Chapter 2

The Motivating HIV+PTB Cohort Example

2.1 Introduction

The aim of this chapter is to introduce the dataset that has motivated the contributions of this thesis. We present here the epidemiological problem and the initial exploratory steps approaching the dataset and the survival analysis when some part of the covariates is missing. Since Chapter 3 reproduces our paper "*CD4+ lymphocytes and tuberculin skin test as survival predictors in pulmonary tuberculosis HIV-infected patients*" the dataset is again included though in that chapter we perform several statistical analysis, from the complete case method to an approach via imputation and bootstrapping techniques to deal with the missing data problem.

2.2 The HIV+PTB dataset

In the city of Barcelona, the acquired immunodeficiency syndrome (AIDS) and the tuberculosis (TB) diseases are covered by the active epidemiological surveillance system of the Epidemiology Service of the Municipal Institute of Health (ESMIH). Specifically, the "Prevention and Control of Tuberculosis Programme" takes care, since 1986, of compiling data provided by doctors, as well as results of microbiological

analysis, hospital discharges, mortality data and linkage between AIDS and TB Registers. We are collaborating with the ESMIH since 1994 in several research projects and epidemiological data analysis.

In June 1995, a relevant data appeared: 3429 cases of AIDS had been diagnosed; from them 1097 (32%) were TB patients and three epidemiological goals arised:

1. To study of the progression of AIDS in TB patients
2. To find survival predictors in HIV+TB-infected patients
3. To elucidate if the response to the tuberculin skin test is a good prognosis for survival

The HIV+PTB cohort was integrated by 494 HIV-infected patients with pulmonary tuberculosis (PTB), with or without extrapulmonary infection, residents in Barcelona city, and diagnosed between January 1st 1992 and December 31st 1994. All of them started treatment against TB at the moment of the TB diagnosis. The survival time for each patient was established as the number of days between the date of TB diagnosis and death –or December 31st 1994 for those patients who were still alive on this date–. So, patients alive at December 31st 1994 and those lost of follow-up were considered to be right censored. More details about the selection criteria can be found in Chapter 3, Section 3.2.

For each patient, sociological variables as well as clinical variables listed in Table 2.1 had been recorded at the beginning of the study. Tables 2.2 and 2.4 present, as illustration, the value of the categorical and continuous variables for 10 arbitrary patients of the sample. The full dataset is in Appendix I. Descriptive statistics of all the variables are in Table 2.3, for the categorical covariates, and Table 2.5 for the continuous ones.

In addition to the mentioned variables, in order to analyze the missing data mechanism and study the conditional distributions, we create the following variables:

a) Response indicators for each variable with missing data

b) Standardization of the continuous variables Y and $CD4$

c) Categorization of AGE and $CD4$ based on the quartiles distribution and

Name	:	Description
SE	:	Patient sex: 0 = Female, 1 = Male
RE	:	Inner city resident indicator: 0 = Others, 1 = Inner city
BA	:	Bacteriology test: 0 = Negative, 1 = Positive,
		2 = Positive culture only
AI	:	AIDS diagnosis indicator: 0 = No, 1 = Yes
RA	:	Radiological pattern: 0 = Normal, 1 = Cavitary, 2 = Non cavitary
PR	:	Prison history indicator: 0 = No, 1 = Yes
TR	:	Treatment against tuberculosis history indicator: 0 = No, 1 = Yes
CO	:	Final conclusion: 0 = Pending , 1 = Recovered, 2 = Chronic,
		3 = Death, 4 = Emigration, 5 = Giving up
TG	:	HIV transmission group: 1 = IVDU[1], 2 = Homosexual man,
		3 = Hemophilic hemotransfused, 4 = IVDU and homosexual,
		6 = Heterosexual
AL	:	Alcohol addiction indicator: 0 = No, 1 = Yes
HL	:	Homelessness indicator: 0 = No, 1 = Yes
TB	:	Site TB: 0 = Pulmonary, 1 = Mix
AGE	:	Age (in years)
Y	:	Survival time (in days). Time to death or on study time
δ	:	Death/censoring indicator:
		0 = Alive/Censored, 1 = Death/Non-censored
CD4	:	T-CD4 lymphocytes counts (in percentage)
CD8	:	T-CD8 lymphocytes counts (in percentage)
IND	:	Index-ratio between CD4 and CD8 (direct measurement)
PPD	:	Tuberculin skin test result: 0 = Negative, 1 = Positive
MM	:	Reaction to the tuberculin skin test (in millimeters)

Table 2.1: *Names and description of the variables in the HIV+PTB dataset. Missing values are coded as NA*

[1] *IVDU: Intravenous drug user*

SE	RE	BA	AI	RA	PR	TR	CO[1]	TG[2]	AL	HL	TB
1	0	2	1	2	0	1	3	1	1	0	1
1	0	0	1	2	1	1	3	1	0	0	0
1	0	1	1	2	1	0	3	1	0	0	1
1	0	1	0	2	0	1	1	1	0	0	0
1	1	2	0	2	1	1	5	4	1	1	0
...	...	...	...	...	...	...	...	...	...	...	...
0	1	NA	1	2	0	0	0	NA	0	0	0
1	0	1	1	1	0	0	0	1	1	0	0
1	0	1	1	1	0	1	0	1	1	0	0
1	0	NA	1	2	1	0	0	1	0	0	0
1	0	0	1	2	0	0	0	2	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...

Table 2.2: *Categorical covariates in the HIV+PTB dataset for 10 arbitrary cases*

Value	SE	RE	BA	AI	RA	PR	TR	CO[1]	TG[2]	AL	HL	TB
0 :	16.4	68.8	16.6	35.2	6.1	70.4	79.4	30.4	22.2	62.8	95.5	76.3
1 :	83.6	30.6	39.1	64.8	19.4	29.6	20.6	36.0	72.5	37.2	4.5	23.7
2 :	–	–	34.0	–	72.3	–	–	33.4	–	–	–	–
NA :	–	0.6	10.3	–	2.2	–	–	0.2	5.3	–	–	–

Table 2.3: *Overall percentages for the values of the categorical covariates in the HIV+PTB dataset*

[1] *Summarized as 0 = Non recovered, 1 = Recovered, 2 = Other*

[2] *Summarized as 0 = Other, 1 = Exclusively IVDU*

d) Dichotomization of the $CD4$ based on the mean: 0 corresponds to $CD4 \leq 14$ and is the most immnunosuppressed level and 1 corresponds to $CD4 > 14$ and is the least immunosuppressed level.

AGE	Y	δ	CD4	CD8	IND	PPD	MM	
41	121	1	10	53	.19	0	NA	
36	17	1	NA	NA	NA	NA	NA	
31	365	1	15	55	.28	0	NA	
26	1030	0	16	69	.23	1	10	
40	526	0	NA	NA	NA	NA	NA	
...	...	...	...	...	...	...	...	
29	47	0	77	47	.17	0	NA	
35	121	0	23	60	2.11	1	NA	
30	9	0	20	53	.37	NA	NA	
33	31	0	NA	NA	NA	1	10	
46	16	0	13	47	.27	0	NA	
	...	...	...	...	...	...	...	...

Table 2.4: *Continuous variables in the HIV+PTB dataset for 10 arbitrary cases (δ and PPD binary variables are included for completeness)*

	AGE	Y	δ	CD4	CD8	IND	PPD	MM
Min:	17	1	–	1	7	0.01	–	3
1st Qu.:	28	156.8	–	5	53	0.07	–	10
Mean:	33.4	414.9	0.362	13.9	61.1	0.23	0.392	15.4
Median:	32	344.5	–	11	63	0.17	–	15
3rd Qu.:	36	630.5	–	18	72	0.30	–	20
Max:	66	1082	–	93	89	1.69	–	35
Std Dev.:	7.8	306.8	0.481	13.1	14.4	0.24	0.489	6.3
NA %:	–	–	–	38.9	39.5	38.9	50.4	86.8

Table 2.5: *Descriptive statistics for the continuous variables in the HIV+PTB dataset (δ and PPD binary variables are included for completeness)*

For each variable we fit a Cox proportional hazard model. Table 3.1, in Chapter 3, summarizes the most significant results. We observe that only the $CD4$ covariate (continuous and dichotomized), the result of the tuberculin test and age (continuous and stratified by quartiles) are significant. For all of them, in order to verify if the proportional hazard hypothesis holds, we test that the relative hazard is constant. In order to do that, we check that the relative hazard is independent on the time t in which it is estimated. We extend the initial model for the covariate, say X, by adding either the interaction $X \cdot t$ or $X \log t$ and the new modelizations for the hazard at time t for the value $X = 1$ respect to the hazard for the group $X = 0$, at the same time t, are expressed as

$$h(t; X, Xt) = \exp(\beta_1 X + \beta_2 X \cdot t) h_0(t)$$

and

$$h(t; X, X \log t) = \exp(\beta_1 X + \beta_2 X \cdot \log t) h_0(t).$$

It follows that the relative hazards are, respectively

$$\exp(\beta_1 X + \beta_2 X \cdot t) = \exp(\beta_1 X) \cdot \exp \beta_2 X \cdot t$$

and

$$\exp(\beta_1 X + \beta_2 X \cdot \log t) = \exp(\beta_1 X) \cdot t^{\beta_2 X}.$$

An hypothesis test on $\beta_2 = 0$ allows to decide the validity of the Cox model (Collett, 1994). For all these covariates the Cox model becomes correct.

In fact, looking at Table 3.1, $CD4$ and PPD are the best prognosis variables at univariate level. With respect to the $CD4$ covariate, in general, for each unit increment in the $CD4$ percentage there is a reduction of 5% in the hazard function. Note that the hazard in the least immunosuppressed group is a 35.4% of the hazard for the most immunosuppressed group. With respect the PPD covariate, we find that the positivity to the tuberculin skin test has a protective effect; specifically, the hazard for the positive tuberculin group is a 41% of the hazard for the negative tuberculin group. It is also possible to test that there is not significative differences between the negative tuberculin group and the group for whose the PPD covariate is missing (see Figure 3.3, part b), in Chapter 3).

2.3 The missing data problem

Looking at Tables 2.3 and 2.5 we can see that the percentage of missing data is, in general, lower than 10%. However, there is a large amount of missing data in both the variable $CD4$ (38.9%) and in the PPD variable (50.4%). Table 2.6 shows a table of contingency for the dichotomized $CD4$ and PPD values (including the code NA). Only 157 (31.8%) cases are complete in these two variables. We name this sample *observed subsample*.

	$CD4\%$			
	≤ 14	> 14	NA	Totals
PPD Negative	80	22	47	149
	(16.2/53.7/41.5)	(4.5/14.8/20.2)	(9.5/31.5/24.5)	(30.2)
PPD Positive	18	37	41	96
	(3.6/18.8/9.3)	(7.5/38.5/33.9)	(8.3/4.3/21.4)	(19.4)
PPD NA	95	50	104	249
	(19.2/3.8/49.2)	(10.1/20.1/45.9)	(21.1/41.8/54.2)	(50.4)
Totals	193	109	192	494
	(39.1)	(22.1)	(38.9)	(100)

Table 2.6: *Table of contingency for the values in the dichotomized CD4 % and PPD. Percentages in parentheses (overall/by rows/by columns)*

If we apply a χ^2 test to the previous table we obtain $\chi^2 = 37.16$ ($df = 4$) for the entire sample and $\chi^2 = 29.90$ ($df = 1$) for the observed subsample. So, in both cases there is a strong dependence between the immunosuppression level and the result of the tuberculin skin test (p-value $< 10^{-6}$). Looking at the percentage in the observed subsample we can infer that the dependence is positive, in the sense of that low values of $CD4$ are correlated with negative results of PPD and high values of $CD4$ are correlated with positive results of PPD. Boxplot in Figure 2.1 supports this positive dependence; note that the conditional distribution of the $CD4$ values

given that the value of PPD is missing is quite similar to the corresponding to the negative tuberculin group.

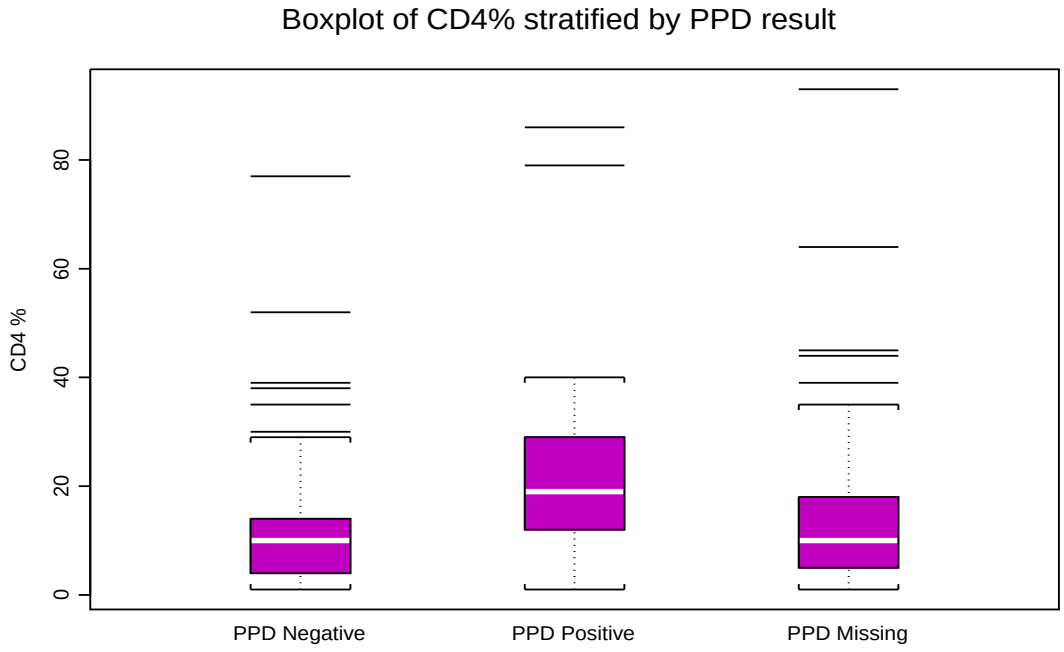

Figure 2.1: *Boxplot of the covariate CD4% stratified by the result of the tuberculin skin test (PPD)*

Other important issue, related with the source of missingness is the "anergic patient effect". In practice, it is well know, between doctors, that when a patient is much immunosuppressed, then he/she can become anergic, that is, the patient does not react to some intracutaneous tests, in particular to the tuberculin skin test. In our dataset, all the patients are PTB, so, potentially, all of them must have a positive PPD. Therefore, due to the fact that there are other variables to show the PTB disease (*e.g.,* the bacteriology test or the radiological pattern), in order to avoid some false negative some doctors skip the tuberculin skin test. In Chapter 4, we will consider the subsample integrated by the patients with observed Radiology and Bacteriology results (418 cases) in order to use these covariates as a surrogates for the missing values and thus to improve the efficiency of the estimators.

If we compute the Kaplan–Meier estimator for the entire sample and for the

observed subsample we obtain the curves shown in Figure 2.2. The picture is very interesting because we could think that the similarity between the survival estimates implies that the observed subsample is representative of the entire sample and, so, the missing data mechanism is MCAR or MAR. This impression is, however, false because when plotting the stratified survivals we know that the non-response pattern in the PPD covariate may be non-ignorable (Figure 3.3). The sensitivity analysis performed at the end of Chapter 6 confirms that the non-response to the PPD covariate can not be considered MAR.

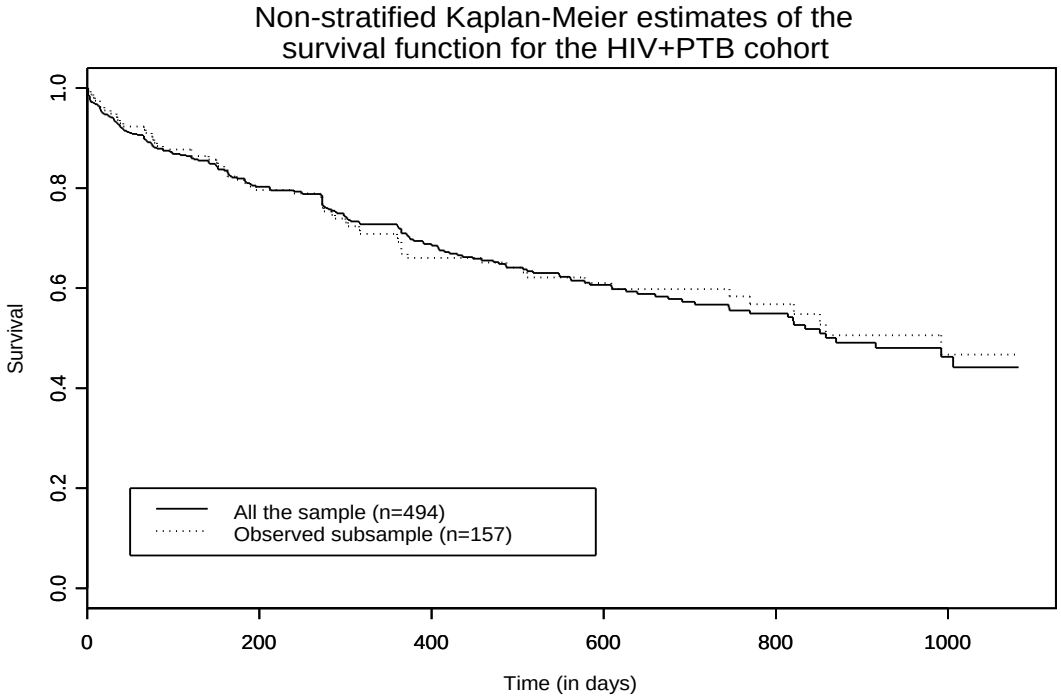

Figure 2.2: *Kaplan–Meier estimates of the survival function for all the sample (solid line) and for the observed subsample (dotted line) for the HIV+PTB cohort*

After these nonparametric and semiparametric methods, we fit a Weibull model to the observed subsample. Figure 2.3 shows the Weibull hazard plots in order to validate the goodness-of-fit of the model. Results in detail of the respective estimates via maximum likelihood for the scale and the shape parameters can be found in the complete case data analysis in Chapter 3, Section 3.3.

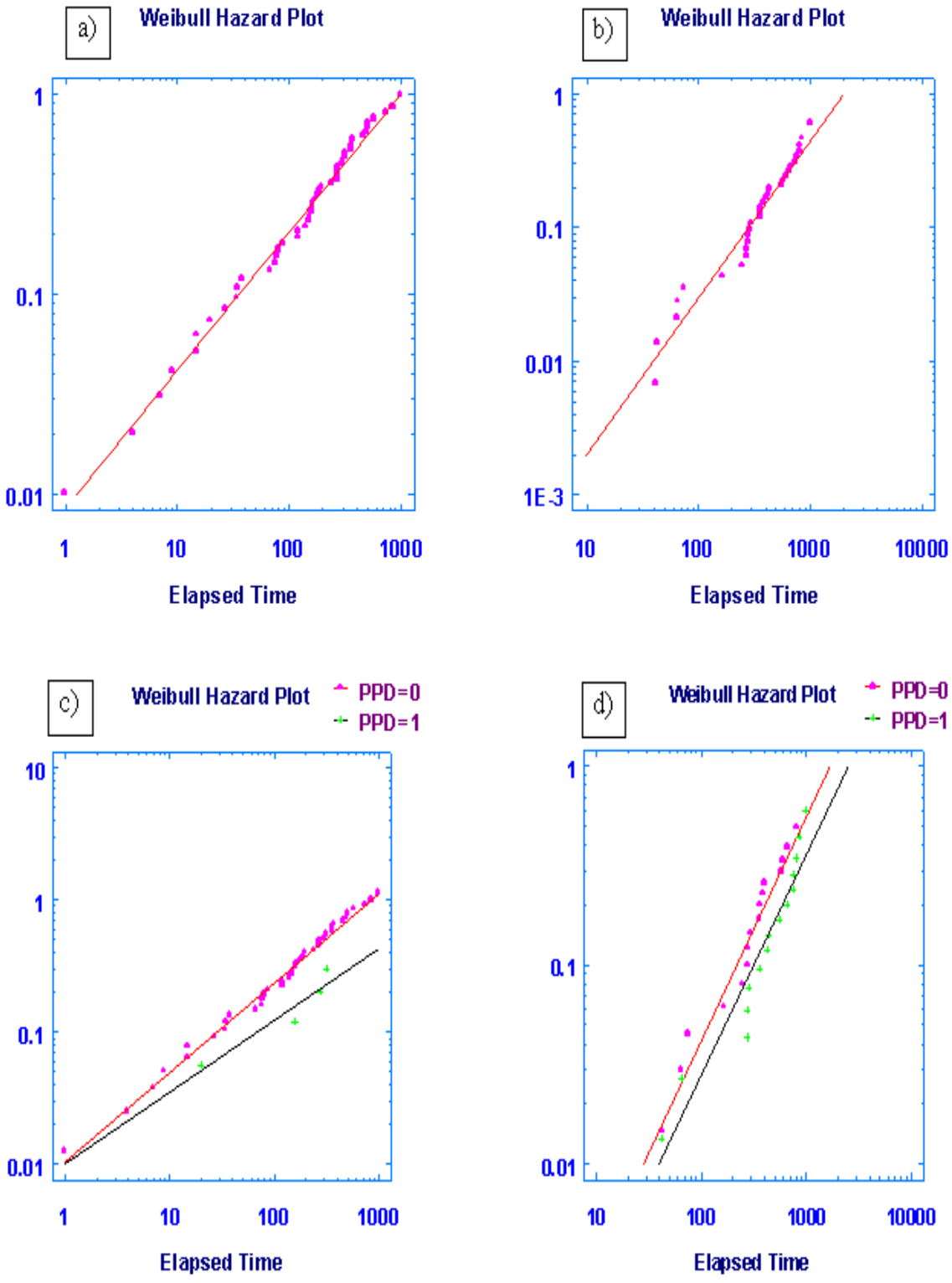

Figure 2.3: *Weibull hazard plots for a) $CD4 \leq 14$, b) $CD4 > 14$, c) $CD4 \leq 14$ stratified by PPD and d) $CD4 > 14$ stratified by PPD*

2.4 A naive pointwise lower/upper bound for survival estimates

Suppose that for a given binary covariate X (in our case the dichotomized $CD4$ or PPD) we are interested in estimating $S_{X=0}(t)$ or, analogously, $S_{X=1}(t)$, from a sample with an important proportion of missing in the covariate X.

A first approach is to obtain a lower/upper bound for these quantities based on considering the worst/best imputed dataset from the observed information. In other words, it means to impute the missing values according to the following algorithm:

The worst scenario for $X = 0$ –and the best for $X = 1$– at time t, compatible with the observed data comes from allocating all the missing values in X to the group $X = 0$ if $Y \leq t$ and $\delta = 1 \rightarrow X = 0$, and to $X = 1$ otherwise, where Y is the observed survival time and δ is the censoring indicator.

Suppose that at some time we allocated "a" cases in one of the previous steps. With the previous allocations, terms in the Kaplan–Meier estimation expression for $X = 1$ at time t get improved because we increase the number of individuals at risk with $X = 1$ with censored data, and therefore we improve the survival for category $X = 1$ (because $1 - \frac{d}{r+a} > 1 - \frac{d}{r}\ \forall a$). In a similar way, individuals allocated to $X = 0$ can reduce the survival for $X = 0$ because $1 - \frac{d+a}{r+a} \leq 1 - \frac{d}{r} \Leftrightarrow r \geq d, \forall a$. By symmetry, in a similar way, we obtain the best scenario for $X = 0$ –and the worst for $X = 1$– at time t, by replacing in the algorithm $X = 1$ by $X = 0$ and viceversa.

In general, the corresponding intervals will be as wider as bigger the proportion of missing would be. If the proportion of missing is high, which is the case in our HIV+PTB dataset, the lower/upper bounds are not very informative. Table 2.7 shows these intervals for the covariates $CD4$ and PPD in our cohort. In particular, these intervals contain all the results coming from the sensitivity analysis in Chapter 6, Table 6.4.

If we compute at each uncensored time the respective bounds for both categories, we obtain the plots in Figure 2.4. Once more, survival functions obtained with the semiparametric method introduced in Chapter 6 (see Figure 6.5) will be in their respective bands. This naive approach will be informative only if, on one hand, there is a significative difference between the stratified survivals and, on the other

hand, there is a small proportion of missing values in the covariates of interest.

Time (in days)	$CD4 = 0$	$CD4 = 1$	$PPD = 0$	$PPD = 1$
90	[0.756, 0.911]	[0.770, 0.964]	[0.691, 0.949]	[0.678, 0.990]
180	[0.653, 0.863]	[0.705, 0.954]	[0.590, 0.917]	[0.596, 0.985]
270	[0.598, 0.834]	[0.677, 0.954]	[0.540, 0.897]	[0.561, 0.985]
360	[0.514, 0.779]	[0.605, 0.926]	[0.457, 0.853]	[0.486, 0.959]
450	[0.431, 0.721]	[0.540, 0.909]	[0.380, 0.811]	[0.410, 0.937]
540	[0.387, 0.685]	[0.528, 0.909]	[0.340, 0.784]	[0.393, 0.936]
630	[0.336, 0.642]	[0.503, 0.898]	[0.301, 0.757]	[0.359, 0.925]
720	[0.300, 0.613]	[0.485, 0.897]	[0.277, 0.743]	[0.332, 0.913]
810	[0.287, 0.602]	[0.458, 0.880]	[0.268, 0.732]	[0.309, 0.884]
900	[0.251, 0.569]	[0.365, 0.809]	[0.228, 0.696]	[0.244, 0.837]
990	[0.243, 0.566]	[0.350, 0.806]	[0.220, 0.694]	[0.234, 0.834]
1080	[0.198, 0.526]	[0.328, 0.803]	[0.192, 0.658]	[0.200, 0.785]

Table 2.7: *Lower-upper bounds for the estimation of the stratified survival for the covariates $CD4$ and PPD based on the re-allocation, at each time, of the individuals with missing covariates to the worst-best option. Results shown every three months*

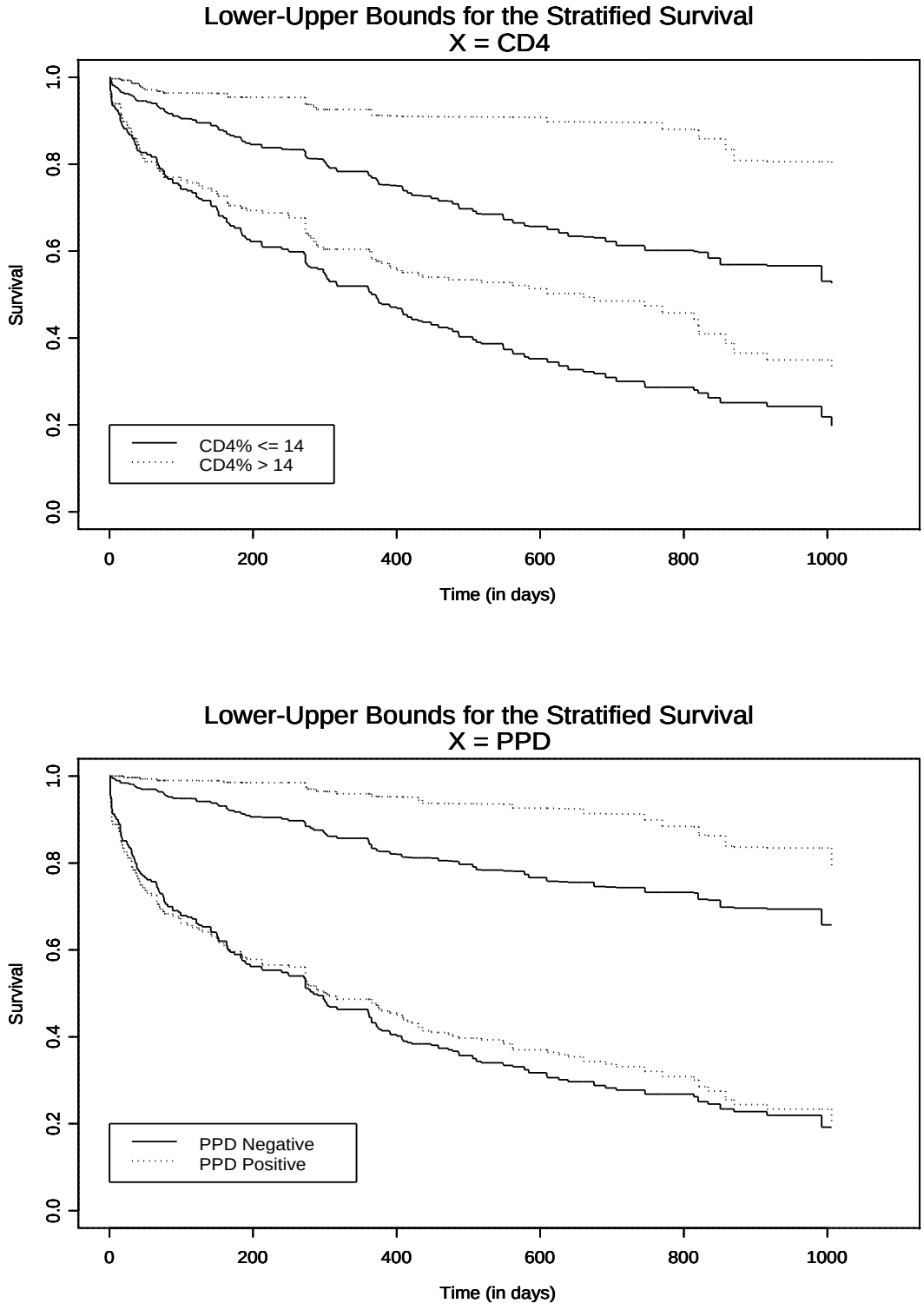

Figure 2.4: *Lower-upper bounds for the estimation of the stratified survival for the covariates $CD4$ and PPD based on the allocation of missing values to the worst-best case, at each death-time*