalanced.
- Last observation carried forward (LOCF): LOCF uses the last value observed before dropout, regardless of when it occurred. Let $Y_a = \{y_{a,1}, y_{a,2} \; ....y_{a,n}\}$ the vector of all potential observations of patient A who dropouts at time k (with k<n), respectively at time 1,2,...n. If for this patient observations $y_{a,k+1}$,.... $y_{a,n}$ are not available (missing), the LOCF methos will replace them by $y_{a,k}$.

In 1998 Cozzi Lepri *et al.* published a paper on HIV clinical trials, reviewing the LOCF method and comparing it with a new method called adjusted mean change from baseline (AMCB). In AMCB method the mean change from baseline is used as characteristic to ″match″ similar patients. We may consider a research study in which investigators are interested at 4-weekly changes in viral load usually reported as a difference between viral load at time point X minus viral load at time point Y. When values are dropouts after a certain time, the missing observations are replaced by values taken from another patient, whose change in log_{10} HIV-RNA level is closest to the value the patient has at the time of dropouts. AMCB is a particular case of ″nearest-neighbor″ or ″hot deck″ imputation, consisiting of replacing missing data by values taken from another patient with similar characteristics, in case of HIV we may think to the value of HIV-1 RNA level that is closest to the value the patient had at the time of dropouts.

So we summarize the hot deck imputation in the following way: indicate with I the number of patients with no missing outcomes values. Consider a patient B who dropouts at time k and another patient C with complete information and similar HIV-1 RNA level at time k. Let $Y_b = \{y_{b,1}, y_{b,2} \; ....y_{b,n}\}$ and $Y_c = \{y_{c,1}, y_{c,2} \; ....y_{c,n}\}$ be the vector of all potential observations of patients B and C, respectively. If the observations $y_{b,k+1}$,.... $y_{b,n}$ are missing, substitute them by $y_{c,k+1}$,.... $y_{c,n}$ (with k<n), chosen minimizing the difference in absolute terms in HIV-1 RNA levels for patient B and patient C at time of dropout $abs(y_{b,k} - y_{c,k}) \quad with \; i = 1, ..., I.$

The nearest neighbour permits to reduce the bias respect to LOCF. The last one assumes constancy over time that may not be justified for HIV data as HIV-1 RNA levels seem to substantial decline in the first two or four weeks from starting an antiretroviral drug.

In 1999, Le Corfec *et al.* reported a study research in which the constancy hipothesis under LOCF is longer valid. Based on Rabould and Montaner study (1997), he suggests that ″HIV-1 RNA maintains relatively constant for weeks 2 to 24 since after an initial decline, plasma RNA levels remain flat for several months or slowly increase for most patients″

Another attempt to compare missing methods in clinical trials in HIV is due to Kelleher *et al.* (2001). They focused on different ways to evaluate HIV-1

RNA level:

- The change of HIV-1 RNA from baseline through week 48 for patients randomized in two different treatments groups A and B. In this case, three different approaches for treating missing data were introduced: available case analysis or pairwise deletion that omits cases which do not have data on a variable used in the current calculation only, complete case analysis and LOCF.
- The proportion of subjects with HIV RNA < of a certain value. Also in this case three approaches were evaluated: worst case scenario considers missing measurements as failures (NC=F), complete case analysis and composite. Composite is an approach suggested by the regulatory authority and define the case in which missings are set equal to failure: i) No confirmed HIV RNA response (two consecutive HIV-1 RNA $<$ 400 c/mL, ii) HIV RNA rebound (two consecutive HIV-1 RNA $>=$ 400 c/mL, after response, iii) Discontinuation of treatment, iv) AIDS event or death.
- The time to treatment failure based on HIV-1 RNA measurements for two treatments using a Kaplan Meier or a Proportional Cox model approach.

2.4.2 2000-2010

In a more recent paper Huson *et al.* (2007) reviewed the newer and older method mechanisms recomended in clinical trials for treating missing values. The major part of these techniques were previously described as LOCF and hot deck imputation with the only exception of:

- baseline carried forward: In a clinical trial where the primary endpoint is the change in HIV-1 RNA level from baseline to week 24 (early virological response at 24 weeks), this method was described by Huson as follow: "it consists at setting the change in HIV RNA level from baseline to 24 weeks to zero for all patients who withdrew from the study prior to week 24". Let $Y_a = \{y_{a,1}, y_{a,2} \;....y_{a,n}\}$ the vector of all potential observations of patients A who dropouts at time k (with k<n),

if the observations $y_{a,k+1}$,.... $y_{a,n}$ are missing, substitute them by $y_{a,1}$ so the change from baseline to the end of the study will be equal to zero.

- Multiple imputation: as pointed out by Huson "It consists of imputing a value for change in HIV-1 RNA levels from baseline to week 24 for all patients who withdrew from the study prior to week 24". This method was used for a no monotone missing data structure. A detailed description of this method will be given in section 2.7.

2.4.3 EMEA-FDA

In the last years the European Medicines Agency (EMEA) and the US Food and Drug Administration (FDA) authorities have given some guidelines for treating missing data. Complete case analysis was recommended in the guidelines offered by the European medicine agency. Furthermore, the complete case analysis violated the intention to treat principle which states: a clinical trial should be based on the initial treatment intent, not on the treatment eventually administrated. Simply use the completers is convenient when we have missing completely at random data. Otherwise the FDA Division has traditionally viewed LOCF as the preferred method of analysis. LOCF produce unbiased estimates in case we assume missing completely at random. It is a better approach compared to complete case analysis conducting to shorter confidence intervals but could be used if the assumption of stability does hold.

In general the LOCF approach seem to be the most used method to handle missing data in HIV. Some authors refer to LOCF as a conservative method. For patients in whom conditions are expected to deteriorate, as in the case of a clinical trial in HIV for patients in an advanced state of the infection, the LOCF is very likely to give optimistic results. For example in the case of a new antiretroviral treatment compared to the drugs currently used. If we have more missing values at 24 weeks for the older treatment and we impute values using the LOCF (for example values of HIV-1 RNA at baseline), the treatment comparison may be biased in favor of the new therapy. Other authors state that considering LOCF as conservative is a common misunderstanding, as there are situations in which this strategy is anticonservative. For instance, consider a randomized trial with the issue to evaluate couples at high risk of HIV infection. We might expect a reduction in risky behavior even in the absence of the randomized experiment. Therefore the LOCF will result in values that look worse than truly are. Differential rates of missing data across the treatment and control groups will result in biased treatment effect estimates that are anticonservative.

2.5 Exploratory data analysis (EDA)

The first step to investigate the nature of the missing values in the HIV outcomes, as for example the viral load, consists at summarising the different types of missings data for the general group of patients and also for the intervention and the control group during the follow up time. To explore if there is a correlation structure of missing observations during time, we may built a series of χ^2 tests to asset if there is a statistically significant association between the presence of missing data at baseline, 12, 24, 36 and 48 weeks. The missing data pattern could be evaluated using different strategies:

- **A logistic regression**: is a generalized linear model used for binomial regression in which two components are specified: a categorical response variable R with two levels (1 =missing and 0 =observed) and a logit link function:

$$log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 time_i + \beta_2 cov2_i + \beta_3 cov3_i + ...\beta_4 g(Y_{ij}) + \varepsilon_i \quad (2.12)$$

$\varepsilon_i \sim N(0, \sigma_\tau)$

- Where $p_{ij} = E(R_{ij} \mid X_i) = P(R_{ij} = 1 \mid X_i)$ and R_{ij} is the presence of the i-th individual measured at time j, with distribution $R_{ij} \sim Bernoulli(p_{ij})$. $g(Y_{ij})$ is the variable response or a trasformation of the variable (after an imputation for instance).

In case of longitudinal studies, the general linear model (GLM) could be changed in favor of a generalized estimation equation model (GEE). The GEE (Linag and Zeger, 1986) method allows for the correlation between observations. In the ordinary logistic regression model, standards errors are based on the assumption that the proposed correlation structure is correct. However, GEE as the property that even if this structure is incorrect, the fixed effect estimates are still consistent. Nevertheless, the naive standard errors may be improved by the use of a sandwich estimator. This gives us the robust standard errors. Confidence interval under sandwich standard errors in GEE, are in general shorter then the confidence intervals in logistic regression models. So we may fit a GEE model with the presence of missing as response to explore the dependence between R and time, some of the covariates known to be associated with the response (cov1, cov2, cov3) and $g(Y_{ij})$:

- If the final model only include the constant we may conclude that missings have a MCAR structure.
- If the final model include some covariates, we will assume a MAR structure.
- If we may include the observed outcomes $g(Y_{ij})$ as variable response, we may assume a MNAR structure.

This approach present a problem: missingness at random is relatively easy to handle, simply include in the model all variables that affect the probability of missingness. Unfortunately, we generally cannot be sure whether data really are missing at random, or whether the missingness depends on unobserved predictors or the missing data themselves. We generally must make assumptions, or check with reference to other studies. In practice, we typically try to include as many predictors as possible in a model so that the missing at random assumption is reasonable.

- **Sensitivity analysis**: fitting different models valid under different scenarios of nonresponse mechanism (MCAR and MAR) and compare the estimations and the confidence intervals. For example compare the estimates obtained from a logistic regression valid under MCAR and a mixed models valid under MAR.

2.6 Simple stochastic imputation

In section 2.4, we have analyzed different simple methods used in the literature to handle missings data in clinical trials. Some of them are used more frequently than other, in particular we may distinguish:

- LOCF : is a simple imputation method used to complete the values of a dataset that have not been recorded with the last measured observation from a certain time point. Under LOCF the estimated mean and variance are biased and consequentlly the treatment estimates are also biased.

- missings as failures: assigning the worst possible value of the outcomes to dropouts for a negative reason (treatment failure). In the Example 1 in section 2.2 the primary endpoints were evaluated considering the missing observations equal to failures outcomes. The use of this approach may be a reasonable starting point in case of virological failure but in case of adverse event for intolerability of a drug is no longer valid.

- missings as success: assigning the best possible value of the outcomes to dropouts for a positive reason (treatment cure)

  The exploration of the ″worst″ and the ″best″ scenarios are common sensitivity analysis approaches for binary outcomes in clinical trials and permit to create a lower and upper bounds for the intervention effect under study. The ″best″, the ″worst″ scenarios and LOCF are unbiased methods under a MCAR assumption.

- Regression model imputation: We begin considering all the variables to be used in the analysis. Then a regression model (logit model in case of binary outcomes) with the outcome of interest as response variable is fitted. We get predictions from the model and use them by randomly assign to the missing values. Finally we use this to impute missing outcomes.

2.7 Multiple Imputation

In the previous section we have described some single imputation methods as for example LOCF and the regression analysis. The multiple imputation (MI) method of missing data was firstly proposed by Little (1982) and applied in Survey studies, the crucial difference respect to a single stochastic imputation based on completing the data once, is that the imputation process is repeated a small number of times. As for LOCF and regression analysis, it permits to analyze the data sets as we would have done if no data were missing, generating multiple copies of the original data set and replacing missing values by randomly generate values. The key idea of MI is to use the data from units where both the outcome Y and the vector of variables X are observed, together with the rest of the X's, to learn about the relationship between Y and X. Once we have imputed m complete data sets, we should analyze each of them in the usual way (i.e. using the model intended for the complete data). We obtained m estimates of the original quantity of interest, Q. Let denote these estimates Q_1,..., Q_m. So, each Q could represent a regression coefficient from a regression model of interest. We may obtain the average of m complete data estimates in the usual way:

$$\hat{Q_{MI}} = \sum_{j=1}^{m} \frac{\hat{Q}_j}{m} \tag{2.13}$$

The variance is composed by two components:

- within imputation variance: which is the average of the m variances

$$\bar{\sigma}_{\omega}^2 = \sum_{j=1}^{m} \frac{\hat{\sigma}_{\mathrm{J}}^2}{m} \tag{2.14}$$

- between imputation variance:

$$\bar{\sigma}_b^2 = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{Q}_j - \bar{Q}_{MI}) \tag{2.15}$$

The total variance is aproximately:

$$T = \bar{\sigma}_{\omega}^2 + \bar{\sigma}_b^2 \left(1 + \frac{1}{m}\right) \tag{2.16}$$

The tests and the confidence interval are based on a Student's approximation:

$\left(\hat{Q_{MI}} - Q\right) / \sqrt{T} \sim t_{\nu}$

Unless rates of missing information are unusually high, there tends to be little or no practical benefit to using more than five to ten imputations. Rubin (1987), showed that the relative efficiency (RE) of an estimate based on m imputations to one based on an infinite number of them is approximately:

$$RE = (1+\lambda/m)^{-1}$$

where λ is the rate of missing information.

λ				
m	10%	30%	50%	70%
3	0.98	0.95	0.93	0.90
5	0.99	0.97	0.95	0.94
20	0.99	0.99	0.98	0.97
∞	1	1	1	1

As we may observe from the table above with a small percentage of missing values and m=3 or m=5, we may obtain a nearly fully efficiency. This information can be summarized also in terms of standard deviation. Let $\sqrt{1+\lambda/m}$ the standard deviation. With $\lambda = 50\%$ missing information, an estimate based on m=5 imputations has a standard deviation that is only about 5% wider than one based on $m = \infty$, with $\sqrt{1+0.5/5} = 1.049$ versus 1.

In the literature the most famous multiple imputation model are:

- Multivariate normal model, firstly introduced by Shafer (1997). It assumes that data are normally distributed. Skewed variables were firstly transformed as no normality distribution can lead to a bias into the analysis.

- Chained equation model using the chained equation approach. The advantage of the chained equation model is that perform a series of univariate regressions rather than a single multivariate normal model so it can be easier to estimate. Moreover variables are not assumed to have a normal distribution, so the regression model can be replaced by some generalized linear model (GLM) for no normal responses. This approach can be summarized in 5 steps: 1) for each variables we fill missing values with randomly chosen observed values 2) once we have the all dataset completely full, we start with the multiple imputation mechanism. We firstly asset a single regression model involving a single response variable, the original missing data for the first variable and multiple predictors corresponding to the other variables in the dataset that were previously imputed as described in step 1. 3) the 'filled in' values in the second variable are discarded. These missing values are then imputed using the regression imputation on all the other variables. 4) this process is repeated for the all number of variables introduced for the imputation analysis. 5) this process is continued for several cycles.

The common assumption for these two methods is that missings have a MAR structure.

2.8 Hot deck multiple imputation

We consider an Hot deck (HD) multiple imputation approach as presented in Tang *et al.* (2005) for Survey Studies using a predictive mean matching method and the approximate Bayesian bootstrap for missing values. The HD method is the non parametric version of the multiple imputation approach.

First of all, it is required to indroduce some useful terms:

i) Propensity score: is the probability of a unit (such as a person) being assigned to a particular condition in a study (for example a treatment), given a set of known covariates. In clinical trials for evaluating if a new treatment is better than the traditional one, we may want to asses the conditional probability of being treated with the new intervention (T=1) given some backgrounds variables X_1,X_2,......,X_p:

$$PS = P(T = 1|X_1, X_2, ..., X_p)$$

We may estimate the propensity score through the logistic regression model:

$Ln(PS/1 - PS) = \beta_0 + \beta_1 X_1 + \ ....\beta_p X_p$ where PS is the propensity score

Solving respect to PS:

$$PS = exp(\beta_0 + \beta_1 X_1 + \ ....\beta_p X_p)/(1 + exp(\beta_0 + \beta_1 X_1 + \ ....\beta_p X_p))$$

In case of missing data, we are interested at evaluating the probability of being observed respect to being not observed. In section 2.3, we have defined R_i a dummy variable that assumes a value equal to zero if the outcome is observed at time j and one if the outcome is not observed at time j.

$$PR_1(X_0) = P(R_1 = 0|X_0)$$

$$PR_2(X_0, Y_1) = P(R_1 = 0|X_0, Y_1)$$

$$PR_T(X_0, Y_1, ...Y_{T-1}) = P(R_1 = 0|X_0, Y_1, Y_{T-1}....)$$

Where X_0 is a vector of baseline characteristics, Y_1 a vector of the response outcome at time 1, Y_2 a vector of the response outcome at time 2, and finally Y_{T-1} a vector of the response outcome at time T-1.

The first equation means that the probability of response at time 1 is conditioned to the baseline covariates vector X_0 into the dataset. The second equation means that the probability of response at time 2 is conditioned to the baseline characteristics (X_0) and to the outcome vector at time 1 (Y_1). The last equation means that the probability of response at time t depends on the baseline characteristics and on the outcome vector at time 1,2 up to T-1. We may employ the same procedure for obtening the probability of no response:

$$PR_1(X_0) = P(R_1 = 1|X_0)$$

$$PR_2(X_0, Y_1) = P(R_1 = 1|X_0, Y_1)$$

$$PR_T(X_0, Y_1, ...Y_{T-1}) = P(R_1 = 1|X_0, Y_1, Y_{T-1}....)$$

We may reconduct the last equation to equation 2.6 in case we consider a probability mass function instead of a probability density function. The probability of being missing depends on the covariates and the oserved outcomes. So in case of

HD nultiple imputation we are assuming a MAR structure.

In the second step, the propensity scores were stratified basing on quartiles and an approximate Bayesian bootstrap was introduced:

i) In each quantiles, we define with nobs the number of observed outcome and with nmiss the number of not observed outcomes. So we randomly sample with replacement n_1 values from the observed responses nobs.

ii) We draw the n_0 = n - n_1 missing outcomes randomly with replacement from the potential set of observed outcomes created in step i.

iii) we repeated step i and ii in each time.

The propensity score method is valid under a monotone missing data structure.

Chapter 3

Analysis of a clinical trial in HIV

The aim of this chapter is to introduce the dataset that has motivated the contribution of this thesis.

3.1 The HIV dataset

International treatment guidelines recommend an antiretroviral therapy conteining 2 nucleoside reverse transcriptase inhibitors (NRTIs) and a boosted PI or a non nucleoside reverse transcriptase inhibitor (NNRTI) in treatment-naive patients. In particular the 2007 european guidelines (EACS) consider a treatment consisting of lopinavir/ritonavir (Lopivavir/r) or Efavirenz as the best choices. The first one for its high antiviral potency, long durability (low risk of resistance) and its acceptable tollerance. The other for its low pill burden, which make adherence easier, and the high number of patients who achieve viral suppression. Therefore, a multicenter randomized double blind control trial of phase III called LAKE (Negredo *et al.*, 2010), was performed. A number of 116 patients were randomly assigned in a ratio of 1:1 to two drugs:

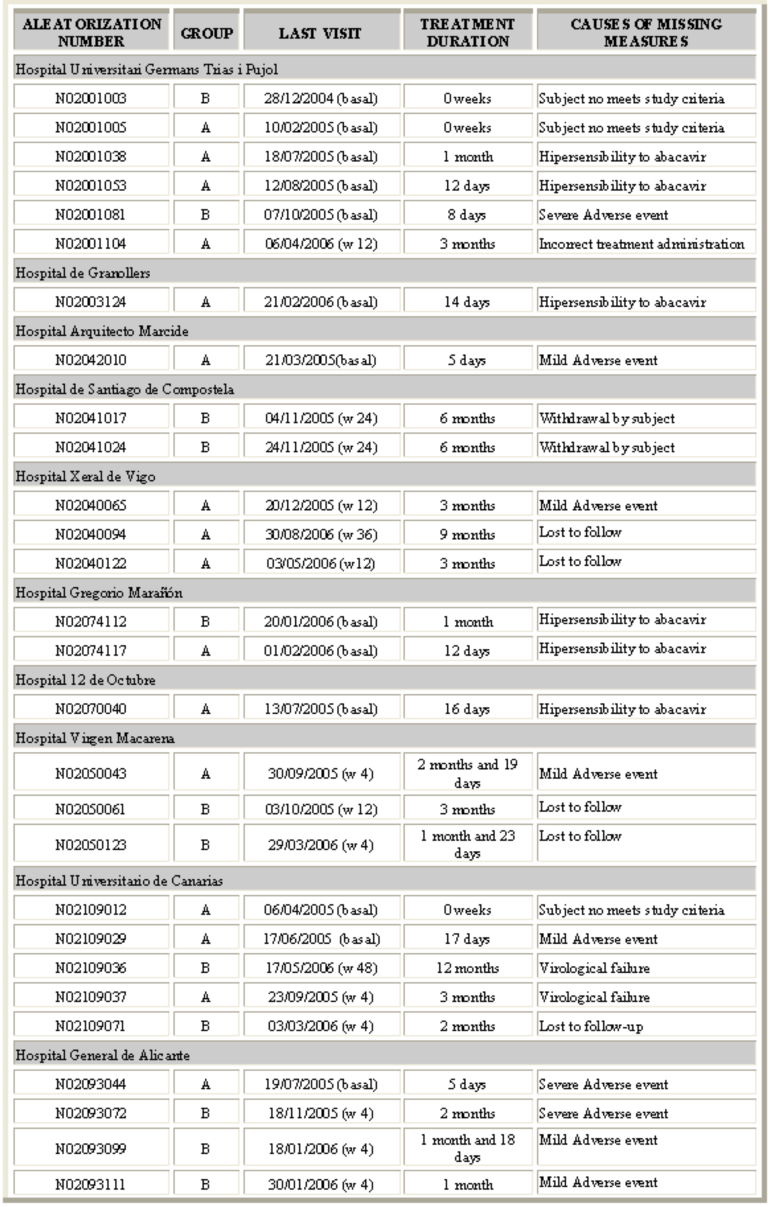
- efavirenz + abacavir (600 mg) /lamivudine (300 mg) once daily =EFV + KIVEXA
- Lopinavir (400 mg, 3 capsules) +ritonavir(100 mg twice a day)= KALETRA + KIVEXA

The clinical trial was planned to have 5 visits: basal, at 12 weeks, 24 weeks, 36 weeks and 48 weeks (corresponding to the end of the drug administration).

Patients were recluted from 19 study centers in Spain between March 2005 to March 2006 from people aged 18 years or above, with HIV-1. The primary endpoint was to evaluate the percentage of participants with HIV-1 RNA < 50 copies/mL at week 48. The second endpoint was to evaluate the virological failure and changes in CD4 at time 48 weeks. In the course of this thesis we will focus on the primary endpoint.

3.2 Missing Causes

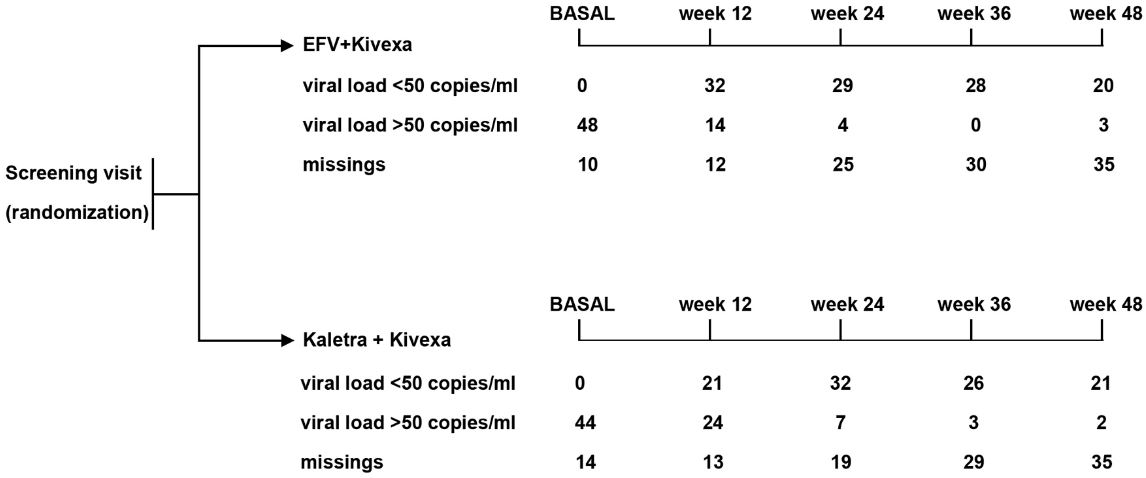
A number of 58 patients were assigned to the Efavirenz group and the other 58 patients to the Lopivavir/r group. Reasons for discontinuation are classified as: virological failure, adverse events (mild, moderate and severe based on the intensity), hypersensibility reaction, death or any other causes (voluntary discontinuation, simplification, etc) and are sumarized in figure 3.1.



ALEATORIZATION NUMBER	GROUP	LAST VISIT	TREATMENT DURATION	CAUSES OF MISSING MEASURES
Hospital Universitari Germans Trias i Pujol				
N02001003	B	28/12/2004 (basal)	0 weeks	Subject no meets study criteria
N02001005	A	10/02/2005 (basal)	0 weeks	Subject no meets study criteria
N02001038	A	18/07/2005 (basal)	1 month	Hipersensibility to abacavir
N02001053	A	12/08/2005 (basal)	12 days	Hipersensibility to abacavir
N02001081	B	07/10/2005 (basal)	8 days	Severe Adverse event
N02001104	A	06/04/2006 (w 12)	3 months	Incorrect treatment administration
Hospital de Granollers				
N02003124	A	21/02/2006 (basal)	14 days	Hipersensibility to abacavir
Hospital Arquitecto Marcide				
N02042010	A	21/03/2005(basal)	5 days	Mild Adverse event
Hospital de Santiago de Compostela				
N02041017	B	04/11/2005 (w 24)	6 months	Withdrawal by subject
N02041024	B	24/11/2005 (w 24)	6 months	Withdrawal by subject
Hospital Xeral de Vigo				
N02040065	A	20/12/2005 (w 12)	3 months	Mild Adverse event
N02040094	A	30/08/2006 (w 36)	9 months	Lost to follow
N02040122	A	03/05/2006 (w12)	3 months	Lost to follow
Hospital Gregorio Marañón				
N02074112	B	20/01/2006 (basal)	1 month	Hipersensibility to abacavir
N02074117	A	01/02/2006 (basal)	12 days	Hipersensibility to abacavir
Hospital 12 de Octubre				
N02070040	A	13/07/2005 (basal)	16 days	Hipersensibility to abacavir
Hospital Virgen Macarena				
N02050043	A	30/09/2005 (w 4)	2 months and 19 days	Mild Adverse event
N02050061	B	03/10/2005 (w 12)	3 months	Lost to follow
N02050123	B	29/03/2006 (w 4)	1 month and 23 days	Lost to follow
Hospital Universitario de Canarias				
N02109012	A	06/04/2005 (basal)	0 weeks	Subject no meets study criteria
N02109029	A	17/06/2005 (basal)	17 days	Mild Adverse event
N02109036	B	17/05/2006 (w 48)	12 months	Virological failure
N02109037	A	23/09/2005 (w 4)	3 months	Virological failure
N02109071	B	03/03/2006 (w 4)	2 months	Lost to follow-up
Hospital General de Alicante				
N02093044	A	19/07/2005 (basal)	5 days	Severe Adverse event
N02093072	B	18/11/2005 (w 4)	2 months	Severe Adverse event
N02093099	B	18/01/2006 (w 4)	1 month and 18 days	Mild Adverse event
N02093111	B	30/01/2006 (w 4)	1 month	Mild Adverse event

Figure 3.1: *Missing causes*

As shown in chapter 2, not all subjects included in the study, will complete the 48 weeks follow-up. Focusing on the primary endpoint the number of patients with viral load < 50 copies/mL and $>= 50$ copies/mL for the two treatments during time are summarized in the following follow-up sequence:



Screening visit
(randomization)

EFV+Kivexa	BASAL	week 12	week 24	week 36	week 48
viral load <50 copies/ml	0	32	29	28	20
viral load >50 copies/ml	48	14	4	0	3
missings	10	12	25	30	35

Kaletra + Kivexa	BASAL	week 12	week 24	week 36	week 48
viral load <50 copies/ml	0	21	32	26	21
viral load >50 copies/ml	44	24	7	3	2
missings	14	13	19	29	35

Figure 3.2: *Lake Study design*

The number of missing values for viral load increase over time in the two treatment groups. The distribution of missings over time for the two drugs is quite similar at 12, 24 and 48 weeks. In the other weeks we may observe some difference. The distribution of response, defined as measuring a viral load < 50 copies/mL, and of no reponse, defined as measuring a viral load $>= 50$ copies/mL, are also heterogenous between groups.

3.3 Descriptive analysis

For each patient, sociological variables were recorded at the beginning of the study. According to the duration of the trial, the viral load and the CD4 were evaluated at baseline, 12, 24, 36 and 48 weeks:

TABLE 3.1: Variables Description

Name	Description
Sex	Patient sex =1 male, =2 female
Age	Age in years
Group	EFV + Kivexa, Kaletra + Kivexa
CD4A0	T-CD4 lymphocytes counts at time 0
CD4A12	T-CD4 lymphocytes counts at week 12
CD4A24	T-CD4 lymphocytes counts at week 24
CD4A36	T-CD4 lymphocytes counts at week 36
CD4A48	T-CD4 lymphocytes counts at week 48
cv500	Percentage of load viral at time 0,$<= 50$,> 50
cv5012	Percentage of load viral at week 12,$<= 50$,> 50
cv5024	Percentage of load viral at week 24,$<= 50$,> 50
cv5036	Percentage of load viral at week 36,$<= 50$,> 50
cv5048	Percentage of load viral at week 48,$<= 50$,> 50
Infection time	Time from the HIV infection measured in months

In table 3.2, we presented some statistics for the continuous and the categorical variables considered in the study.

TABLE 3.2: Summary Statistics

Variable	Mean (sd)	Median (Inter. Range)	MIN-MAX	Freq.(%)
Sex(Man)				95 (86.36)
Age	38.04 (8.28)	37 (32.25-43)	20-59	
Group (EFV + Kivexa)	58 (50)			
CD4A0	192.6(123.32)	188(89-283)	3.3-569	
CD4A12	417.8(672.96)	333(209-459)	45.8-6604	
CD4A24	362.9(192.046)	331.5(193.6-499.8)	83.8-806	
CD4A36	392.2(202.879)	375(223-519)	13.8-900	
CD4A48	431.2(243.95)	398.5(268.2-509)	15.1-1169	
cv500$<= 50$				0 (0)
cv5012$<= 50$				53 (58.24)
cv5024$<= 50$				61 (84.72)
cv5036$<= 50$				54 (94.74)
cv5048$<= 50$				41 (89.13)
Infect.time(months)	27.82 (53.090)	6.067(2.083-23.47)	0-285.30	

Descriptive statistics by group are summarized in the table below:

TABLE 3.3: Summary Statistics by group of treatment

Variable	EFV + Kivexa (N=58)	Kaletra + Kivexa(N=58)
Age*	38.68(8.49)	37.4(8.082)
Infection time (in months)**	13.2(3.317-38.75)	4.32(1.725-17.02)
Sex(Man)***	48(85.71)	47(87.04)

* Values are Mean (standard deviation) for variables normally distributed.

** Values are Median (interquartile range) for variables no normally distributed.

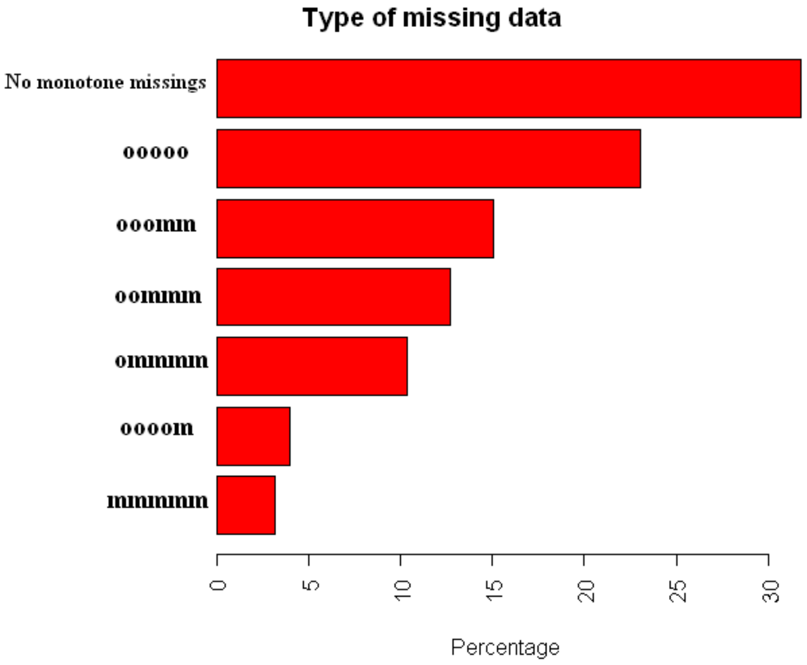
*** Values are number of cases (proporions) for categorical variables.

Time infection could be thought as a confounder variable that should be taken into account when we analyze the effect of the treatment. Patients following an EFV + Kivexa are patients with higher infection time and so with poor health respect to patients treated with Kaletra + Kivexa. Moreover in the presence of a no good administration of the treatment, EFV + Kivexa patients are more resistent. The variable infection time could be responsible for lack of effect for EFV + Kivexa compared to Kaletra + Kivexa group. In the following analysis we adjust for this variable.

The table above shows homogeneity in the mean age and the same proportion of men included for the two treatments. It seems to be a difference in the infection time. In particular patients in the EFV + Kivexa tend to have an higher infection time respect to patients in the Kaletra + Kivexa group.

3.4 Study of the missing patterns

Results in the study were presented every 12 weeks starting from baseline up to 48 weeks. We firstly summarized the different types of missings data for the variable viral load during time, respectively at baseline, 12, 24, 36 and 48 months.



Where o is the observed outcome and m the missing outcome

More than 45% of missings have a monotone structure, the 25% of values are observed during the follow-up time and almost the 30% of missings have a non monotone missing structure.

We also summarize the missing patterns for type of administrated drug:

TABLE 3.4: Missing pattern for group of treatment

Missing pattern	EFV + Kivexa	Kaletra + Kivexa
ooooo	16 (27.59)	13(22.41)
oooom	3 (5.17)	2 (3.45)
ooomm	8 (13.79)	11 (18.97)
mmmmm	2 (3.45)	2 (3.45)
no monotone missings	11(18.97)	19 (32.76)
Total	58	58

Values are frequency (%)

As we may see from table 3.4, almost 28% of patients in EFV + Kivexa and 22% in Kaletra + Kivexa have a complete pattern of response during time. The second line shows a missing data structure in which values are missing at 48 weeks, in the third line values are missings at 36 weeks. Finally, we may observe the same number of cases (2) and percentages (3.45) for a missing data pattern corresponding to missing observations during the all follow up-time.

3.5 Different missing treatments

In chapter 2, we have described different simple methods used to handle missing data in clinical trials. In this section we will apply some of them to the Lake study according to two of the prefered principles to the analysis of clinical trials: intention to treat analysis and protocol analysis.

- Intention to treat principle: "everyone randomized should be included into the analysis". Based on this principle we consider three approaches: LOCF, Missings=Failure and Missings=Success. LOCF was applied to the data and the percentage of viral load < 50 copies/mL and is described in the following table:

  TABLE 3.5: LOCF

Group	Baseline	12 Weeks	24 Weeks	36 Weeks	48 Weeks
EFV+Kivexa	0	48	68.97	75.86	70.69
Kaletra+Kivexa	0	36.21	63.79	68.97	68.97

  Values are expressed as % of patients with viral load < 50 copies/mL

  The table 3.5 shows as EFV+Kivexa treatment is more effective than Kaletra+Kivexa during the all follow-up time.

  In the case of a poor outcome assumption, we set missing values equal to failures.

  TABLE 3.6: Assuming poor outcome(Missing = Failure)

Group	Baseline	12 Weeks	24 Weeks	36 Weeks	48 Weeks
EFV+Kivexa	0	55.17	50	48.28	34.48
Kaletra+Kivexa	0	36.21	55.17	54.17	36.21

  Values are expressed as % of patients with viral load < 50 copies/mL

With the assumption of Missings=failure, we are assuming that a patient withdraws because does not believe in the advantage of the treatment as he has not responded up to the time of dropout. The table 3.6 shows as Kaletra+Kivexa treatment is more effective than EFV+Kivexa starting from 24 weeks.

In the case of a good outcome assumption, we set missing values equal to success. This hypothesis is less plausible than the previous one because it is difficult to think that severe patients could control the HIV virus before the end of the treatment and no require any other drug administration.

TABLE 3.7: Assuming good outcome (Missing = Success)

Group	Baseline	12 Weeks	24 Weeks	36 Weeks	48 Weeks
EFV+Kivexa	17.24	75.86	93.10	100	94.83
Kaletra+Kivexa	24.14	58.62	87.93	93.75	96.55

Values are expressed as % of patients with viral load < 50 copies/mL

The table 3.7 shows as the Kaletra+Kivexa treatment is more effective than EFV+Kivexa at baseline (24.14% versus 17.24%) and at the end of the treatment administration (96.55 % versus 94.83 %).

- On treatment analysis (or protocol analysis): to avoid diluition of treatment effect, we also perform an analysis by treatment actually received or complete case analysis:

TABLE 3.8: On treatment analysis

Group	Baseline	12 Weeks	24 Weeks	36 Weeks	48 Weeks
EFV+Kivexa	0	69.57	87.88	100	86.96
Kaletra+Kivexa	0	46.67	82.051	89.66	91.3

Values are expressed as % of patients with viral load < 50

As we may observe from the table 3.8, in case of on treatment analysis, the percentage of response is higher for EFV+Kivexa respect to Kaletra+Kivexa treatment during time up to 36 weeks. At 48 weeks, the Kaletra+Kivexa treatment will be more effective than the other one. We may conclude that the EFV+Kivexa experiments fast effectiveness, but at the time of the evaluation of the primary endpoint Kaletra+Kivexa will be prefered as permit to reach 91.3 % of patients with viral load < 50 copies/mL.

The results presented in tables 3.5-3.8 are quite ambiguous. For some analysis EFV+Kivexa is more effective than Kaletra+Kivexa. For others we may state the opposite conclusion. Conclusion about treatment effectiveness is related to missingness treatment. In section 3.6.2, we analyze the same models considering some adjusting variables.

3.6 Exploratory data analysis

As observed in section 2.5, to explore if there is a structure of missings observations during time, we built a series of χ^2 tests to check the correlation between missingness over different time points. We found a statistically significant association between the proportion of missing data for viral load at week 12 and week 24 (p-value = 1.907e-04), at week 24 and week 36 (p-value = 6.485e-08), at week 36 and week 48 (p-value $< 2.2e-16$). We did not find any statistically association between missings at baseline and 24 months.

Following section 2.5 the missings data pattern could be evaluated using the logistic regression for independent observations and the GEE models in case we assume a correlation structure:

3.6.1 Logistic regression

We specify a categorical response variable R with two levels (1 =missing and 0 =observed viral load during follow-up), a logit link function and a series of covariates known to be associated with the presence of missing like time of infection, time, sex, age of the patient and treatment for clarifying this dependency. We set male and EFV+KIVEXA as reference categories

$$log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 time_i + \beta_2 InfectionTime_i + \beta_3 Female_i + \beta_4 age_i \\ + \beta KALETRA + KIVEXAi + \varepsilon_i$$

with the error term $\varepsilon_i \sim N(0, \sigma_\tau)$

Where $p_{ij} = E(R_{ij} \mid X_i) = P(R_{ij} = 1 \mid X_i)$ and R_{ij} is the presence of the i-th individual measured at time j, with distribution $R_{ij} \sim Bernoulli(p_{ij})$.

To allow account for correlation between observations, we fit a GEE model with the presence of missing for the different time points as response with unstructured correlation matrix structure. We started evaluating the dependency of missings from time and treatment:

TABLE 3.9: GEE with treatment and time

Variable	Estimate	SANDWINCK SE	P-value
Intercept	-1.991	0.249	1.443e-15
TREAT.(Kaletra + Kivexa)	-0.0324	0.180	0.857
Time	0.491	0.0679	4.645e-13

From table 3.9, we may observe as the missing structure depends on the visit times. In this case, we could conclude that missing data are not completely at random, because the missingness does depend on variables in the database (time).

We also consider a model with sex and treatment as covariates adjusted to the number of values not missing for the sex variable:

TABLE 3.10: GEE with sex, treatment and time

Variable	Estimate	SANDWINCK SE	P-value
Intercept	-1.987	0.254	4.885e-15
TREAT.(Kaletra + Kivexa)	-0.0733	0.186	0.693
Time	0.485	0.069	2.201e-12
Sex(Female)	-0.00869	0.285	0.9756

Sex and treatment are not associated with the presence of missing for the variable viral load.

As we did not observe any dependency respect to sex, we decide to exclude this variable from the analysis. Finally we fit a model with infection time, treatment effect and time.

TABLE 3.11: GEE with treatment, time and infection time

Variable	Estimate	SANDWINCK SE	P-value
Intercept	-1.788	0.277	1.141e-10
TREAT.(Kaletra + Kivexa)	-0.00428	0.197	0.983
Time	0.415	0.072	8.521e-09
Infection Time	-0.0026	0.00183	0.156

Introducing the time of infection we may observe a statistically significant dependence of missing respect to visit times. A model adjusted for age was also fitted but we did not find any association with the missing response. We also evaluated the association between CD4 during time and the presence of missing for viral load as described in the literature. We observed missings data in CD4 corresponding to missing data in the viral load during time. This is an evidence of the lack of blood analysis and so we may conclude than the correlation structure is not fiable. So we decided to not consider the CD4 for our analysis.

3.6.2 Sensitivity analysis

The sensitivity analysis consists at comparing the model coefficients obtained under various models valid under different missing structures: MCAR or MAR. We first started with a logistic regression model that is valid under a MCAR structure and under MAR. The coefficient estimations is based on the likelihood function. In particular logistic regression is equivalent to a GEE under the hypothesis of independency between observations. As we have shown in paragraph 3.6.1, the generalized estimating equations procedure (GEE) requires that missing data depend only on covariates or that they be missing completely at random (MCAR) otherwise GEE regression parameter estimates are biased. In case of MAR assumption we should introduce a weighted generalized estimating equations (WGEE) to have unbiased estimation of parameters or we should specify a good model in which the covariates explain the missing observations. With the logistic regression model, we use the observed data ignoring all missing measurements for the response (viral load) and the other covariates (time, infection time, age, sex and treatment):

logit($\mathrm{Y}_{i=1} \mid time_i, treatment_i, age_i, infectiontime_i, sex_i) = \beta_0 + \beta_1 time_i + \beta_2 treatment_i + \beta_3 age_i + \beta_4 infectiontime_i + \beta_5 sex_i + \varepsilon_i$

with the error term $\varepsilon_i \sim N(0, \sigma_\tau)$

The model above allows for:

- differences between group of treatment
- differences between sex
- differences in age
- linear changes in the log odds of infection over time with slope β_4

We are interested at evaluating the viral response to the treatment during time, so we fix cv$>= 50$ copies/mL as the reference category. Also we have considered EFV+Kivexa respect to Kaletra+ Kivexa and female respect to male.

TABLE 3.12: Logistic regression model

Variable	Estimate	Standard error	P-value
Intercept	-2.970	0.852	0.000487
TREAT.(EFV + Kivexa)	0.8216	0.336	0.0144
Time	1.700	0.189	$< 2e - 16$
Time Infection	-0.000451	0.00293	0.877
Female	-0.944	0.481	0.0494
Age	-0.0263	0.0199	0.187

We first see from Table 3.12 that treatment, sex and time are all statistically significant predictors of viral load. In this model, increasing time is associated with an increasing response defined as viral load < 50 copies/mL (1.700), female sex is associated with a decreased response (-0.944), and EFV+Kivexa treatment is associated with an increased response respect to Kaletra+Kivexa.

The ordinary logistic regression model permits to analyze an association model with just fixed effects so a model with both fixed effects and random effect in the constant or in the slope, called mixed model is going to be used. With this model we may:

- evaluate the random effects into the intercept that allow for heterogeneity between individuals and represents the part of omitted subject specific covariates that causes some subjects to be more prone to the EFV + Kivexa than others.
- evaluate the random effect into time, allow the model to take account of different slopes between individuals.

Under a mixed model we are assuming a MAR structure.

We first started considering a mixed model with a random intercept:

logit(Y$_{i=1}$ $| \, time_i, treatment_i, age_i, infectiontime_i, sex_i) = (\beta_0 + b_{i0}) + \beta_1 time_i + \beta_2 treatment_i + \beta_3 age_i + \beta_4 infectiontime_i + \beta_5 sex_i + \varepsilon_i$

with the error term following a multivariate normal distribution with mean 0 and variance $\sigma_\tau \quad \varepsilon_i \sim N(0, \sigma_\tau),$ and random effects following a multivariate normal distribution with mean 0 and variance σ_0 , $b_{i0} \sim N(0, \sigma_0)$.

TABLE 3.13: Mixed logistic model with random effect in the constant

Variable	Estimate	Standard error	P-value
Intercept	-2.973	0.853	0.00049
TREAT.(EFV + Kivexa)	0.8219	0.336	0.0144
Time	1.701	0.189	$< 2e-16$
Time Infection	-0.000451	0.00293	0.878
Female	-0.945	0.481	0.0495
Age	-0.0263	0.0199	0.187

Results presented in table 3.13 are quite similar to results displayed in table 3.12. We may observe some difference in the treatment effect 0.8216 versus 0.8218 and for time 1.700 versus 1.701.

We can employ the likelihood ratio approach for testing the random effect on the constant. A random effect in the constant is not statistically significant (p= 0.497)

We also considered a mixed model with a random intercept and slope:

logit(Y$_{i=1}$ $\mid$ $time_i, treatment_i, age_i, infectiontime_i, sex_i) = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})time_i + \beta_2 treatment_i + \beta_3 age_i + \beta_4 infectiontime_i + \beta_5 sex_i + \varepsilon_i$

with the error term and random effects $\varepsilon_i \sim N(0, \sigma_\tau)$, $b_{i0} \sim N(0, \sigma_0)$ and $b_{i1} \sim N(0, \sigma_1)$

Where the random effects are represented by b_{i0} and b_{i1} .

TABLE 3.14: Mixed logistic model with random effect on constant and slope

Variable	Estimate	Standard error	P-value
Intercept	-6.499	1.417	4.63e-06
TREAT.(EFV + Kivexa)	1.187	0.477	0.0128
Time	3.764	0.480	4.66e-15
Time Infection	0.00179	0.00429	0.676
Female	-0.137	0.659	0.835
Age	-0.042	0.0298	0.158

Table 3.14 displays that treatment and time are statistically associated with response. In particular an increasing time is associated with an increasing response (3.764) and EFV+Kivexa treatment is associated with an increased response respect to Kaletra+Kivexa (1.187).

We employed the likelihood ratio approach for testing the random effects. Both random effects into intercept and slope are statistically significant (p< 0.00001).

With a mixed model with random effect in the constant and in the slope, the total variability is decomposed in:

- Individual variability: 12.278
- Time variability: 4.678
- Residuals variability

An alternative approach to the available case analiysis examined in tables (3.12-3.14) is the imputation analysis. In this case we include all the randomized subjects by imputing the missing values with various imputation strategies: LOCF, missings =Failure, missings=Success, logit model imputation, multiple imputation and hot deck multiple imputation. We first start with 4 ways of imputation:

1) Logit model imputation: with this approach we take into account the significant dependence of missing values of the main outcome with the observation data of the treatment variable and the others covariates. We built a logit regression with dichotomous response for modelling this relationship that allows us to impute values on missing observations.

TABLE 3.15: Mixed logistic model with missings simple Imputation

Variable	Estimate	Standard error	P-value
Intercept	-3.068	0.8307	0.000221
TREAT.(EFV + Kivexa)	0.939	0.327	0.004053
Time	0.889	0.0898	$< 2.e - 16$
Time Infection	-0.0007564	0.00313	0.809
Female	0.136	0.460	0.767
Age	0.00973	0.0202	0.6306

2) LOCF

TABLE 3.16: Mixed logistic model with LOCF Imputation

Variable	Estimate	Standard error	P-value
Intercept	-3.133	1.614	0.0522
TREAT.(EFV + Kivexa)	0.859	0.657	0.1908
Time	1.585	0.151	$< 2.e - 16$
Time Infection	0.00345	0.0062	0.556
Female	-0.843	0.918	0.578
Age	-0.0468	0.040	0.358

As we may observe from table 3.16 according with the results presented in table 3.5, the effect of Kaletra+Kivexa and EFV+Kivexa are quite similar during time (p=0.1908, no statistically significant).

3) Missings=success

TABLE 3.17: Mixed logistic model with Missings=success

Variable	Estimate	Standard error	P-value
Intercept	-2.883	0.888	0.00116
TREAT.(EFV + Kivexa)	0.617	0.341	0.0703
Time	2.294	0.229	$< 2.e - 16$
Time Infection	-0.000566	0.00331	0.864
Female	0.0477	0.461	0.918
Age	-0.0309	0.0215	0.150

We first see from Table 3.17 that time is still significant but we may not assume differences in treatment effectiveness.

4) Missings=failure

TABLE 3.18: Mixed logistic model with Missings=Failure

Variable	Estimate	Standard error	P-value
Intercept	-1.321	0.889	0.137
TREAT.(EFV + Kivexa)	0.458	0.365	0.210
Time	0.479	0.104	4.32e-06
Time Infection	0.00228	0.00336	0.498
Female	-0.233	0.525	0.657
Age	-0.0262	0.0228	0.250

In case of a worst scenario, EFV + Kivexa drug is no longer prefered to Kalexa + Kivexa treatment.

1) is valid under a MCAR structure and could permit to create a complete dataset that will be analyzed with a regression model. In the first step we estimate the mean vector and the covariance matrix for the complete cases assuming a logistic regression model. In the second step the conditional mean from the regression of the missing components on the observed measurements is calculated and substituted for the missing values. Using this method permit to have some advantages and disadvantages:

Advantages

- The point estimation maintains

- Reduce the covariances distortion
- The size of the sample is preserved

Disadvantage

- The model should be correctly specified and this highly complicated to reach

2),3) and 4) permit to create a complete dataset that will be analyzed with a mixed model with both random effects in the intercept and in the constant who permits to have unbiased estimates under a MAR structure.

Another approach used to handle missing data, was showen in section 2.7 and is known with the name of multiple imputation. With the multiple imputation the missing data for each variable are predicted using existing values from other variables.

TABLE 3.19: Multiple Imputation for time of infection

Observation	imput.1	imput.2	imput.3	imput.4	imput.5
1	10.3	3.267	6.033	4.90	0.667
9	0.967	41	23.10	0.70	52.233
23	10.667	14.567	43.20	18.30	52.267
26	10.30	36.267	213.967	3.267	30.80
35	41	0	1.567	13.20	1.367
44	181.60	181.60	1.233	285.30	36.50

In table 3.19 we may observe some of the imputed values for the variable time of infection in 5 imputed datasets.

The library mice (multiple imputation by chained equation) in the R environment permits to impute any incomplete data specified into the model. The default methods used in mice are predictive mean matching for numeric data, logistic regression for two categories and polytomic logistic regression for categorical variables.

Once the imputed dataset have been created (we created 5 datasets, we have seen in section 2.7 that with 5 imputations we reach a nearly fully efficiency), we fit 5 logistic regression models for the viral load. Once the analysis have been completed for each of the 5 datasets, we combined the parameter estimates using Rubin's rules for obtaining an overall set of estimates. We first have built a table for evaluating the percentage of patients with viral load < 50 copies/mL in the two treatments during the 48 weeks.

TABLE 3.20: Multiple Imputation analysis

Group	Baseline	12 Weeks	24 Weeks	36 Weeks	48 Weeks
EFV + Kivexa	0	67.24	79.31	96.55	86.21
Kaletra + Kivexa	0	43.10	82.76	93.10	86.21

Values are expressed as % of patients with viral load < 50 copies/mL

With multiple imputation the two treatments are equal effective at 48 weeks. We may observe some differences from baseline up to 36 weeks. The parameter estimates for the pool multiple imputation are reported in the following table:

TABLE 3.21: Pool Multiple Imputation

Variable	Estimate	Standard error	P-value
Intercept	-0.00524	0.0760	0.945
TREAT.(EFV + Kivexa)	0.0718	0.0289	0.0136
Time	0.225	0.0101	$< 2e - 16$
Time Infection	-0.00053	0.000272	0.0516
Female	-0.130	0.0420	0.00206
Age	-0.000152	0.00178	0.932

By using multiple imputation analysis sex, time and treatment are statistically significant.

3.7 Hot-Deck analysis

The hot-deck method was already described in section 2.8. In this section, I implement the hot-deck imputation method in the R environment. We may summarize the procedure in 4 steps:

Step 1) We start the process creating a R_i variable that assumes value 1 if the variable viral load is missing at baseline and 0 otherwise. For our dataset, at time 0 the number of patients with viral load < 50 is equal to zero. In this particular case, we simplify the analysis assigning to the missing observations the value=viral load $>= 50$.

Step 2) At week 12, we repeat the process. We have created a R_i variable that will assume value equal to 1 if the viral load is missing at week 12 and 0 otherwise. We fit a logistic regression model in $R_{i=1(week=12)}$ and we introduce the same covariates considered in the previous models. We also expect to have a dependency respect to viral load at baseline. In our particular case the viral load has just one category (viral load $>= 50$) so will not appear into the model:

logit(R$_{i=1})_{week=12} = \beta_0 + \beta_1 time_i + \beta_2 treatment_i + \beta_3 age_i + \beta_4 infectiontime_i + \beta_5 sex_i + \varepsilon_i$

From the logistic regression, we estimate the probability of response (viral load < 50) and also the propensity score as described in section 2.8. The propensity score will be divided into quartiles. In each quartile there will be a number of missings outcomes (nmiss) and a certain number of observed outcomes (nobs) We randomly sample the missing outcomes from the observed outcomes that were previously randomized with replacement (to be adjusted to the length of the missings outcomes). We do the same for each quartile.

Step 3) We repeat the same process at week 24. In this case we will fit the following model:

logit(R$_{i=1})_{week=24} = \beta_0 + \beta_1 time_i + \beta_2 treatment_i + \beta_3 age_i + \beta_4 infectiontime_i + \beta_5 sex_i + \beta_6 * viraload_{iweek=0} + \beta_7 * viraload_{iweek=12} + \varepsilon_i$

where $viraload_{iweek=12}$ was imputed in step 1 and step 2.

Step 4) We repeat this analysis for each time up to 48 weeks.

Let consider 6 missing values for the first quartile at week 12. We assigned randomly with replacement the 6 values from the 8 observed outcomes:

With the same process, we have obtained the following table at week 12:

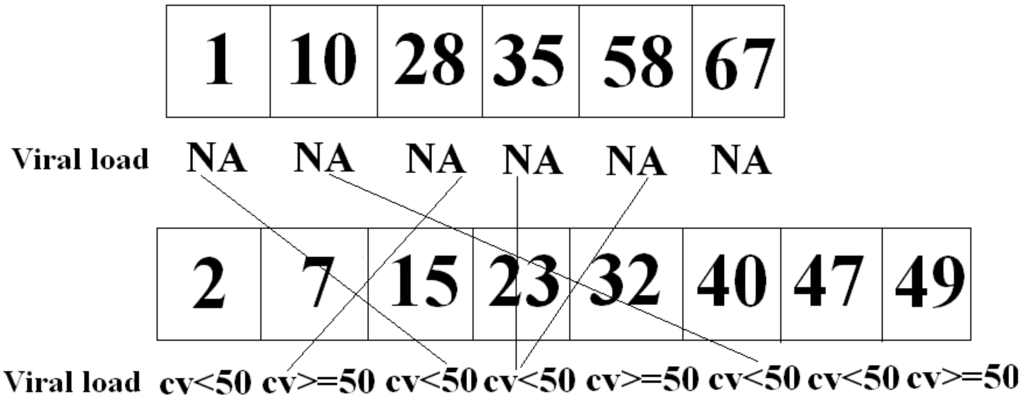


Figure 3.4: *HOT-DECK*

TABLE 3.22: Hot deck imputation at week 12

Category	at start	1 Quartile	2 Quartile	3 Quartile	4 Quartile
cv<50	53	58	59	63	64
cv$>=50$	38	39	40	942	47
NA	25	19	17	11	5

Running the step 1) 2) and 3) for each time, we will be able to explore the effect of the treatment using a logistic regression model for the parameter estimations.

We first have built a table for evaluating the percentage of patients with viral load < 50 in the two treatments during the 48 weeks.

TABLE 3.23: Hot-Deck analysis

Group	Baseline	12 Weeks	24 Weeks	36 Weeks	48 Weeks
EFV + Kivexa	0	70.69	93.10	100	94.83
Kaletra + Kivexa	0	47.17	85.42	92.10	93.75

Values are expressed as % of patients with viral load < 50 copies/mL

EFV + Kivexa seems to be more effective during the all follow-up time in case of hot deck multiple imputation.

The parameter estimates, standard error and P-values are summarize in table

3.24:

TABLE 3.24: Hot-Deck Multiple imputation

Variable	Estimate	Standard error	P-value
Intercept	-10.1905	1.422	7.65e-13
TREAT.(EFV + Kivexa)	1.042	0.428	0.0149
Time	4.005	0.4505	$< 2e - 16$
Time Infection	0.00178	0.0039	0.647
Female	-0.221	0.605	0.715
Age	0.0451	0.0253	0.0750

By using Hot-deck multiple imputation analysis we found time and treatment statistically associated with the viral response.

3.8 Conclusions

We have reviewed different techniques to handle missing data in clinical trials. Some of these methods are valid under a MCAR assumption other under a MAR structure. The following table summarizes the odds ratio and the confidence intervals obtained using the different missing approaches:

TABLE 3.25: OR AND CI 95%

Model	OR	Lower CI (95%)	Upper CI (95%)
GLM	2.274	1.178	4.390
GLMM	3.276	1.287	8.339
SIMPLE IMP.	2.559	1.846	3.550
LOCF	2.362	0.652	8.559
MISSING=SUCCESS	1.853	0.950	3.616
MISSING=FAILURE	1.580	0.773	3.2309
MULTIPLE IMP.	1.088	1.348	4.859
HOT-DECK IMP.	1.225	2.836	6.566

The table above shows as EFV + Kivexa treatment seems to be more effective than EFV + Kivexa in the major part of the approaches. In case of GLM and GLMM, we have greater estimates for treatment effect and wider confidence intervals because we are excluding some cases (complete case analysis). With GLMM, we have obtained a greater Odds ratio, so we are overestimating the EFV+Kivexa treatment. In case of LOCF, missing=failure and missing=success, the two treatments appear to have the same effectiveness. The conditional mean imputation has the advantage to maintain the size of the sample and the estimation mean parameters but it should be correctly specified and this highly complicated to reach. Under the multiple imputation we have obtained lower estimates and also shorter confidence intervals respect to the other methods. With multiple imputation we are not just imputing the missings response but also the missings in the covariates. Finally by the hot-deck multiple imputation, we have obtained the same results in terms of direction of the effect but wider confidence intervals compared to the multiple imputation approach. We see that the coefficients obtained under a MAR structure and a MCAR structure are quite different, that it might be inferred that the missingness does affect the response.

The validity of the treatment methods depends of some assumptions about the missing structure: the simple imputation methods (LOCF, missings=failure and missing=success) are valid under a MCAR structure while under a MAR structure are considered invalid. The direct likelihood method without using imputation strategies (mixed logistic regression model without considering any missing imputation approach) and the multiple imputation approaches are both valid under a MAR assumption. Finally, selection models and the sensitivity analysis are valid under a MNAR structure of missing data. In case of sensitivity analysis we are considering different methods to handle missing data. Based on it and comparing the estimates we may conclude that the effectiveness of the EFV + Kivexa treatment is higher but not for all methods and we may build two intervals: a set of parameter estimates (region of ignorance) and a set of interval estimates (region of uncertainty). The region of interval is (1.088-3.276) and the region of uncertainty is (0.652-8.559).

Chapter 4

Concluding remarks

This master thesis gave me the opportunity to improve my skills in longitudinal data analysis and learn about the different methods used to handle missing data in clinical trials. The overall result of my work is a R function, which I have created, to implement the Hot-deck multiple imputation up to date only available in SAS environment. Moreover this work generate guidelines that should be followed by researchers working in clinical trials (in particular in the Lluita Fundacio') who have to analyse data sets with missing data:

- First of all researchers should evaluate if the percentage of the missing is negligible (less than 5%) which lead to unbiased estimates. In case of a percentage of missing higher than 5 %, researchers should investigate the missing data structure to confirm if the missing data are MCAR, MAR or MNAR fitting a GEE model with the presence of missing as response. If the model just include the constant, they may conclude that missing have a MCAR structure. If the final model includes some covariates, they should assume a MAR structure. In the case in which the model also include the observed outcomes a MNAR structure could be assumed. Unfortunally we generally cannot be sure whether data are actually missing at random, or the missingness depends on unobserved predictors or the missing data themselves.

- Make assumptions about the missingness: instead of exploring the missing structure, researchers could analyze the different missing data techniques to handle missing data, making assumptions on the missing structure and evaluate if the different approaches are valid. Based on it, simple imputation methods as LOCF, the worst case scenario (missing=failure) and the best scenario missing=success are valid under a MCAR structure while under a MAR structure are considered invalid. The multiple imputation approach is valid under a MAR assumption and the sensitivity analysis under a MNAR structure.

My recommendation for researchers working in clinical trials in HIV is to avoid simple imputation methods invalid under a MAR structure. The biases in the estimates may lead to false conclusions. I suggest to consider methods that are valid under a MAR structure (multiple imputation methods) or a MNAR structure (sensitivity analysis).

We conclude that the effectiveness of a treatment in a clinical trial is affected by the nature of the missing data and the preference for a particular method instead of another may affect the conclusions of a trial.

Future steps of this work will be to simulate different scenarios of missing percentage to evaluate from which percentage results will not be affected, simulate different sample sizes and to adopt different rules to generate missing data. With the simulates studies we could evaluate the treatment effectiveness under different scenario and missing data treatments.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Nuria Perez Alvarez for the continuous support of my study and research and for the moral support and also I would like to thank Professor Guadalupe Gomez for their valuable assistance in the preparation and completion of this study.

I thank the MEIO students in UPC in particular Moisés Gómez Mateu, Marc Dell'Olmo and Nuria Planell Picola for the stimulating discussions and for all the fun we have had in the last three years.

Last but not the least, I would like to thank my family: my husband Fabrizio for his support and patience and my little child Gaia who is on the way.

Bibliography

[1] AIDS HIV Information From AVERT (www.avert.org).

[2] Carpenter JR, Kenward MG, White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res. 2007. 16(3):259-75.

[3] ClinicalTrials.gov

[4] Cozzi Lepri A, Smith GD, Mocroft A, Sabin CA, Morris RW, Phillips AN. A practical approach to adjusting for attrition bias in HIV clinical trials with serial marker responses. AIDS. 1998. 12(10):1155-61.

[5] Dybul M, Fauci AS, Bartlett JG, Kaplan JE, Pau AK. Panel on Clinical Practices for Treatment of HIV. Guidelines for using antiretroviral agents among HIV-infected adults and adolescents. Ann Intern Med. 2002 Sep 3. 137(5 Pt 2):381-433.

[6] EACS Guidelines for the Clinical management and treatment of HIV-infected adults in Europe.

[7] Echeverría P, Negredo E, Carosi G, Gálvez J, Gómez JL, Ocampo A, Portilla J, Prieto A, López JC, Rubio R, Mariño A, Pedrol E, Viladés C, del Arco A, Moreno A, Bravo I, López-Blazquez R, Pérez-Alvarez N, Clotet B. Similar antiviral efficacy and tolerability between efavirenz and lopinavir/ritonavir, administered with abacavir/lamivudine (Kivexa), in antiretroviral-naïve patients: a 48-week, multicentre, randomized study (Lake Study). Antiviral Res. 2010 Feb. 85(2):403-8.

[8] European Medicine Agency (EMEA), Guidelines on missing data in confirmatory clinical trials.

[9] Global Reports on the Global AIDS Epidemic 2010 (www.unaids.org/globalreport).

[10] Huson LW, Chung J, Salgo M. Missing data imputation in two phase III trials treating HIV1 infection. J Biopharm Stat. 2007. 17(1):159-72.

[11] Kelleher T, Thiry A, Wilber R and Cross A. Missing data methods in HIV clinical trials: Regulatory guidance and alternative approaches. Drug Information Journal. 2001. 35:1363-1371.

[12] Le Corfec E, Chevret S, Costagliola D. Visit-driven endpoints in randomized HIV/AIDS clinical trials: impact of missing data on treatment difference measured on summary statistics. Stat Med. 1999. 18(14):1803-17.

[13] Liang KY and Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986. 73 (1):13-22.

[14] Little RJA, Models for nonresponse in sample surveys. Journal of the american statistical associacion. 1982. 77: 237-250.

[15] Little RJA. and Rubin DB. Statistical Analysis with Missing Data. 1987

[16] Little RJA and Rubin DB. Statistical Analysis with Missing Data, 2nd edition. 2002.

[17] Methodological Challenger in Biomedical HIV Prevention trials Board on Global Health (2008)

[18] Molenberghs G. and Kenward M. Missing data in clinical Studies. 2007

[19] William Myers R. Handling missing data in clinical trials: an overview. Drug information Journal. 2000. 34: 525-533.

[20] Rabould JM and Montaner JSG .Issues in the design of trials of therapies for subjects with human immunodeficiency virus infection that use plasma RNA level as an outcome. J Infect Dis. 1997 Mar. 175(3):576-82.

[21] Rubin DB. Inference and missing data. Biometrika. 1976. 63: 581-592.

[22] Rubin DB. Multiple imputation for non response in Surveys. 1987

[23] Schafer JL. Analysis of Incomplete Multivariate Data. (1997)

[24] Schafer JL. Multiple imputation: a primer. Statistical methods in medical research 1999; 8. 3-15.

[25] Shih W. Problems in dealing with missing data and informative censoring in clinical trials. Curr Control Trials Cardiovasc Med. 2002. 3(1):4.

[26] Smith KY, Patel P, Fine D, Bellos N, Sloan L, Lackey P, Kumar PN, Sutherland-Phillips DH, Vavro C, Yau L, Wannamaker P, Shaefer MS; HEAT Study Team. Randomized, double-blind, placebo-matched, multicenter trial of abacavir/lamivudine

or tenofovir/emtricitabine with lopinavir/ritonavir for initial HIV treatment. AIDS. Jul 2009.31;23(12):1547-56.

[27] Tang L, Juwon S, Thomas R. A comparison of imputation methods in a longitudinal randomized clinical trial. Statist. Med. 2005. 24: 2111-2128

[28] UNAIDS: http://www.unaids.org/en/

[29] Volberding PA, Deeks SG. Antiretroviral therapy and management of HIV infection. Lancet. 2010 Jul. 3;376(9734):49-62.

[30] Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials. 2004. 1(4):368-76.