CHANGES IN CAUSES OF DEATH IN HIV/AIDS PATIENTS IN BRAZIL IN THE HAART ERA

by

Antonio Guilherme Fonseca Pacheco

MD, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil, 1995

MSc, Escola Nacional de Saúde Pública, Rio de Janeiro, Brazil, 2001

Submitted to the Graduate Faculty of

Department of Epidemiology

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2008

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This dissertation was presented

by

Antonio Guilherme Fonseca Pacheco

It was defended on

July 23, 2008

and approved by

Dissertation Advisor: Lee H. Harrison, MD
Professor
Department of Medicine, School of Medicine
Departments of Epidemiology and Infectious Diseases and Microbiology
Graduate School of Public Health
University of Pittsburgh

Stewart Anderson, PhD, MA
Professor
Department of Biostatistics, Graduate School of Public Health
University of Pittsburgh

Leland Yee, PhD
Assistant Professor
Department of Epidemiology, Graduate School of Public Health
Department of Medicine, School of Medicine
University of Pittsburgh

Lawrence Kingsley, DrPH, MPH
Associate Professor
Departments of Epidemiology and Infectious Diseases and Microbiology
Graduate School of Public Health
University of Pittsburgh

Lee H. Harrison, MD

CHANGES IN CAUSES OF DEATH IN HIV/AIDS PATIENTS IN BRAZIL IN THE

HAART ERA

Antonio Guilherme Fonseca Pacheco, PhD

University of Pittsburgh, 2008

The introduction and widespread use of highly active antiretroviral therapy

(HAART) in the past decade has changed the profile of HIV/AIDS epidemic. Marked

decreases in mortality and morbidity have been reported in low- and high-income

settings. Recently, significant relative increases in non-AIDS-associated conditions in

HIV-infected individuals have been reported in developed countries.

In Brazil, where access to HAART has been universal for all eligible patients

since 1996, a steep decrease in mortality among HIV/AIDS patients has also been

documented, but the rates have been stable since 1999. So far, no data have been

available about the pattern of non-HIV-related mortality in these patients. In these

studies, we assessed temporal changes in causes of death among HIV-infected

patients in Brazil.

In the first paper using Brazilian national mortality data, we demonstrate that,

between 1999 and 2004, the odds of having conditions not usually considered to be

related to HIV-infection among individuals who had HIV/AIDS listed on their death

certificate significantly increased over time. Specific diseases that increased were

cardiovascular diseases and diabetes mellitus.

In the second paper, we studied temporal trends in cause of death in a cohort of HIV-infected patients in Rio de Janeiro. Results showed an increase of non-AIDS causes of death and a decrease of AIDS causes of death in HIV-infected patients, which appeared to be driven by an aging of the population.

This cohort study was facilitated by a linkage algorithm that was developed to recover vital status from patients lost to follow-up. In a third paper that validated the algorithm, sensitivity and specificity were found to be 95% and 100%, respectively. In addition, the use of the algorithm led to a 50% increase in the observed mortality rate.

These findings have major public health and programmatic implications for developing countries that are scaling-up access to antiretroviral therapy. In the HAART era, HIV infection has become a manageable disease and is now associated with an increase in chronic illness. Public health measures that are not normally targeted to this population need to be included in their regular care, such as smoking cessation, lipid-lowering drugs, and hypertension control.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 BACKGROUND AND LITERATURE REVIEW

The introduction and widespread use of highly active antiretroviral therapy (HAART) in the past decade deeply changed the profile of HIV/AIDS epidemic. Marked decreases in mortality and morbidity have been reported in both low- and high-income settings [1-4], turning a deadly disease into a manageable chronic condition. One immediate consequence of the increased survival of HIV/AIDS patients is allowing more time at risk for causes of death that may or may not be related to the underlying infection.

Several studies in developed countries both population-based and cohorts have documented these changes, showing that although HIV/AIDS-related causes of death remain at the top of the list, other conditions, including Diabetes Mellitus (DM), cardio-vascular diseases (CVD), cancer, liver and renal diseases have been increasingly reported in HIV-infected individuals [5-9]. Using combined information on HIV/AIDS surveillance with vital statistics data in New York City, Sackoff et al. [6] showed an increase in the proportion of non-HIV/AIDS-related causes of death from 19.8% to 26.3% between 1999 and 2004, with CVD ranking second among them and DM reaching 4.4% among women. In another population-based study in the U.S., Selik et al., showed that patients who had HIV/AIDS mentioned on the

death certificate, had CVD increased from 4% to 7.7% from 1987 to 1999, with a steeper increase between 1996 and 1999 [9].

Similar trends have also been documented in cohort studies. In the CASCADE Collaboration, CVD and DM as a group increased from 1.3% in the pre-HAART era to 4.3% in post-HAART era [7]. In the HIV Outpatient Study, Palella et al. [8] showed that non-HIV/AIDS-related causes of death increased from 13.1% to 42.5% between 1996 and 2004, with CVD being the leading cause of death in this group in 2004. Crum et al., studying an HIV cohort of U.S. military beneficiaries, also showed increases in cardiac disease (from 8% to 22%) and DM (from 0% to 3%) being listed on the death certificates in the pre- and post-HAART era, respectively [5]. In a more recent paper, Lau et al. showed that the risk of non-HIV-related deaths was even higher than HIV-related causes for individuals with CD4+ counts greater than 200 cells/mm^3 in the HAART era [10].

Even though some controversy exists about the association between the use of antiretroviral therapy (ART) and cardiovascular diseases [11], growing evidence has been pointing to increased risk of ART use, especially protease inhibitors (PI) with cardiovascular events.

A possible contributing role of uncontrolled HIV replication on the risk of non-HIV-related deaths has also been suggested [12], even though evidence so far only corroborates this finding for liver-related disease death rates [13] and hepatitis C virus coinfection death rates [14].

1.1 ROLE OF ART IN CVD RISK FACTORS

After the description of peripheral lipodystrophy, hyperlipidemia and insulin resistance in patients taking PI-containing ART in 1998 [15], known risk factors for CVDs, several studies have shown an association between ART and dyslipidemia, especially PI-containing regimens [16-21].

In a retrospective study of 200 patients, 40 patients were treated with dual nucleoside reverse transcriptase inhibitors (NRTI) plus saquinavir alone, 40 patients received NRTI and indinavir alone, 40 subjects received NRTI plus ritonavir alone, 40 patients received NRTI with (13 cases) or without (27 patients) nonnucleoside reverse transcriptase inhibitors (NNRTI), and 40 patients did not receive antiretroviral therapy, Manfredi et al. [17] showed that hypertrygliceridaemia was significantly more frequent and severe in patients taking ritonavir vs. indinavir (P<0.001), and in subjects given indinavir vs. all remaining patients (either treated or not) (P<0.001), while isolated saquinavir use was associated with higher triglyceride levels than NRTI-NNRTI treatment alone, or no antiretroviral therapy (P<0.03).

Results from the Data Collection on Adverse events of Anti-HIV Drugs (D:A:D) study [16] showed similar results, with increased prevalence of elevated total cholesterol (> 6.2 mmol/l) observed among subjects receiving an NNRTI but no PI [odds ratio (OR), 1.79; 95% confidence interval (CI), 1.45–2.22], PI but no NNRTI (OR, 2.35; 95% CI, 1.92–2.87), or NNRTI + PI (OR, 5.48; 95% CI, 4.34–6.91) compared to the prevalence among ART-naïve subjects.

High Framingham scores with HAART use were found in a Norwegian cross-sectional study with 714 patients [19]. The prevalence of a 10-year estimated CHD risk of >20% was 11.9% in patients on HAART compared to 5.3% in controls (P=0.004). The same group also studied differences in the components of the metabolic syndrome comparing HIV-infected patients with or without HAART, compared to HIV-negative individuals [20]. They found that in non-overweight subjects the prevalence of the metabolic syndrome was 15% in HAART-treated patients, 2% in HAART-naïve (p=0.019) and 2% in controls (p=0.020). The prevalence of insulin resistance in non-overweight subjects was also higher in HAART treated than in controls, 39% vs 18% (p=0.012). In non-overweight patients with lipodystrophy the metabolic syndrome was diagnosed in 21% and insulin resistance in 49%.

Cheseaux et al. showed that hyperlipidemia was also present in HIV-infected children treated with PI-containing regimens [21]. Plasma cholesterol levels were significantly elevated in those taking PI, being more pronounced for those given ritonavir (from 3.3 ± 0.7 mmol/L to 6.3 ± 2.8 mmol/L; p=0.03) than for nelfinavir (from 3.0 ± 0.7 mmol/L to 4.9 ± 1.0mmol/L; p<0.001). Cholesterol levels exceeded 10.0 mmol/L in 6% of PI-treated children and culminated at 13.8 mmol/L. Even though the authors conclude that these changes have minimal risk for heart disease development in children, it corroborates the effect of PIs over lipid disorders.

Also of note is the fact that HIV infection per se may be involved in lipid profile changes in infected individuals, and that treatment may have a modulation

effect over it. Riddler et al. demonstrated in 50 men form the MACS study that significant declines occurred in total (TC), low-density lipoprotein (LDL) and high-density lipoprotein (HDL) cholesterol levels after HIV seroconversion. Treatment with HAART elevated the levels LDL back to pre-seroconversion levels, had no effect on HDL and increased TC to a mean excess of 20mg/dl in three years. Even though the authors conclude that these levels of TC are expected, the study lack power to make more refined comparisons [18].

Even though still controversial, other risk factors for CVD include DM [22] and hypertension [23].

1.2 ROLE OF ART IN CVD

Table 1.1 describes 11 articles on the relationship between ART and PI-containing ART use and risk of cardiovascular diseases. Even though the first three studies listed failed to establish an association between PIs and cardiovascular disease [24], heart disease [25] and myocardial infarction (MI) [26], respectively, there is growing evidence that ART and especially PI-containing ART in fact increase the risk of such events. Two studies from D:A:D established the correlation of ART and cardiovascular events [27] and also PIs and MI [28]. The results point to an increased risk of these conditions of about 26% per year of exposure. Two other studies pointed out the association between PI use and MI, In the French Hospital Database [29], PI was associated with a relative hazard of 2.56, and in the HOPS with a hazard ratio of 6.5, despite the fact that they had

relatively low power to detect the effect. Heart disease was also shown to be associated both with ART [30] and PI [31] use. In the former study PI use doubled the risk of heart disease in young patients, whereas in the latter both coronary disease and MI had significantly higher incidence in patients taking PI. Data from the HIV Insight study alone [32] showed an association between PI intake for more than 60 days and cardio- and cerebro-vascular events (adjusted HR=1.71, 95%CI=1.08,2.74). In another study combining data from this study with the Athena study [33] use of PI was found to be associated with MI (HR=1.19, 95%CI=1.01,1.40).

Table 1.1. Studies of the association between PI and CVD

Study	No. Patients (follow-up)	Events	Risk with PIs?	Main Findings	Reference
Veterans administration cohort	36 766 (mean: 40 months)	1 207 admissions for CCVE	No	No association between PI (HR=1.23, 95%CI=0.78,1.93) or NNTRI (HR=0.88, 95%CI=0.63,1.22)	[24]
Kaiser Permanente	4 159 (median: 4.1 years; mean: 3.6 years)	72 hospitalizations for CHD events	No	No significant increase in risk before and after PI (6.2 vs. 6.7 events/1000py)	[25]
Meta-analysis of 30 RTs	10 986 (variable)	29 MIs	No	PIs did not increase risk relative to NRTIs (RR=1.69, 95%CI=0.54,7.48)	[26]
California Medicaid cohort	28 513 (median: 2.25; mean: 2.50)	1 360 CHD events	NS	ART treatment in young (18-33 years) individuals increased risk for CHD (RR=2.06, 95%CI=1.42,2.99)	[30]

Table 1.1. Continued:

French hospital database	34 976 (mean: 30.2 months; median without MI:33 months; median with MI: 28 months)	60 MIs	Yes	PIs associated with higher risk of MI (RH=2.56, 95%CI=1.03,6.34)	[29]
Italian cohort	1 551 (median: 36 months	25 CHD events	Yes	PI-containing regimens associated with higher incidence of CHD (9.8/1000py vs. 0.8/1000py, p<0.001) and MI (5.1/100py vs. 0.4/1000py, p<0.001)	[31]
HOPS	5 672 (mean: 3.2 years)	21 MIs	Yes	PI associated with higher risk of MI (OR=7.1, 95%CI=1.6,44.3; adjusted HR=6.5, 95%CI=0.9,47.8)	[34]
D:A:D	23 468 (median: 1.6 years)	126 MIs	NS	Duration of combination ART increases the risk of MI (relative rate per year of exposure=1.26, 95%CI=1.12,1.41)	[28]
D:A:D	23 468 (mean: 1.54 years)	207 CCVE	NS	Duration of combination ART increases the risk of CCVE (relative rate per year of exposure=1.26, 95%CI=1.14,1.38)	[27]
HIV Insight + Athena	18 603 (median: 3.49 years)	318 AtDs (92 MIs) and 2044 deaths	Yes	PI-containing regimen associated with increased risk of MI (HR=1.19, 95%CI=1.01,1.40)	[33]
HIV Insight	7 542 (median = mean = 3.5 years)	127 CCVE	Yes	PI-exposed group had higher rates of CCVE (9.8 vs. 6.5/1000py, p<0.001) and PI exposure ≥ 60 days associated with higher risk of CCVE (adjusted HR=1.71, 95%CI=1.08,2.74)	[32]

ART = antiretroviral therapy; CCVE = cardio- and cerebrovascular events; CHD = cardiovascular heart disease; D:A:D = Data Collection on Adverse Events of Anti-HIV Drugs study; HOPS = HIV OutPatient Study; MIs = myocardial infarctions; NNRTI = non-nucleoside reverse transcriptase inhibitor; NRTI = nucleoside (nucleotide) reverse transcriptase inhibitors; PI = protease inhibitor; NS = not studied; RTs = Randomized trials; RR = risk ratio; HR = hazard ratio; RH = relative hazard; AtD = atherosclerotic disease

The current first-line PI-containing regimen for naïve HIV-infected patients, lopinavir boosted with ritonavir [35-37] has been shown to impact on CVD risk factors [38]. Even though newer PIs such as atazanavir did not show increase for CVD risk when used without ritonavir boost [39], conflicting results have been reported when ritonavir is used [40, 41]. Some other newer PIs that would be more 'lipid-friendly', such as tipranavir and darunavir still need more studies to define their impact on lipid profile of their chronic use [42].

These factors taken together have major public health implications, because the profile of HIV-infected patients is changing and that care for these patients should include other measures to minimize the risk for these diseases, including smoking cessation and weight control.

In Brazil, where access to HAART has been universal for all patients who qualify since 1996, a steep decrease in mortality among HIV/AIDS patients has also been documented, but the rates have been fairly stable between 1999 and 2004 [3, 4]. So far, no data are available about the transition of non-HIV-related mortality in these patients. This is public health policy issue, both in terms of medication schemes and assessment of patients.

The objective of this project is to study the profile of non-HIV-related causes of death in HIV-infected patients in Brazil, both in the population and individual level and its association with ARV use.

2.0 SPECIFIC AIMS

1. To study changes in causes of death in HIV/AIDS patients in Brazil over time in the HAART era at the national level using death certificate data.

2. To study changes in causes of death in HIV/AIDS patients over time in the HAART era in a cohort of HIV+ patients in Rio de Janeiro, using death certificate data, combined with a standardized algorithm for ascertainment of cause of death.

   o Describe HIV- and non-HIV-related death rates and the proportion of non-HIV-related causes of death over time.

   o Search for factors associated with increased non-HIV-related causes of death (e.g. CD4 levels, age)

3. To develop a computerized algorithm to perform linkage between cohort databases and other databases that contain relevant information for research, that yield high sensitivity and specificity, focusing especially on significantly reducing losses to follow-up using death certificate databases in Brazil.

3.0 PAPER #1: INCREASE IN NON-AIDS RELATED CONDITIONS AS CAUSES OF DEATH AMONG HIV-INFECTED INDIVIDUALS IN THE HAART ERA IN BRAZIL

Antonio G. Pacheco[1,2*], Suely H. Tuboi[2,3], José C. Faulhaber[3], Lee H. Harrison[2], Mauro Schechter[4]

[1] Department of Epidemiology and Quantitative Methods in Health, National School of Public Health/Scientific Computing Program, FIOCRUZ, Brazil, [2] Infectious Diseases Epidemiology Research Unit, Graduate School of Public Health and School of Medicine, University of Pittsburgh, PA, USA, [3] Projeto Praça Onze, Hospital Escola São Francisco de Assis, Universidade Federal do Rio de Janeiro, Brazil, [4] AIDS Research Laboratory, Hospital Universitario Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Brazil

3.1 ABSTRACT

Background: In 1996, Brazil became the first developing country to provide free and universal access to HAART. Although a decrease in overall mortality has been documented, there are no published data on the impact of HAART on causes of death among HIV-infected individuals in Brazil. We assessed temporal trends of mortality due to cardiovascular diseases (CVD), diabetes mellitus (DM) and other conditions generally not associated with HIV-infection among persons with and without HIV infection in Brazil between 1999 and 2004.

Methodology/Principal Findings: Odds ratios were used to compare causes of death in individuals who had HIV/AIDS listed on any field of the death certificate with those who did not. Logistic regression models were fitted with generalized estimating equations to account for spatial correlation; co-variables were added to the models to control for potential confounding. Of 5,856,056 deaths reported in Brazil between 1999 and 2004 67,249 (1.15%) had HIV/AIDS listed on the death certificate and non-HIV-related conditions were listed on 16.3% in 1999, increasing to 24.1% by 2004 ($p<0.001$). The adjusted average yearly increases were 8% and 0.8% for CVD ($p<0.001$), and 12% and 2.8% for DM ($p<0.001$), for those who had and did not have HIV/AIDS listed on the death certificate, respectively. Similar results were found for these conditions as underlying causes of death.

Conclusions/Significance: In Brazil between 1999 and 2004 conditions usually considered not to be related to HIV-infection appeared to become more likely causes of death over time than reported causes of death among individuals who

had HIV/AIDS listed on the death certificate than in those who did not. This observation has important programmatic implications for developing countries that are scaling-up access to antiretroviral therapy.

3.2 INTRODUCTION

The introduction and widespread use of highly active antiretroviral therapy (HAART) has had a profound impact on the HIV/AIDS epidemic, turning a fatal disease into a manageable chronic condition. As a consequence, marked decreases in AIDS-related morbidity and mortality have been reported in both low- and high-income settings [1-4, 43]. However, concerns have been raised about the consequences of prolonged exposure to antiretroviral drugs, with some evidence for an association between the use of protease inhibitors (PIs) and diabetes mellitus (DM) [22] and cardiovascular diseases (CVD) [16, 28]. It has also been recently suggested that HIV itself may play a contributing role in the pathogenesis of some these conditions [12]. Several recent studies conducted in developed countries have documented changes in mortality patterns after the introduction of HAART. Although causes of death traditionally associated with HIV/AIDS continue to play a predominant role, other conditions, including DM, CVD, cancer, liver and renal diseases have been increasingly reported [5-9, 44, 45].

In 1996, Brazil became the first developing country to provide free and universal access to HAART. In 2004 147,500 patients were receiving HAART through the Brazil public health system [46]. In 2005 WHO estimated that ART

coverage in Brazil was in excess of 80% [47]. Although a decrease in overall mortality has been documented [3, 4], there are no published data on the impact of HAART on causes of death among HIV-infected individuals in Brazil.

To our knowledge, there are no population-based studies that have investigated temporal changes in causes of death among HIV/AIDS patients in developing countries after the introduction of HAART. In the present study, we assessed temporal trends of overall mortality and of selected conditions usually considered not to be related to HIV-infection as causes of death between 1999 and 2004 in individuals who had and who did not have HIV/AIDS listed on their death certificate.

3.3 RESULTS

A total of 5,856,056 deaths were reported in Brazil between 1999 and 2004. Of these, 67,249 (1.15%) had HIV/AIDS reported in any field of the death certificate, corresponding to a stable rate of approximately 6.4 cases/100,000 inhabitants per year during the study period (p-value for trend=0.67, Figure 3.1A, solid circles). According to official figures from the Brazil Ministry of Health [48], the death rate associated with HIV/AIDS decreased from 9.3/100,000 in 1996 to 6.26/100,000 in 1999, and, according to our data, it has remained relatively stable until 2004 (Figure 3.1A, open circles).

In the HIV-group, non-HIV-associated causes of death were listed in 16.3% of the death certificates in 1999 and steadily increased to 24.1% in 2004, whereas

in the non-HIV group this ranged from 67.4% to 72.1% in the same period. After adjustment for age, gender and state of residence, the average yearly increase of non-HIV-associated causes of death in the HIV group was 7.98% (95%CI=6.65,9.33; p-value<0.001) and 2.98% for the non-HIV group (95%CI=2.22,3.75; p-value<0.001). The slopes of the increase for the two groups were significantly different (p-value<0.001). Figure 3.1B shows temporal trends for the HIV and for the non-HIV groups, with year as a categorical variable, setting 1999 as the baseline year. In this model, the OR for having a non-HIV-associated condition listed on the death certificate in 2004 compared to 1999 was 1.6 (95%CI=1.48,1.72; p-value<0.001) for the HIV group, and 1.19 (95%CI=1.14,1.23; p-value<0.001) for the non-HIV group and (Table 3.1); the interaction with year as a categorical variable was also highly significant (p-value<0.001).

CVD increased in the HIV group from 4.3% in 1999 to 6.4% in 2004 (adjusted average increase of 7.79% per year; (95%CI=5.74,9.66; p-value<0.001). In the non-HIV group, an increase of 0.80% per year was observed (95%CI=0.28,1.33; p-value=0.002), from 36.2% in 1999 to 39.3% in 2004. Compared to 1999, the OR for having CVD listed on the death certificate in 2004 was 1.5 (95%CI=1.34,1.68; p-value<0.001) for the HIV group and 1.07 (95%CI=1.04,1.10; p-value<0.001) for the non-HIV groups (Table 5.2). Temporal trends for both groups are shown in Figure 5.2A; both interactions between year as a continuous or a categorical variable were significantly different between the two groups (p-value<0.001).

A similar phenomenon was observed for DM, in which an adjusted annual increase of 12.3% (95%CI=8.13,16.56;, p-value<0.001), from 0.6% in 1999 to 1.4% in 2004 was observed in the HIV group, in contrast with an adjusted increase of 2.83% per year in the non-HIV group (95%CI=1.95,3.72,, p-value<0.001), from 6.2% in 1999 to 7.7% in 2004 (Figure 5.2 B, Table 5.3). Compared to 1999, the OR for having DM listed on the death certificate in 2004 was 2.16 (95%CI=1.65,2.82; p-value<0.001) for the HIV group and 1.2 (95%CI=1.13,1.26; p-value<0.001) for the non-HIV group. The increase in the HIV group was again steeper than in the non-HIV group, both with year as a continuous or as a categorical variable (p<0.001).

In the analysis of underlying causes of death among HIV-infected patients, non-HIV-related conditions, CVD and DM increased significantly from 1999 to 2004, a result that corroborates our findings when all causes listed on death certificates were analyzed (data not shown).

3.4 DISCUSSION

The online availability of all death certificates issued in Brazil provided the opportunity to compare reported causes of death in 1999, the first year in which death certificates contained primary, secondary, and contributing causes of death according to ICD-10 codes, with later years among individuals who had HIV/AIDS listed on the death certificate.. To our knowledge, this is the first report on temporal

changes in causes of death among HIV/AIDS patients at the population level in a developing country in the HAART era.

The present study suggests that, in Brazil, similar to what has been reported from developed countries, mortality patterns among HIV/AIDS patients are changing in the HAART era. We found that, in comparison to 1999, there was a steady and significantly larger increase in the frequency with which conditions not usually associated with HIV/AIDS were listed as causes of death increased for individuals who also had HIV listed on the death certificate than in individuals who did not, from 16.3% in 1999 to 24.1% in 2004, representing 14,746 deaths in the period. In particular, listing of CVD or DM as causes of death represented 3,746 and 744 deaths, respectively, both appearing to become more likely causes of death over time in individuals whose death certificate also included HIV/AIDS than in those who did not. Additionally, during the same period there were statistically significant increases in these conditions as underlying causes of death among individuals who had HIV/AIDS mentioned on their death certificates.

We speculate that theses changes are not explained by aging of the population alone, given that the mean age of death in the non-HIV group increased marginally more than the mean age of death in the HIV group (data not shown). Thus, certain potentially preventable and/or treatable conditions, such as CVD and DM, may have played significant roles in these changes, given that the proportion of death certificates in which these conditions are listed increased significantly faster in individuals for whom HIV was listed on the death certificate than in those for whom it was not listed.

16

Our results are in agreement with reports from developed countries where the sharp decrease in mortality following the introduction of HAART was accompanied by significant changes in mortality patterns among HIV-infected individuals. In these countries, after a steep decrease in mortality rates following the introduction of HAART, mortality rates have been reasonably stable since the late 1990's. For example, in the United States, mortality rates declined abruptly in 1994/1995, but remained stable from 1998 onwards, at approximately 7 deaths/100,000 population [49, 50]. In countries where HIV prevalence is well defined and thus could be used as the denominator, a steady increase in the proportion of deaths attributed to conditions that generally are not attributed to HIV infection, such as CVD and DM [5-9, 43], has been reported. We were not able to perform similar analyses, given the absence of reliable estimates of HIV prevalence in most regions of Brazil.

Our results are also in agreement with what has been reported in population-based studies conducted in developed countries. For example, combining information on HIV/AIDS surveillance in New York City with vital statistics data, Sackoff et al. [6] showed an increase in the proportion of non-HIV/AIDS-related causes of death from 19.8% to 26.3% between 1999 and 2004, with CVD ranking second overall and DM reaching 4.4% of all non-HIV-related causes of death in women. In another population-based study in the U.S., Selik et al., using a similar approach to ours but limiting the analysis only to those who had HIV/AIDS listed on the death certificate, showed that CVD increased from 4% to 7.7% from 1987 to 1999, with a steeper increase between 1996 and 1999 [9].

Similar trends have also been documented in cohort studies. In the CASCADE Collaboration, CVD and DM as a group increased from 1.3% in the pre-HAART era to 4.3% in post-HAART era [7]. In the HIV Outpatient Study, Palella et al. showed that non-HIV/AIDS-related causes of death increased from 13.1% to 42.5% between 1996 and 2004 in 12 clinics in the United States, with CVD being leading cause of death in this group in 2004 [8]. Crum et al., studying an HIV cohort of U.S. military beneficiaries, also showed increases in cardiac disease (from 8% to 22%) and DM (from 0% to 3%) being listed on death certificates in the pre- and post-HAART era, respectively [5]. In a simulation study, Braithwaite et al predicted that HIV-infected patients will be increasingly dying of co-morbidities not related to HIV infection. For example, they predicted that a 30 years old patient with a high CD4 count and low viral load would have a 45% chance of dying from non-HIV related causes, and that up to 35% of these would be CVD [44]. Given the absence of reliable estimates of HIV prevalence in most regions of Brazil, we were not able to provide estimates of actual trends among HIV-infected individuals. Nonetheless, we believe that our findings are likely to reflect a changing mortality pattern among HIV/AIDS patients in Brazil that could be associated with widespread availability of HAART [51], a pattern similar to that in developed countries. It should be noted that it is estimated that over 80% of HIV infected patients who, according to international guidelines, were in need of treatment in Brazil were on antiretroviral therapy during the study period[47].

A major strength of our study is that we analyzed all available data from all death certificates issued in a large developing country over a six year period, which

allowed us to investigate temporal trends in causes of death by comparing individuals who had HIV/AIDS listed in their death certificates to those who did not. Furthermore, the use of any mention of conditions on the death certificate as opposed to only the underlying cause of death allowed us to overcome one of the limitations of the current ICD system, which does not cover some diseases associated with HIV [45], as well as to capture the contributing effect of other conditions. Since HIV-infected individuals now live longer [5, 8] and thus have longer periods of time at risk for chronic conditions associated with aging, an increase in the frequency of CVD and DM is to be expected. There are also data that indicate an association between certain antiretroviral drugs as well as time on therapy and risk for CVD [28, 52]. Additionally, recent data suggest that HIV replication may be associated with increased levels of pro-inflammatory markers which, in turn, are involved in the pathogenesis of CVD [12].

Our study has several limitations. Most importantly, we only analyzed data from death certificates, which may lack sensitivity and specificity for medical conditions [53-55]. Nonetheless, by using an approach akin to MOR [56], a strategy that is commonly utilized in studies that investigate occupational hazards, we were able to estimate relative risks by comparing individuals who had HIV/AIDS cited in their death certificates to those who did not. The main difference from the classical use of MOR was that the focus of our analysis was not to compare risks between groups, but to compare temporal trends among them. This approach was chosen due to the lack of reliable estimates of the prevalence of HIV infection in Brazil, particularly age and gender distribution. Additionally, as is the case for all

19

population-based studies, we cannot exclude the influence of unknown confounders that may have contributed to increased reports of non-HIV associated causes of death or to changes over time in the frequency in which HIV/AIDS is listed on death certificates. The latter, if present, probably did not play a significant role, given the stable rate of HIV/AIDS being listed on death certificates, which in turn is in agreement with reports that indicate that HIV-related mortality has remained stable since 1999 [3, 4]. Information on the proportion of those who had HIV/AIDS listed on their death certificate and who were or had been on antiretroviral therapy is not available, as well as other important confounders such as smoking and other risk factors for CVD, and thus could not be assessed in our analyses. However, since HAART is freely and universally available for all those who qualify for treatment according to treatment guidelines that are virtually identical to those used in developed countries [36, 37], it is likely that the vast majority of those who had HIV/AIDS listed on the death certificate were or had been on HAART.

In conclusion, this is the first study to examine changes in mortality patterns among HIV-infected individuals in Brazil after the introduction of HAART. Of particular importance is the finding that conditions that potentially may be prevented and/or treated, such as CVD and DM, have appeared to become more likely causes of death over time among HIV-infected individuals than in the general population. The immediate implication of this finding is that the clinical management of persons living with HIV/AIDS should include prevention, diagnosis, and treatment of chronic conditions, such as CVD and DM. At the program level,

the Brazilian network of treatment facilities will need to increase its capacity to diagnose and manage co-morbidities that can influence outcome. This, in turn, reinforces the need to integrate HIV/AIDS programs with other public health programs, in order to establish a healthcare infrastructure that is capable to take the necessary measures to prevent, diagnose, and treat these conditions [57]. Finally, for other developing countries that are just starting to scale-up their programs, lessons learned from the Brazilian experience might help to prepare in advance to changes in morbidity and mortality patterns that will likely occur once HAART becomes widely available.

3.5 MATERIALS AND METHODS

In Brazil the death certificate is a standardized form that is filled out by a physician. Until 1998, death certificates only included the underlying cause of death. Since 1999, death certificates, besides containing demographic information, include primary, secondary, and contributing causes of death according to the International Classification of Diseases 10th revision (ICD-10) codes [53]. All death certificates issued in Brazil are entered in datasets without personal identifiers and are available online at

http://tabnet.datasus.gov.br/tabdata/sim/dados/cid10_indice.htm. This national database is known as the Brazilian Mortality Information System (Sistema de Informações sobre Mortalidade [SIM]).

In this study, we investigated trends for all causes of death mentioned on death certificates in Brazil between 1999 and 2004. During the study period Brazil had a population of approximately 173 million people [58]. We compared temporal trends in causes of death among individuals who had HIV/AIDS listed in any field of the death certificate (ICD-10 codes B20-B24, Z21), referred to as the HIV group, with those who did not have HIV/AIDS reported on the death certificate (the non-HIV group) using an approach akin to the mortality odds-ratio (MOR) focusing on temporal trends, not on the risk of dying with the studied causes, which would be the classical use of MOR [56].

We conducted three separate analyses, in which the outcomes were defined as the presence or absence in any field of the death certificate of: [I] non-HIV-associated causes, defined as non-HIV-related neoplasms (C00-C80, except C46 – Kaposi's sarcoma), DM (E10-E14), CVD (I00-I99), except cardiac arrest (I46), digestive diseases (K00-K93), genital-urinary diseases, (N00-N99) and external causes (S00-Y98); [II] CVD (I00-I99), except cardiac arrest (I46); and [III] DM (E10-E14). Additionally, we also examined these outcomes as the underlying cause of death, as defined by the World Health Organization (WHO), in individuals who had HIV/AIDS mentioned on their death certificates [59]. The underlying cause of death, which consists of only one cause per death certificate, is coded by local health departments based on standard rules [60], and is the official figure reported by the Brazil Ministry of Health to WHO for mortality statistics.

Logistic regression models were fitted and co-variables were added to the models to control for possible confounding, including age group (<15; 15-29; 30-39;

40-49; 50-59; >=60 years); gender; year of death; and state of residence. Age groups were used to avoid low numbers in some regions of the curves. Generalized estimating equations (GEE) were used to fit the logistic models and to account for spatial correlation structures in the data (i.e. using state of residence as clusters), which were assumed to be interchangeable for this analysis.

Year of death was treated either as a continuous (linear) or categorical variable in the models. In the former case, linear trends are reported as percentages per year, while in the latter, odds ratios are used to compare yearly changes to the baseline year of 1999. Differences in slopes in temporal trends were tested by an interaction term between HIV status and year. Reference groups were <15 years for age, female for gender, and São Paulo State for state of residence.

All analyses were performed in R for Windows v. 2.4.1 [61], using the package 'geepack' for GEE estimation [62].

3.6 Figures
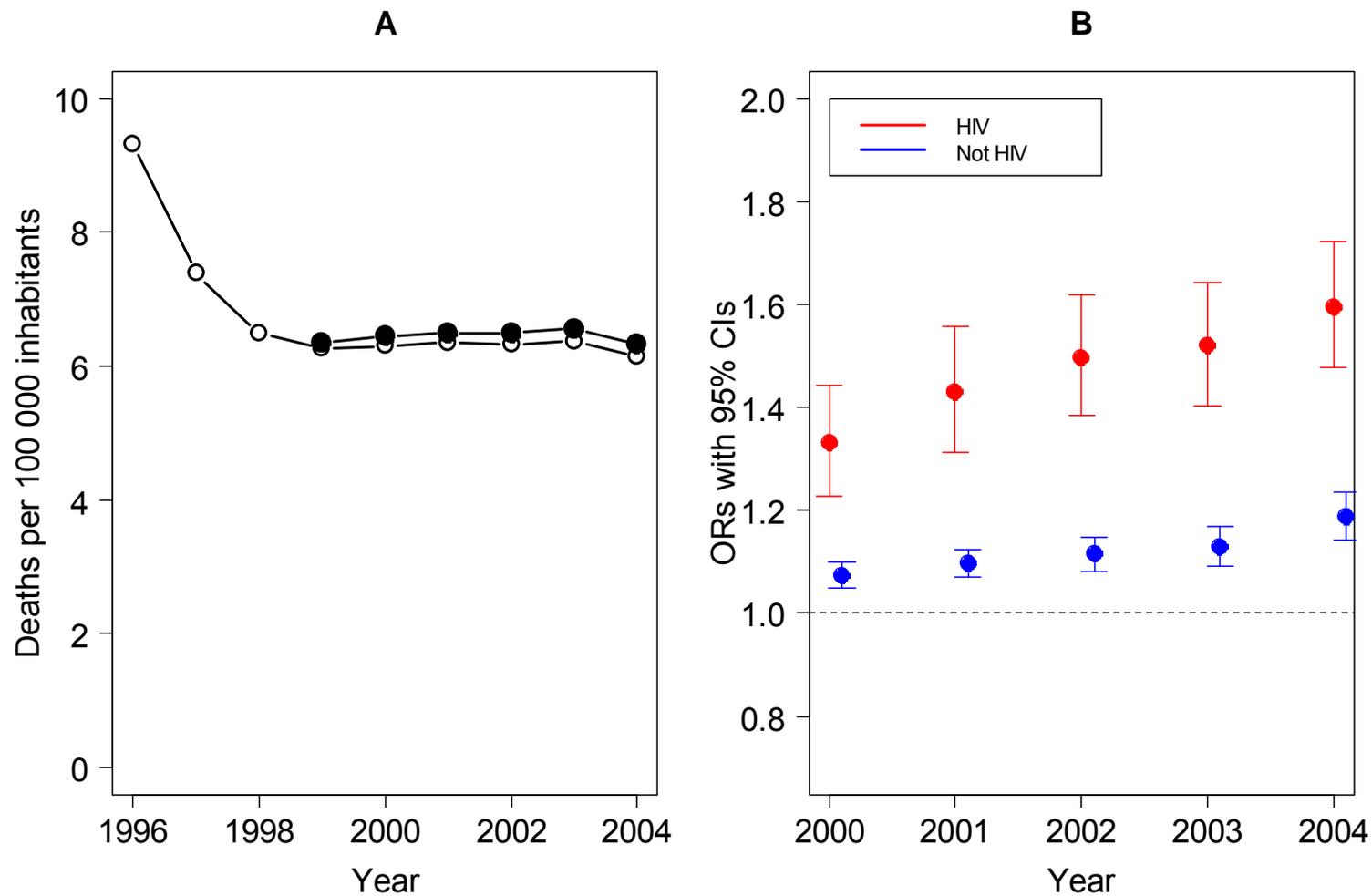
Figure 3.1. Death Rates and non-HIV-related causes of death. A – Death rates per 100 000 inhabitants of individuals that had HIV/AIDS listed on the death certificate, 1999-2004 (p-value for trend = 0.67, solid circles) and AIDS mortality as an underlying disease from 1996 to 2004 (open circles). B – Odds ratios and 95% confidence intervals of non-HIV-related causes of death listed on the death certificate in individuals who had and who did not have HIV listed on the death certificate. Slopes of trends are significantly different between the two groups (p-value<0.001)
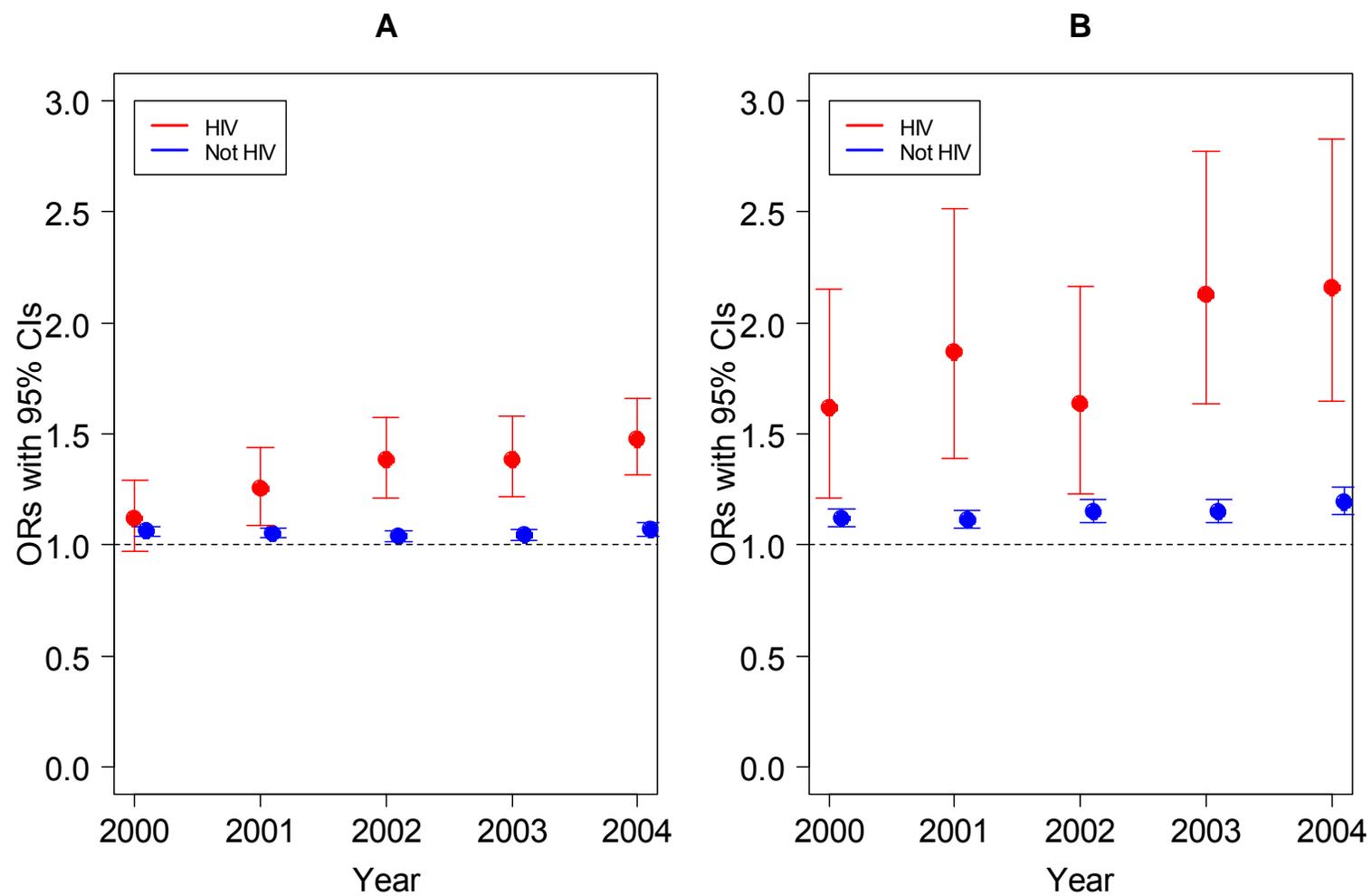
Figure 3.2. CVD and DM in HIV-infected/AIDS patients. Odds ratios and 95% confidence intervals comparing the chance of having the disease listed on the death certificate over time, compared with 1999. A – CVD; B – DM. Slopes for trends are significantly different between groups for CVD (p-value<0.001) and DM (p-value<0.001)

3.7 Tables

Table 3.1. Description and odds ratios of having a non-HIV related condition* in any field of the death certificate stratified by reported HIV status**

Variable		non-HIV (N=5788807)			HIV (N=67249)		
		n (%)	OR (95% CI)	P-value	n (%)	OR (95% CI)	P-value
Outcome	Non-HIV-related	4059422 (70.12%)	-	-	14746 (21.93%)	-	-
Year of death	1999	625767 (67.43%)	Reference		1774 (16.6%)	Reference	
	2000	649305 (69.39%)	1.07 (1.05-1.1)		2317 (21.09%)	1.33 (1.23-1.44)	
	2001	665383 (70.02%)	1.1 (1.07-1.12)		2494 (22.24%)	1.43 (1.31-1.56)	
	2002	685628 (70.58%)	1.11 (1.08-1.15)	<0.001	2628 (23.11%)	1.5 (1.38-1.62)	<0.001
	2003	702692 (70.93%)	1.13 (1.09-1.17)		2737 (23.56%)	1.52 (1.4-1.64)	
	2004	730647 (72.15%)	1.19 (1.14-1.23)		2796 (24.58%)	1.6 (1.48-1.72)	
Gender	Female	1659581 (68.4%)	Reference		4369 (20.98%)	Reference	
	Male	2399841 (71.37%)	1.08 (1-1.17)	0.015	10377 (22.35%)	1.06 (1.01-1.11)	0.03

Table 3.1. Continued:

	<15	97573 (19.46%)	Reference		232 (14.14%)	Reference	
	15-29	352801 (83.44%)	20.44 (16.83-24.81)		2057 (15.9%)	1.15 (1-1.33)	
	30-39	262805 (76.66%)	12.99 (11.2-15.07)		5387 (20.39%)	1.5 (1.31-1.71)	
Age group	**40-49**	390424 (76.88%)	12.79 (11.14-14.69)	<0.001	4186 (24.59%)	1.89 (1.65-2.16)	<0.001
	50-59	535089 (78.34%)	13.85 (12.1-15.86)		1881 (29.81%)	2.47 (2.14-2.86)	
	60+	2400943 (72.69%)	10.39 (9.07-11.9)		974 (35.63%)	3.25 (2.81-3.76)	
	Unknown	19787 (71.11%)	9.29 (7.93-10.88)		29 (15.85%)	1.17 (0.76-1.8)	

*non-HIV-related neoplasms (C00-C80, except C46 – Kaposi's sarcoma), DM (E10-E14), CVD (I00-I99), except cardiac arrest (I46), digestive diseases (K00-K93), genital-urinary

diseases, (N00-N99) and external causes (S00-Y98)

**Adjusted for state of residency

Table 3.2. Description and odds ratios of having CVD* in any field of the death certificate stratified by reported HIV status**

Variable		non-HIV (N=5788807)			HIV (N=67249)		
		n (%)	OR (95% CI)	P-value	n (%)	OR (95% CI)	P-value
Outcome	CVD	**2196228 (37.94%)**	-	-	**3746 (5.57%)**	-	-
Year of death	1999	341452 (36.8%)	Reference		491 (4.59%)	Reference	
	2000	360784 (38.56%)	1.07 (1.05-1.09)		579 (5.27%)	1.14 (1-1.3)	
	2001	368185 (38.75%)	1.06 (1.04-1.09)		647 (5.77%)	1.25 (1.1-1.43)	
	2002	375371 (38.64%)	1.05 (1.02-1.07)	<0.001	721 (6.34%)	1.38 (1.22-1.57)	<0.001
	2003	386117 (38.97%)	1.05 (1.02-1.07)		750 (6.46%)	1.39 (1.23-1.57)	
	2004	404072 (39.9%)	1.07 (1.04-1.1)		800 (7.03%)	1.5 (1.34-1.68)	
Gender	Female	1067296 (43.99%)	Reference		1241 (5.96%)	Reference	
	Male	1168685 (34.76%)	0.79 (0.75-0.82)	<0.01	2747 (5.92%)	0.98 (0.91-1.05)	0.96

Table 3.2. Continued:

	<15	23306 (4.65%)	Reference	101 (6.15%)	Reference
	15-29	33154 (7.84%)	1.75 (1.43-2.14)	480 (3.71%)	0.59 (0.47-0.75)
	30-39	62250 (18.16%)	4.46 (4.01-4.97)	1331 (5.04%)	0.79 (0.62-0.99)
Age group	**40-49**	162441 (31.99%)	9.22 (8.54-9.95)	1103 (6.48%)	1.01 (0.8-1.28)
	50-59	289154 (42.33%)	14.33 (13.34-15.39)	590 (9.35%)	1.51 (1.18-1.93)
	60+	1661862 (50.31%)	19.56 (18.18-21.05)	379 (13.86%)	2.37 (1.85-3.04)
	Unknown	3814 (13.71%)	3.53 (2.48-5.02)	4 (2.19%)	0.35 (0.11-1.16)

p-value <0.001 (first group); <0.001 (second group)

*CVD: ICD10 I00-I99, except cardiac arrest (I46)

**Adjusted for state of residency

Table 3.3. Description and odds ratios of having DM* in any field of the death certificate stratified by reported HIV status

Variable		non-HIV (N=5788807)			HIV (N=67249)		
		n (%)	OR (95% CI)	P-value	n (%)	OR (95% CI)	P-value
Outcome	DM	**411862 (7.11%)**			**744 (1.11%)**	-	-
Year of death	1999	57312 (6.18%)	Reference		63 (0.59%)	Reference	
	2000	65525 (7%)	1.12 (1.08-1.16)		109 (0.99%)	1.62 (1.21-2.15)	
	2001	66886 (7.04%)	1.12 (1.08-1.16)		130 (1.16%)	1.87 (1.39-2.52)	
	2002	71025 (7.31%)	1.15 (1.1-1.2)	<0.001	117 (1.03%)	1.63 (1.23-2.17)	<0.001
	2003	72933 (7.36%)	1.15 (1.1-1.2)		161 (1.39%)	2.13 (1.63-2.77)	
	2004	78181 (7.72%)	1.2 (1.13-1.26)		164 (1.44%)	2.16 (1.65-2.82)	
Gender	Female	235268 (9.7%)	Reference		240 (1.15%)	Reference	
	Male	176594 (5.25%)	0.59 (0.55-0.64)	<0.01	504 (1.09%)	0.9 (0.77-1.04)	0.19

Table 3.3. Continued:

	<15	617 (0.12%)	Reference		0 (0%)	Reference	
	15-29	3309 (0.78%)	Reference		42 (0.32%)	Reference	
	30-39	6811 (1.99%)	4.86 (3.96-5.95)		171 (0.65%)	2.15 (1.53-3.03)	
Age group	**40-49**	22329 (4.4%)	10.59 (8.74-12.82)		244 (1.43%)	4.68 (3.37-6.49)	
	50-59	57967 (8.49%)	20.94 (17.32-25.32)	<0.001	163 (2.58%)	8.58 (6.15-11.97)	<0.001
	60+	320444 (9.7%)	23.07 (19.04-27.96)		122 (4.46%)	15.24 (10.96-21.19)	
	Unknown	385 (1.38%)	3.89 (2.06-7.33)		2 (1.09%)	4 (1-16.09)	

*DM: ICD10 E10-E14, except cardiac arrest (I46)

**Adjusted for state of residency

4.0 PAPER # 2 - TEMPORAL CHANGES IN CAUSES OF DEATH AMONG HIV-INFECTED PATIENTS IN THE HAART ERA IN BRAZIL

Manuscript under preparation

4.1 ABSTRACT

Objective: To assess temporal trends of mortality due to non-AIDS-related conditions in a cohort of HIV infected individuals in Rio de Janeiro, Brazil between 1997 and 2006.

Design: Retrospective cohort study.

Methods: Death rates were analyzed in two-year periods using Poisson models and survival models accounting for competing risks were used to assess association of co-variables. A standardized algorithm was used to ascertain causes of death.

Results: Of the 1,538 eligible patients, 226 (14.7%) died during the study period. Overall, 98 (43.4%) were classified as non-AIDS-related causes. Opportunistic infections were the leading causes of death (37.6%). AIDS-related death rates declined significantly over time (p<0.01). In the most recent period (2005-2006) the rate of non-AIDS related causes of deaths was higher than that of AIDS-related causes of death. CVD increased significantly over time

(p=0.04). There was a 46% increase in risk per 10 years of baseline age for AIDS-related causes of death, with a significant and negative interaction over time (p=0.02).

Conclusions: Our results indicate that in the HAART era there has been a significant change in causes of death among HIV-infected patients in Rio de Janeiro, Brazil. The major driving force behind this change appears to be longer survival due to the significant decrease in deaths due to AIDS-related causes. As access to HAART improves and HIV becomes a chronic condition, integration with other public programs, such as smoking cessation and control of hypertension will become critically important for the long term success of HIV/AIDS programs in developing countries.

Key words: HIV, AIDS, Brazil, causes of death

4.2 INTRODUCTION

In the past decade, the introduction and widespread use of highly active antiretroviral therapy (HAART) has led to marked decreases in mortality and morbidity in low- and high-income countries [1-4], turning a deadly disease into a manageable chronic condition.

In developed countries, population-based and cohort studies have demonstrated that, although HIV/AIDS-related conditions remain the most frequent causes of death, other conditions, including diabetes mellitus (DM), cardiovascular diseases (CVDs), cancer, liver and renal diseases, have become

increasingly frequent in HIV-infected individuals [5-9]. In fact, in some populations, such as patients with CD4+ counts greater than 200 cells/mm^3, non-HIV-related causes of death are now more frequent than HIV-related ones [10].

Due to the increased survival, the HIV-infected population tends to become older [63], a factor that in itself may contribute to the increased frequency of non-HIV/AIDS-related causes of death. An association between certain HAART regimens, particularly those containing protease inhibitors or abacavir, and risk of CVD and conditions such as the metabolic syndrome has been reported [11, 24-34]. A possible contributing role of uncontrolled HIV replication on the risk of non-HIV-related deaths has also been suggested [12].

In Brazil, where free access to HAART has been universal for all patients who qualify for treatment under locally developed guidelines since 1996, although mortality rate among HIV-infected patients decreased sharply after the introduction of HAART, it has remained stable since 1999 [3, 4]. Recently, based on data from the national databases of death certificates, we reported a significant increase between 1999 and 2004 in non-HIV-related conditions, including CVD and DM, as causes of death, [64].

In this study, we describe changes in temporal trends of causes of death in a cohort of HIV-infected patients in Rio de Janeiro, Brazil.

4.3 METHODS

4.3.1 Data sources

All patients included in the study were followed in the Rio de Janeiro cohort. The Rio de Janeiro cohort is a participating site of the ART-LINC collaboration (http://www.ispm.ch/artlinc.html) [1, 65] and of CCASAnet (http://ccasanet.vanderbilt.edu) of the IeDEA network (http://www.iedea-hiv.org). Patients who entered the cohort between 1997 and 2006, were aged 16 years or older, and had at least one follow-up visit were eligible for the present study. Laboratory data, including CD4 counts and viral load, are electronically transferred from the hospital database. Clinical data are abstracted from hospital charts immediately after consultations by trained data abstractors who fill out specially designed forms that contain all relevant fields and variables.

Vital status was ascertained either by abstracting information from medical charts or by linkage with the Rio de Janeiro Mortality Database, using a previously validated algorithm [66]. Information from death certificates obtained through the linkage algorithm, which included primary and contributing causes of death, was used in the ascertainment of causes of death as described below.

4.3.2 Ascertainment of causes of death

Causes of death were determined using the CoDe (Coding Causes of Death in HIV) protocol [67]. Briefly, extensive data were collected from all available sources, including death certificates, medical records, autopsy reports, and information obtained from family members by a trained physician or health

care provider. Two independent reviewers attributed causes of death based on the completed CoDE forms. If the attributed cause of death was the same for both reviewers, it was considered to have been established. If there was disagreement between the two reviewers, or both had coded the cause of death as unknown or unclassifiable, the case was referred to a third reviewer for final classification.

Causes of death were classified as AIDS-related or non-AIDS-related based on the presence or absence of an AIDS-defining condition, according to the Centers for Disease Control and Prevention Classification [68], as a primary or a contributing cause of death. All deaths with unknown primary and contributing causes for which at least one CD4 cell count was available within 6 months prior to the date of death were classified as AIDS-related if CD4 counts were < 200 cells/mm^3 and as non-AIDS-related if ≥ 200 cells/mm^3. Deaths that did not meet any of the above criteria were classified as of unknown cause.

Specific underlying causes of death were further categorized according to whether or not they were AIDS-related as follows: a) AIDS-related (according to CDC definitions): opportunistic diseases; cancer; both cancer and opportunistic diseases; and undefined disease(s); b) non-AIDS–related: hepatitis/liver-related; infectious diseases; non-AIDS-related cancer; external causes; CVD; other diseases; and unknown.

4.3.3 Statistical analysis

Descriptive statistics are presented for the variables considered comparing living patients with deceased patients due to AIDS-related, non-

AIDS-related, and unknown causes of death. For continuous variables, non-parametric Kruskal-Wallis tests were used followed by pairwise Wilcoxon tests with Holm's multiple comparisons corrections; for discrete variables, overall Fisher exact tests were performed.

Two-year rates of AIDS- and non-AIDS-related causes of death and CVD per 100 person-years (py) were calculated and compared over time to detect temporal changes. Poisson regression models were employed and correlation and overdispersion were handled by using quasi-Poisson corrections for variance estimation [69].

Survival analysis regression models were performed in the context of competing risks [70, 71]. This approach makes it possible to account for differences between non-informational censoring due to study termination or independent losses to follow-up and informational censoring, when "losses" are due to death by causes other than the one(s) studied, allowing for correct modeling of the main outcome and competing risk(s) [71-73]. Classically, event-specific hazards are modeled separately for the event of interest and the competing risk, which tends to overestimate the cumulative mortality, since it considers that the competing event does not exist. Another problem with this approach is that the assumption of proportionality across covariates strata will usually be violated and inferences could be biased [72]. A new approach proposed by Fine and Gray [74, 75] allows the estimation of the subdistribution hazards for each event, which takes into account the competing risks, allowing for unbiased estimation of the cumulative incidence functions (CIFs) and is implemented in the 'cmprsk' library [76] in the R software. Adjusted hazard ratios (aHR) of the subdistributions are reported for the models.

Variables analyzed included gender, age at start of observation, age at the end of observation or death, transmission group [heterosexual, men who have sex with men (MSM), intravenous drug users (IDU), other and unknown], baseline CD4 counts, use of HAART and use of PI-containing regimens.

All analyses were performed in R for Windows [61].

Institutional Review Board approval was obtained both from the Hospital Universitário Clementino Fraga Filho, where the cohort is being followed, and the University of Pittsburgh.

4.4 RESULTS

Of the 1,633 potentially eligible patients, 95 (5.8%) females were excluded from the analysis because they were followed only during pregnancy. Of the 1,538 eligible patients, who contributed 7,037 person-years of follow-up, 226 (14.7%) were known to have died during the study period. Of those known to have died, 82 (36.3%) were identified only through using the linkage algorithm with the Rio de Janeiro Mortality database. The mortality rate was 3.2/100 person-years, which is over 50% higher than the rate that would be obtained had these 82 deaths not been identified (2.0/100 py; difference = 1.2; 95%CI = 0.9,1.4; p < 0.01). The median follow-up time was 4.61 years (IQR = 5.63 years) and the loss to follow-up rate, defined as absence of any information for more than one year, was 2.4/100 person-years.

Overall, 111 (49.1%) deaths were considered to be due to AIDS-related causes, whereas 98 deaths (43.4%) were classified as due to non-AIDS-related

causes. Patients who died from non-AIDS related deaths contributed more follow-up time than patients who died from AIDS-related causes (median time = 3.39 py vs. 1.45 py, respectively, p < 0.01) (Table 4.1). Rates of death from AIDS-related causes (1.58/100 py) and from non-AIDS-related causes (1.39/100 py) were not statistically different (p = 0.37).

Baseline CD4 cell counts were significantly lower for deceased patients compared to patients alive at the end of the follow-up period (p < 0.01 for both AIDS- and for non-AIDS-related causes of death). Nonetheless, after multiple comparisons adjustment, the difference in baseline CD4 counts between patients who died from AIDS- and from non-AIDS-related causes of death was not statistically significant (148 vs. 95 cells/mm3, respectively, p = 1).

The distribution of causes of death is shown in Table 4.2. Opportunistic infections were the leading causes of death (37.6%). Among non-AIDS-related causes, infectious diseases were also the most frequent causes (8.4%), followed by external causes (4.9%) and CVD (4.0%).

Death rates remained fairly stable during the study period (p = 0.57), despite a significant decrease when the initial period (1997-1998) was compared to the final period (2005-2006) of observation (p < 0.01) (Table 4.3 and Figure 4.1A). AIDS-related causes of death declined significantly over time (p < 0.01), while non-AIDS-related causes of death increased over time, although not significantly (p = 0.46). Nonetheless, in the most recent period (2005-2006) the rate of non-AIDS related causes of deaths was higher than that of AIDS-related causes of death (1.59/100 py and 1.24/100 py, respectively). Despite the relatively small number of cases, CVD as a cause of death increased significantly over time (p = 0.04).

The cumulative incidence functions for the hazards subdistributions for AIDS-related, non-AIDS-related and unknown causes of death in a competing risks framework is shown in Figure 4.2. At the beginning of the observation period, AIDS-related causes of death had a higher cumulative incidence than non-AIDS related causes of death, but 7 years later the cumulative incidence of non-AIDS related causes of became higher than that of AIDS-related causes of death. In Table 4.4 we present models for the hazard subdistributions for AIDS- and non-AIDS related causes of death, treating age at the start of the observation period as a time-dependent variable and adjusting for baseline CD4 cell count, gender, and risk factor for HIV infection. The models indicate that baseline CD4 cell count is strongly associated with protection for both AIDS- and non-AIDS-related causes of death (10% and 13% per 50 cells increase, respectively) and that intravenous drug use is a significant risk factor for non-AIDS-related causes of death (aHR = 4.96; 95% CI: 2.34,10.52; p < 0.01), but not for AIDS-related causes of death. Although baseline age was not associated with non-AIDS-related causes of death, there was a 46% increase in risk per 10 years of baseline age for AIDS-related causes of death, with a significant and negative interaction over time (p = 0.02), indicating that risk decreased significantly over time.

HAART use during observation time and PI use were not significantly associated with non-AIDS-related causes of death (HR = 0.7; 95%CI: 0.44,1.09; p = 0.11 and HR = 1.23; 95%CI: 0.82,1.86; p = 0.32, respectively).

Using more restrictive criteria that are also part of the CoDe algorithm, that classify deaths according to CD4 cell counts closer to death and if the death was sudden or non-sudden, deaths were further classified as due to

41

immunodeficiency-related causes or to non-immunodeficiency related causes. Death rates due to causes not related to immunodeficiency significantly increased over time (0.54 vs. 0.84 deaths per 100 py, in 1997-1998 and 2005-2006, respectively; p < 0.01), even though death rates due to causes related to immunodeficiency were consistently and significantly higher throughout the study period (data not shown).

4.5 DISCUSSION

Our results show that overall mortality rates in a large cohort of HIV-infected patients in Rio de Janeiro, Brazil, remained fairly stable between 1999 and 2006. This is in agreement with other studies that showed that in Brazil, although mortality rates declined sharply after HAART became widely available, they have remained stable since 1999 [4, 64]. In our study, when AIDS- and non-AIDS-related causes of death are examined separately, a decrease in AIDS-related causes of death and an increase in non-AIDS-related causes of death were observed. When more restrictive criteria were used to classify causes of death, non-immunodeficiency related causes of death increased significantly over time. Even though the temporal trend for non-AIDS-related causes of death did not reach statistical significance, in the most recent study period (2005-2006) non-AIDS-related causes of death became more frequent than AIDS-related causes of death. Despite the small number of cases, during the study period there was a significant increase of CVD as a cause of death (p = 0.04). Thus, the distribution and temporal trends for specific groups of causes of death we encountered are in agreement with reports from developed

countries, in which although opportunistic infections remain the main cause of death, non-AIDS-related causes of death are becoming increasingly common [7, 10].

In a regression model for the hazards subdistribution adjusted for possible confounders, although baseline age was significantly associated with risk for AIDS-related causes of death, the risk significantly decreased over time. On the other hand, baseline age had no significant effect on the risk for non-AIDS-related causes of death. These results indicate that, as patients age because of longer survival times, they become at higher risk of dying from a non-AIDS-related condition. These findings are in accordance with the recent description of HIV cohorts tending to get older [63], as is the case in our cohort, where the median age significantly increased from 33.25 years in 1997 to 39.0 years in 2006 (p < 0.01).

Our results are in agreement with several reports in the literature that have documented changes in causes of death among HIV-infected patients in the HAART era. Using combined information on HIV/AIDS surveillance with vital statistics data from New York City, Sackoff et al. [6] showed an increase in the proportion of non-HIV/AIDS-related causes of death from 19.8% to 26.3% between 1999 and 2004, with CVD ranking second among them. Selik et al. showed that for patients who had HIV/AIDS mentioned on the death certificate in the United States, CVD increased from 4% to 7.7% from 1987 to 1999, with a steeper increase between 1996 and 1999. In the CASCADE Collaboration, CVD and DM as a group increased from 1.3% in the pre-HAART era to 4.3% in post-HAART era [7]. In the HIV Outpatient Study, Palella et al. [8] showed that non-AIDS-related causes of death increased from 13.1% to 42.5% between 1996

and 2004, with CVD being the leading cause of death in this group in 2004. Crum et al., studying a cohort of U.S. military beneficiaries, also showed increases in cardiac disease (from 8% to 22%) and DM (from 0% to 3%) being listed on the death certificates in the pre- and post-HAART era, respectively [5]. We have recently shown that in Brazil between 1999 and 2004 the proportion of non-AIDS-related causes of death reported on death certificates of patients that had HIV/AIDS as one of the causes of death increased significantly, as did CVD [64].

Among the main strengths of our study was the use of a database linkage algorithm that allowed us to determine the vital status of patients who otherwise would be considered to be lost to follow-up, which increased the death rate by 57% and kept the loss to follow-up rate at 2.4/100 person-years. Another strength of our study was the use of the CoDe algorithm [67], which allowed for a standardized classification of causes of death, which would not be possible if only data abstracted from death certificates or medical charts were used.

The main limitation of our study was the relatively small number of deaths (n = 226), which made some of the models we used underpowered for associations. Another limitation was the retrospective nature of the study, which did not allow us to investigate potentially important variables, such as adherence to treatment, smoking status, hepatitis co-infection, and lipid and cholesterol blood levels.

In conclusion, our results indicate that in the HAART era there has been a significant change in causes of death among HIV-infected patients in Rio de Janeiro, Brazil. Although overall death rates remained stable after 1999, deaths

from AIDS-related causes decreased, while deaths from non-AIDS-related causes increased to the point that the latter became more common than the former. The major driving force behind this change appears to be longer survival due to the significant decrease in deaths due to AIDS-related causes. Our findings have major programmatic implications for developing countries. As access to HAART improves and HIV becomes a chronic condition, integration with other public programs, such as smoking cessation and control of hypertension will become critically important for the long term success of HIV/AIDS programs in developing .countries.

4.6 Figures
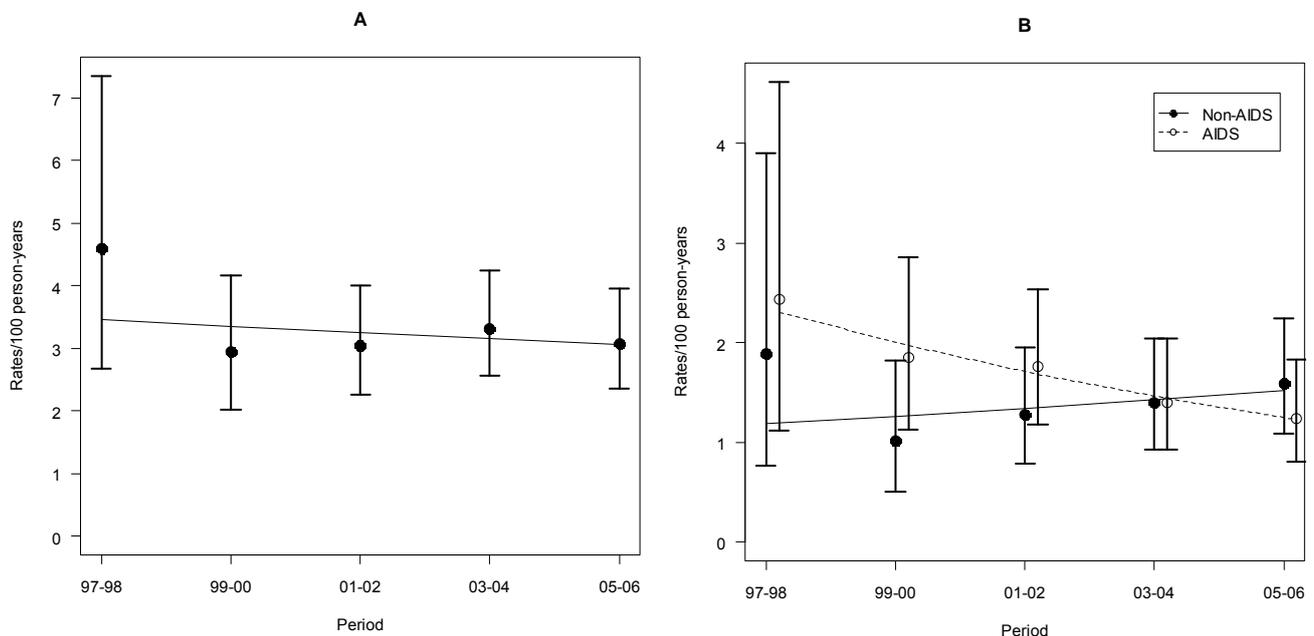
Figure 4.1. Temporal trends of deaths in the cohort. A – Overall deaths; B – AIDS and Non-AIDS-related causes of death
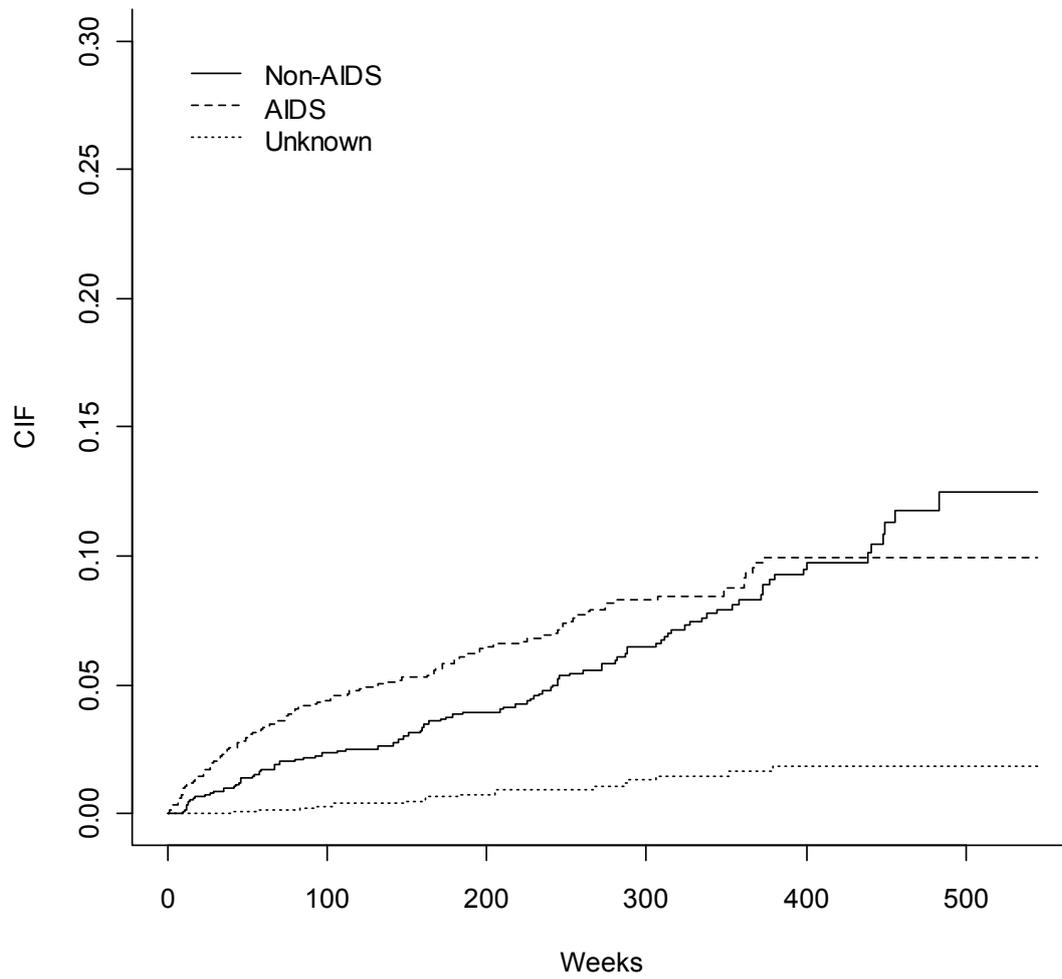
Figure 4.2. Cumulative incidence functions (CIFs) of the subdistributions of deaths associated with AIDS-, non-AIDS-related and unknown causes in a competing risks framework.

4.7 Tables

Table 4.1. General descriptions of variables analyzed

		Alive	AIDS-related[1] Death	Non-AIDS-related[2] Death	Overall p-value[3]	Unknown	Total
PYs	N	1312	111	98	<0.01	17	1538
	Median	4.99	1.45#	3.39#§		3.52	4.61
	IQR	5.8	3.1	4.62		3.52	5.63
	Total	6360.36	250.28	363.14		63.76	7037.54
	%	85.31	7.22	6.37	-	1.11	100
Death rates per 100 person-years		-	1.58	1.39	0.37	0.24	3.21
Age at start	N	1312	111	98	0.02	17	1538
	Median	33.02	35.07	36.37		35.85	33.62
	IQR	14.53	13.29	18.02		15.28	14.81
Age in the end	N	1312	111	98	0.47	17	1538
	Median	38.68	38.02	41.17		37.44	38.7
	IQR	14.97	11.98	15.18		13.95	14.94
Gender	Female	673	46	44	Ref	9	772
	%	51.3	41.44	44.9		52.94	50.2
	Male	639	65	54	0.16	8	766
	%	48.7	58.56	55.1		47.06	49.8
Group	Heterosexual	661	45	48	Ref	7	761
	%	50.38	40.54	48.98		41.18	49.48
	MSM	222	16	6	0.04	5	249
	%	16.92	14.41	6.12		29.41	16.19
	IDU	40	5	6	0.13	1	52
	%	1.14	2.7	9.18		0	1.76
	Other	40	5	6	0.13	1	52
	%	3.05	4.5	6.12		5.88	3.38
	Unknown	374	42	29	0.16	4	449
	%	28.51	37.84	29.59		23.53	29.19
Baseline CD4	N[4]	1248	106	96	<0.01	15	1465
	Median	268.38	95.36#	148.36#		182	250
	IQR	337.05	272.36	299.6		277.71	338.25
HAART ever	No	314	34	26	Ref	6	380
	%	23.93	30.63	26.53		35.29	24.71
	Yes	998	77	72	0.26	11	1158

Table 4.1. Continued:

	%	76.07	69.37	73.47		64.71	75.29
HAART during observaion	No	433	36	31	Ref	6	506
	%	33	32.43	31.63		35.29	32.9
	Yes	879	75	67	0.98	11	1032
	%	67	67.57	68.37		64.71	67.1
PI ever	No	662	57	36	Ref	9	764
	%	50.46	51.35	36.73		52.94	49.67
	Yes	650	54	62	0.07	8	774
	%	49.54	48.65	63.27		47.06	50.33

[1] Pair-wise p-value < 0.05 comparing with alive patients: #

[2] Pair-wise p-value < 0.05 comparing with alive patients: #; comparing with the competing cause: §

[3] Overall p-values for Kruskal-Wallis test in continuous variables and Fisher exact test for discrete ones, including alive and unknown patients. For rates, mid-p exact Poisson tests were used, comparing non-AIDS and AIDS deaths only.

[4] 73 patients did not have a baseline CD4 result

For continuous variables, pair-wise Wilcoxon rank tests with Holm's adjustment for multiple comparisons were also performed.

Ref – Reference group

Table 4.2. Groups of specific causes of death

Groups	N	%
AIDS-related	111	49.1
Opportunistic	85	37.6
Cancer	17	7.5
Both	2	0.9
Other	7	3.1
Non-AIDS-related	98	43.4
Hepatitis/Liver	8	3.5
Infectious	19	8.4
Cancer	8	3.5
External	11	4.9
CVD	9	4.0
Other	43	19.0
Unknown	17	7.5

CVD – Cardiovascular diseases

Table 4.3. Trends of overall, AIDS-related causes, non-AIDS-related causes and CVD

death rates per 100 person-years over time for all patients

	1997-1998	1999-2000	2001-2002	2003-2004	2005-2006	p trend	p 1st and Last [1]
Total	4.59	2.96	3.04	3.32	3.08	0.57	<0.01
AIDS	2.43	1.85	1.76	1.4	1.24	<0.01	0.1
Non-AIDS	1.89	1.02	1.28	1.4	1.59	0.46	0.2
CVD	0	0	0.12	0.26	0.35	0.04	0.05

[1] p-value comparing the first (1997-1998) and the last (2005-2006) periods only

CVD – Cardiovascular diseases considered as either primary or contributing cause of death

Table 4.4. Models for the hazards of the subdistributions of deaths due to AIDS- and to

non-AIDS-related causes including baseline age (per 10 year increase), CD4 cell counts

(per 50 cells increase), gender (reference: female) and risk group (IDU vs. other).

Baseline age treated as a time-dependent variable.

Outcome	Variables	aHR (95% CI)	p
Deaths due to AIDS-related causes	Baseline Age (10 years) [1]	1.46 (1.15,1.85)	<0.01
	Baseline CD4 (50 cells)	0.87 (0.82,0.92)	<0.01
	Male gender	1.07 (0.71,1.59)	0.75
	IDU	1.35 (0.42,4.31)	0.61
Deaths due to non-AIDS-related causes	Baseline Age (10 years) [2]	1.30 (0.94,1.80)	0.12
	Baseline CD4 (50 cells)	0.90 (0.86,0.95)	<0.01
	Male gender	0.84 (0.56,1.26)	0.40
	IDU	4.96 (2.34,10.52)	<0.01

[1] Negative interaction with time (p = 0.02)

[2] Negative interaction with time (p = 0.77)

aHR: Adjusted hazard ratio

5.0 PAPER #3 - VALIDATION OF A HIERARCHICAL DETERMINISTIC RECORD LINKAGE ALGORITHM USING DATA FROM TWO DIFFERENT COHORTS OF HIV-INFECTED INDIVIDUALS AND MORTALITY DATABASES IN BRAZIL

Antonio G. Pacheco[1,2], Valeria. Saraceni[3], Suely H. Tuboi[2,4], Lawrence H. Moulton[5], Richard E. Chaisson[6], Solange C. Cavalcante[3], Betina Durovni[3], José C. Faulhaber[4], Jonathan E. Golub[6], Bonnie King[6], Mauro Schechter[4,7], Lee H. Harrison[2]

[1]ENSP/PROCC – FIOCRUZ, Rio de Janeiro, Brazil

[2] Infectious Diseases Epidemiology Research Unit, Graduate School of Public Health and School of Medicine, University of Pittsburgh, PA, USA

[3]Rio de Janeiro Municipal Health Secretariat, Communicable Diseases Program, Rio de Janeiro, Brazil

[4]Projeto Praça Onze, Hospital Escola São Francisco de Assis, Universidade Federal do Rio de Janeiro, Brazil

[5]Johns Hopkins Bloomberg School of Public Health, International Health, Baltimore, United States

[6]Johns Hopkins University School of Medicine, Center for Tuberculosis Research, Baltimore, United States

[7]AIDS Research Laboratory, Hospital Universitario Clementino Fraga Filho, Universidade Federal do Rio de Janeiro, Brazil

5.1 Abstract

Losses to follow-up are a major source of bias in cohorts of HIV-infected patients, and could lead to mortality underestimation. The authors developed a hierarchical deterministic linkage algorithm primarily to be used by cohorts of HIV-infected individuals to recover vital status information for patients lost to follow-up. Data from patients known to be deceased in two cohorts in Rio de Janeiro, Brazil and from the state mortality database between 1999 and 2006 were used to validate the algorithm. A fully automated procedure yielded a sensitivity of 92.9% and 100% specificity when no information was missing. When combined with clerical review, in a scenario of 5% death prevalence and 20% of missing mother's name, sensitivity reached 96.5%, and specificity, 100%. In a practical application, the algorithm significantly increased the death rates and decreased the rate of loss to follow-up in the cohorts.

The finding that 23.9% of matched records did not have AIDS as cause of death reinforces the need to search all-cause mortality databases and alerts for possible underestimation of death rates. These results indicate that the algorithm is accurate enough to recover vital status of patients lost to follow-up in cohort studies.

MeSH: cohort studies; data collection; HIV; mortality; software validation

5.2 Introduction

Database linkage is the process of comparing records from different databases that contain enough information to determine if these records refer to the same person, or more generally, to the same entity [77].

There are three main types of record linkage: manual (or clerical), deterministic, and probabilistic, which can be combined, depending on the strategy used. The first type consists of manually comparing records in two databases and deciding if they are true matches or not. This was the standard method used before the availability of computers, is often highly labor-intensive, and sometimes is not feasible, particularly when the amount of data is too large. Deterministic methods are classically based on exact match comparisons of either one unique identifier common to both databases (e.g. social security number) or a combination of variables that allow unique discrimination (e.g. name, surname, date of birth, gender) [78-80]. Probabilistic methods are also based on several variables, but comparisons are made based on the prior probability of two records belonging to the same entity and then calculating a maximum likelihood estimator to reach a score of similarity between records [77, 81, 82]. The method to be chosen (or combination of methods) depends on the type of analysis to be carried out on the linked data and the types of databases available [83].

Record linkage is widely used in population-based studies to make inferences about specific outcomes and in cohort studies to make inferences at

the level of the individual [79, 83-86]. Morbidity and mortality databases are often employed for this purpose, given their wide availability and the fact that in general they have enough information to link their records with other databases.

Cohort studies usually use linkage techniques to gather additional information about patients being followed over time. Even cohorts with active follow-up tend to complement their information with external databases in order to minimize under-reporting of conditions (e.g. vital status), including in their protocol a passive follow-up component. In the case of cohorts of human immunodeficiency virus (HIV)-infected patients, morbidity databases (e.g. tuberculosis, cancer) are also important sources of additional information [87-89].

In Brazil, official surveillance and mortality databases contain variables, such as full name, date of birth, and mother's name (either maiden or married last name), that are suitable for linkage procedures, due to their potentially high discriminatory power, particularly when used in combinations.

In the present study, we describe the validation of a new deterministic linkage algorithm that we developed to be used for passive data collection for cohorts of HIV-infected patients. The algorithm has a hierarchical structure and allows for specific errors in names and dates of birth. It can be used in combination with clerical review of records that are not classified as true matches or as non-matches. The main objectives when developing the algorithm were to maximize accuracy and to minimize the need for clerical review.

5.3 Materials and methods

5.3.1 Data sources

Three data sources were used in this study. The Rio de Janeiro cohort database was originally designed to validate the World Health Organization HIV staging system in a developing country [90] and currently comprises information from 2,666 HIV-infected patients being followed at the Clementino Fraga Filho University Hospital in Rio de Janeiro. All patients are aged 16 years or older, and are included only if they have at least one follow-up visit. The loss to follow-up rate between 2000 and 2005 in the Rio de Janeiro cohort was 2.9/100 person-years.

The TB-HIV in Rio study (THRio) is an ongoing cohort study designed to assess the impact of implementing isoniazid prophylactic therapy for HIV-positive patients with indication for prophylaxis in Rio de Janeiro. It has enrolled more than 15,000 patients from 29 clinics where care is provided both for HIV and for tuberculosis [91, 92]. There is still not enough follow-up time to calculate accurately the rates of losses to follow-up in this study.

The third database is the Rio de Janeiro State mortality database from 2000 through 2006 with a total of 835,066 records. The Rio de Janeiro State mortality database is part of the "Sistema de Informação sobre Mortalidade" database, which is the official mortality system in Brazil. The death certificate is a standardized form that is filled out by a physician, and includes demographic information and primary, secondary, and contributing causes of death according to the International Classification of Diseases, among other variables. An

electronic version of these forms was introduced country-wide in 1979. Information that can identify patients, such as name, mother's name, date of birth and address is also recorded and was made available through a special request to the State Health Department. According to the Brazilian Ministry of Health, the mortality system in Rio de Janeiro State has coverage of 100% of the deaths [93], even though the percentage of undefined causes of death remain somewhat high (9.3%) in 2005.

Data linkage between both cohort databases and the mortality database is part of routine procedures for assessment of vital status among patients who are lost to follow up and was approved by the institutional review boards of all involved institutions. Data from patients known to be dead through an independent source (medical charts in general) with identifying variables were used for validation purposes.

To validate the algorithm, test datasets were assembled and then linked with the mortality database and the outcome 'finding a record in the mortality database' was studied. To determine the sensitivity of the linkage to the mortality database for identifying deceased patients, all patients known to be dead and who had complete information on full name, date of birth and mother's name (either maiden or married last name) in the Rio de Janeiro cohort between 2000 and 2005 (53 patients) and in the THRio cohort between 2003 and 2006 (315 patients) were included in the analysis.

To assess specificity, we incorporated into the test database records that were not supposed to be in the State mortality database between 2000 and 2006 and that would be subject to similar typing mistakes as the patients known

to be dead. We chose to use a random sample of control records of patients who died in 1999, a year that was not included in the linkage.

The overall completeness of information for the THRio cohort is 98.3% for full information and 99.7% for name and date of birth. In the Rio de Janeiro cohort, 60% had full information and 100% had at least name and date of birth.

5.3.2 Data pre-processing

The first step in data linkage was pre-processing of data to guarantee that all variables conformed to the same format. For names, all letters were capitalized, and accents and characters other than letters were removed. Last names that denoted the same name such as the individual's father (e.g. 'Junior', 'Filho', etc) were also removed. A specific function (see supplementary data) was developed for this purpose and has the ability to pre-process a string field as a whole, either with Windows-based Latin alphabet encoding (cp1252) and DOS-based alphabet, still used in older ".dbf" files (cp850).

5.3.3 Linkage algorithm

To avoid an exponential growth of the processing time, records were first blocked by a phonetic code (PC) adapted from the original Soundex algorithm [94] to account for Brazilian Portuguese names (see supplementary data for details). Blocks were composed by the combination of either the PC from first and last name together, or the PC from first and last mother's name, or the PC of first name and first mother's name. This last category was employed to account for last names that are difficult to spell and that are recorded equally for

58

the individual and his/her mother in both databases, but are misspelled in one of them.

Records within each block were then compared, using exact comparisons and also allowing for some errors, in a hierarchical fashion, as described below. Errors in name fields were evaluated by the PCs and also by a string similarity score, based on a recursive longer common substring algorithm, implemented in the 'difflib' library from the programming language Python [95]. Dates of birth were allowed to have at most one digit mistake in any position or the common swap between day and month (only if they are exactly the same, but swapped). Score values used in the algorithm as described in Table 5.1 were chosen empirically in the beginning of the algorithm development, using different data sources [municipal surveillance databases of acquired immunodeficiency syndrome (AIDS) and Tuberculosis – data not shown]. The combinations of these measurements and the values for the scores determine several levels of inclusion – which in the present paper are referred to as 'automatic codes' and depend on how much information is available, as shown in Table 5.1. Records with complete information (automatic codes 0 through 7, Table 5.1) are treated independently from records with missing data (automatic codes 8 through 10 when mother's name is missing, Table 5.1). Whenever a pair of records is neither automatically included with one of the inclusion codes described nor automatically excluded with the criteria in Table 5.1, this pair is kept in the final merged database, marked as an unresolved pair for possible further clerical review. The algorithm is hierarchical in the sense that lower codes mean more similar records – 0 and 8 are perfect matches, but codes 0 through 7 are used for records with full information, and thus are more robust

than codes 8 through 10 for records that miss mother's name. The algorithm is not 'greedy' in that the same record in the test database linked with a 0 code to one record could also be linked to another one with a code 7, for example. This feature is important, because the algorithm can also be used for databases with one-to-many relationships, as is the case of tuberculosis surveillance databases. For mortality, which is supposed to have a one-to-one relationship with the cohort databases multiple matches for the same patient can easily be resolved by automatically picking the match with the lowest value, which was done in the present study.

If a pair was neither included nor excluded, it was eligible for clerical review. For records with name only, only perfect matches were considered.

The algorithm was written in Python for Windows [95].

5.3.4 Algorithm validation

We used three different scenarios to validate the algorithm. First we considered a hypothetical situation in which patients lost to follow-up in a cohort of HIV-infected patients would be searched in the mortality database and we assumed that 50% of these lost patients had actually died. Thus, we constructed a database by combining the 368 records of patients known to be deceased in the cohorts with a random sample of 368 records from the 1999 mortality database. In this scenario we compared the accuracy for perfect match between the records in all fields with the automatic inclusion codes, when: i) full information is available for all individuals; ii) only name and date of birth are available; and iii) only name is available. In the second scenario, we

tested the impact on accuracy if all patients in a cohort of HIV infected patients were linked to the mortality database, considering that only 5% of the patients were truly dead, a reasonable percentage for an open cohort of HIV-infected individuals in developing countries, where death rates are generally around 5 per 100 person-years [1]. In this case we made up a dataset with 368 records of patients known to be deceased in the cohorts and by randomly selecting 368/0.05 – 368 = 6,992 patients from the 1999 mortality database. In these two scenarios, ties were resolved automatically by choosing the pair with the lowest score and no manual search was performed – the aim being merely to assess the potential impact on accuracy of missing information in the test database.

In the third scenario, we mimic a situation similar to what one may encounter in practical research with a cohort of HIV-infected patients: assuming 50% prevalence of deaths among patients lost to follow-up and 20% of records missing mother's name. The database was set up for this scenario the same way as it was for the first scenario, but 20% of mother's names were randomly deleted from the test database. In this run, we did not consider records missing date of birth. Unresolved pairs were submitted to clerical review by two independent researchers, and disagreements were resolved by a third reviewer. Also, records with automatic inclusion codes were manually reviewed for quality control purposes. To minimize selection bias, reviewers had no access to the group membership status of records being reviewed.

Sensitivity, specificity, positive predictive (PPV) and negative (NPV) values were calculated for the experiments along with 95% confidence intervals (CI), using appropriate methods [96]. The total number of records in the test databases was used as the denominator for calculations.

For records of patients found in the mortality system, we assessed the proportion of cases where AIDS-related ICD-10 codes (B20-B24) were not mentioned on the death certificate. The coding system for death certificates follows the WHO guidelines [60].

For a preliminary practical application of the algorithm, rates of loss to follow-up are compared before and after the algorithm is used for the Rio de Janeiro cohort and death rates before and after the algorithm is used for both cohorts. Exact Poisson 95% confidence intervals (CI) are presented for the differences. Calculations were done in the R software environment [61].

5.4 Results

The first two scenarios to assess the impact of missing information on the accuracy of the algorithm are summarized in Table 5.2. As expected, sensitivity for exact matches increases when less information is available to link the records, while the addition of the automatic codes without manual review represented a significant increase both when full information is available (from 50.8% to 92.9%) and when mother's name is missing (71.2% to 91.8%). Specificity for both cases was very high and no misclassification was made by the algorithm when the death prevalence was 50% (PPV = 100%). In order to compare the impact of having 5% or 50% of prevalence, PPV and NPV were compared in these two scenarios. While PPV was 100% at 50% prevalence and no misclassifications occurred, it was reduced, as expected, at 5% prevalence.

62

Even though the percentages of automatic codes with full information and missing mother's name are still high (99.4% and 92.6%, respectively), these represented two false positive cases in the first instance, and 27 in the second. Accuracy for records with patients' names only was not as good as with the other variables; even when considered exact matches only (Table 5.2, last column), since sensitivities were lower than automatic codes for the other situations. Specificity in this case was very low, with PPV of only 81.2% in the 50% scenario, yielding 66 false-positive cases, and as low as 19.5% in the 5% scenario reaching 1,175 false-positive cases.

In reference of records to be manually checked, 1,189 of those with full information and 4,146 of those with missing mother's name would have to be searched for a prevalence of 50%, and 9,351 and 48,333 for a prevalence of 5%, respectively.

In the third scenario with 50% prevalence and 20% of the records missing mother's name, the results obtained were: sensitivity: 96.5%, 95%CI: 94.0, 98.1; specificity: 100%, 95%CI: 99, 100; positive predictive value: 100%, 95%CI: 99.0, 100; negative predictive value: 96.6%, 95%CI: 94.2, 98.2. Manual review was performed on the 1,929 pairs that the algorithm was not able to include or exclude as a true match. Of those, 9 pairs were considered true matches for reviewer #1 and 11 for reviewer #2. The two disagreements were submitted to a third reviewer, who considered one of them a true match. Manual review of the automatic codes considered all of them true matches.

The combination of automatic codes and clerical review yielded high sensitivity and specificity, with a PPV of 100% and NPV of 96.6% for this test database.

Among the 355 patients who were found by the algorithm, 85 (23.9%) did not have HIV-AIDS International Classification of Diseases 10th revision codes (B20 – B24) as the underlying cause of death.

Before the algorithm was used, the rate of loss to follow-up in the Rio de Janeiro cohort between 2000 and 2005 was 2.9/100 person-years and it dropped to 2.1/100 person-years after recovery from the mortality system (difference = -0.8, 95%CI: -1.1,-0.6). In the same period the mortality rate was raised with the deaths from the mortality system from 2.2/100 person-years to 3.2/100 person-years (difference = 1.0, 95%CI: 0.7,1.3). For THRio, the death rate in 2006 before the use of the algorithm was 1.2/100 person-years and increased to 4.2/100 person-years after deaths were recovered for all patients in the cohort, using automatic codes only, without manual review (difference = 3.0, 95%CI: 2.7,3.3).

5.5 Discussion

The deterministic algorithm validated in the present study was developed primarily to assist cohorts of HIV-infected patients with active follow-up to improve their performance by searching for patients lost to follow-up in mortality databases. The performance characteristics of the algorithm were excellent, with a sensitivity of over 90% for automatic codes, either in the 5% or in the 50% prevalence scenarios, which was minimally affected when mother's name was not available. These figures were well over the sensitivity for exact matches of 50% and 71% when full information was available and mother's name was missing, respectively. Specificity was close to 100% for all cases, meaning that

not a single pair was misclassified as a false-positive, but when considering records with patient's names only, even for exact matches specificity was unacceptably low (around 82%). These results are in agreement with the study by Quantin et al. [97], who found that date of birth and first and last patient's name would have enough discriminating power, even though their study was carried out using a probabilistic approach and they did not test mother's name as one of the variables. In the 5% scenario, the PPV remained close to 100% in all situations, and there were two false-positives in the full information dataset and an excess of 27 when mother's name was missing, indicating that, even though false-positives are very unlikely (PPV = 98.9% and 92.5%, respectively), caution must be taken when assigning a patient as being deceased.

In the third scenario, with 50% prevalence and having 20% of the records missing mother's name with clerical review increased sensitivity to over 96%, while preserving 100% specificity.

Although sensitivity was high, it was still impossible to find 13 patients reported as deceased in medical charts. Two situations may have concurred for this finding: i) patients indeed were not included in the mortality database or ii) patients were in the database but the algorithm was not able to find their records. In the former case, if the patient is in fact still alive he/she is truly lost to follow up. On the other hand, if the patient is indeed deceased, either the event was not detected by the system or the patient moved out of the state and died elsewhere. In the case of patients who are in the mortality database, the main reasons for not finding a record are major spelling errors – especially for the first letter, which is very sensitive to Soundex-like phonetic algorithms, but also

deletion of the last name in the case of individuals with four or more names –
and incorrectly entered dates of birth.

The amount of records left for manual review suggest that the best option
is to pre-select records to be linked to the databases in order to increase the
number of patients that are supposed to be found, decreasing clerical review.

Even though probabilistic algorithms have been extensively studied and
there is at least one algorithm validated for Brazilian databases [98, 99], we
chose to use a deterministic approach, allowing for some uncertainty on the
variables used. This decision was based on the fact that even though
probabilistic algorithms tend to yield higher sensitivities, they do so by
sacrificing specificity, which is not a major problem when studying population-
based characteristics, given false-positives and false-negatives would tend to
cancel out [83, 100]. Conversely, to infer the vital status of individual patients
being followed in a cohort, this approach is not advisable and deterministic
algorithms are more indicated [83], given that ethical problems may emerge
once his or her vital status is declared to be deceased and the patient shows up
for a subsequent visit. In either case, caution should always be exercised since
bias due to false-positives and negatives, would lead to over- or
underestimation of the parameters being studied, although the impact of false-
positives on overestimation tends to be more severe than of false-negatives on
underestimation [100].

One of the reasons that cohorts with active tracing of patients might
suffer from losses to follow-up is that information on deaths that occur in other
health care units might be out of reach for the clinic where routine care is
provided, especially if the cause of death was not related to AIDS. In cohorts of

patients who are intrinsically prone to high morbidity and/or mortality rates, as is the case of cohorts of HIV-infected patients, this can be particularly problematic. For example, in a study involving 6,498 patients being followed in 18 treatment programmes in lower income countries, the estimated death rate one year after initiation of antiretroviral therapy would increase from 6.4 to 15 per 100 person-years if mortality among loss to follow-up was similar to that observed in patient without antiretroviral therapy [1]. In another report from cohorts in Sub-Saharan countries, 41% of patients had unknown vital status on medical charts and 65% of those, initially considered to be lost to follow-up, were found to be dead after appropriate vital status investigation procedures were applied [101].

A practical application of the algorithm in both cohorts showed very good results. In THRio, applying the algorithm for all patients with the 2006 mortality database, there was a significant increase in the death rate for that year, even using automatic codes only. For the Rio de Janeiro cohort, the impact of the algorithm on patients lost to follow-up was significant both in reducing losses to follow-up and increasing death rate for the period of 2000-2005.

Finally, the fact that almost 24% of the death certificates of cases that were found through our algorithm did not have AIDS International Classification of Diseases 10th revision codes mentioned on them (B20-B24) underscores the need to search in all-mortality databases and not to restrict searches to HIV/AIDS deaths. On the other hand this finding suggests that official HIV/AIDS mortality figures that are based solely on the mortality system might significantly underestimate the true figures, a possibility that should be formally evaluated, which might lead to adjustments in mortality statistics in Brazil.

5.6 Tables

Table 5.1. Classification of Matched Records

Inclusion codes [a]	Patient's name	Date of birth	Mother's name [b]
0	Exact	Exact	Exact
1	Exact	Exact	Same PC
2	Exact	1 error or swap	Exact
3	Exact	1 error or swap	Same PC
4	Score>0.75	Exact	Exact
5	Score>0.75	1 error or swap	Exact
6	Score>0.75	Exact	Same PC + score>0.75
7	Score>0.9	1 error or swap	Score>0.8
8	Exact	Exact	Missing
9	Exact	1 error or swap	Missing
10	Score>0.9	Exact	Missing
	Not missing	> 1 error	Different PC
	Score<=0.9	> 1 error	Score<=0.8
	Not missing	> 1 error	Score<=0.7
Exclusion [c]	Score<0.8	Not missing	Not missing
	Not missing	Day, month and year are different	Missing
	Score<0.8		Missing

PC, phonetic code; [a] After passing the first blocking phase: same PC of first and last patient's name OR same PC of first and last mother's name OR same PC of patient and mother first

names; [b] PC is for mother's name only in this case; [c] Records that are not included nor excluded are left over for clerical review. Score values were chosen empirically (please see text).

Table 5.2. Summary Accuracy Measures and 95% Confidence Intervals for Exact Matches and Automatic Codes When Records in the Test Database Have Full or Partial Information, in Two Hypothetical Prevalence Scenarios: 50% and 5%

| Accuracy | Full Information | | | | No mother's name | | | | Name only [a] | |
| | Exact match [b] | | Automatic codes [c] | | Exact match [b] | | Automatic codes [c] | | Exact match [b] | |
	%	95% CI	%	95% CI	%	95% CI	%	95% CI	%	95% CI
Sensitivity	50.8	45.6, 56.0	92.9	88.3, 94.2	71.2	66.3, 75.8	91.8	88.6, 94.4	77.4	72.8, 81.6
Specificity	100.0	99.0, 100.0	100.0	99, 100.0	100.0	99.0, 100.0	100.0	99.0, 100.0	82.1	77.8, 85.8
PPV (50% prevalence)	100.0	98.0, 100.0	100.0	98.9, 100.0	100.0	98.6, 100.0	100.0	98.9, 100.0	81.2	76.7, 85.1
NPV (50% prevalence)	67.0	62.9, 70.9	93.4	90.5, 95.6	77.6	73.6, 81.3	92.5	89.4, 94.9	78.4	74.0, 82.4
PPV (5% prevalence)	100.0	98.0, 100.0	99.4	97.9, 99.9	98.9	96.7, 99.8	92.6	89.4, 95.0	19.5	17.5, 21.6
NPV (5% prevalence)	97.5	97.1, 97.8	99.6	99.5, 99.8	98.5	98.2, 98.8	99.6	99.4, 99.7	98.6	98.3, 98.9

CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value

[a] Since only name was available in this case, only exact matches were considered; [b] Exact match means perfect match between the available variables in both databases; [c] The codes listed in Table 5.1

6.0 CONCLUSIONS

Temporal changes of causes of death among HIV-infected individuals have been reported in developed countries in the HAART era. Several studies in developed countries have documented these changes, showing that, although HIV/AIDS-related causes of death remain the most common cause of death, other conditions, including diabetes mellitus, cardiovascular disease, cancer, liver and renal diseases have been increasingly reported in HIV-infected individuals [5-9]. In Brazil, where access to HAART has been universal for all patients who qualify since 1996, a steep decrease in mortality among HIV/AIDS patients has been documented, but the rates have been fairly stable between 1999 and 2004 [3, 4]. However, no data are available about the pattern of mortality in these patients. In this study, we investigated the profile of non-HIV-related causes of death in HIV-infected patients in Brazil, both in the population and individual level, resulting in three papers as described below.

In the first paper we used the Brazilian mortality database from 1999 to 2004 and showed that in Brazil, similar to what has been reported from developed countries, mortality patterns among HIV/AIDS patients are changing in the HAART era. We found that, in comparison to 1999, there was a steady and significantly larger increase in the frequency with which conditions not usually associated with HIV/AIDS were listed as causes of death for individuals who also had HIV listed on the death certificate

compared to individuals who did not, from 16.3% in 1999 to 24.1% in 2004, representing 14,746 deaths in the period. In particular, listing of cardiovascular diseases or diabetes mellitus as causes of death represented 3,746 and 744 deaths, respectively, both appearing to become more likely causes of death over time in individuals whose death certificate also included HIV/AIDS than in those who did not. Additionally, during the same period there were statistically significant increases in these conditions as underlying causes of death among individuals who had HIV/AIDS mentioned on their death certificates. To our knowledge, this is the first paper that documents this transition in Brazil.

In the second paper, we showed that overall mortality rates in a large cohort of HIV-infected patients in Rio de Janeiro, Brazil, remained fairly stable between 1999 and 2006, in accordance with what has been described in Brazil and the results of the first paper. However, when AIDS- and non-AIDS-related causes of death are examined separately, though, a different picture emerged, with a decrease in AIDS-related causes of death (from 2.43/100 py in 1997-1998 to 1.24/100 py in 2005-2006; p < 0.01) and an increase in non-AIDS-related causes of death. Although the temporal trend for non-AIDS-related causes of death did not reach statistical significance, by 2005-2006 non-AIDS-related causes of death were more frequent than AIDS-related causes of death (1.59/100 py and 1.24/100 py, respectively). There was also an increase of CVD as a cause of death (p = 0.04). In corroboration, when more restrictive criteria to classify causes of death were used, non-immunodeficiency related causes of death increase over time. Baseline age was associated with risk for AIDS-related causes of death, but the risk decreased over time. However, it had no significant effect on non-AIDS related

causes of death, indicating that, as patients age because of longer survival, they become at higher risk of dying from a non-AIDS related cause of death. These findings were in accordance with the recent description of HIV cohorts tending to get older [63], as is the case in our cohort, where the median age significantly increased from 33.25 years in 1997 to 39.0 years in 2006 (p < 0.01).

The results described in the second paper were facilitated by a linkage algorithm that was created for this project to recover vital status from patients who were lost to follow-up. As a result, the use of this algorithm increased the observed mortality rate from 2.0/100 py to 3.2/100 py (difference = 1.2; 95%CI = 0.9,1.4; p < 0.01) after patients considered lost to follow-up were matched with the mortality database. We developed and validated this hierarchical deterministic algorithm as part of this study to compare records between databases that do not have unique identifiers, using name, date of birth and mother's name as the linking variables. The validation of the algorithm was presented in a third paper. The results demonstrated a sensitivity over 95% and specificity of nearly 100% for the outcome of finding a record in the mortality database. A practical application of the algorithm was described in this paper both in the Rio de Janeiro cohort and in THRio (see Section 5.3.1 for description). In both cases there was a significant increase in death rates (difference = 1.0, 95%CI: 0.7,1.3 and difference = 3.0, 95%CI: 2.7,3.3, respectively). In the Rio de Janeiro cohort, between 2000 and 2005 the loss-to-follow-up rate dropped from 2.9/100 person-years to 2.1/100 person-years after recovery from the mortality system (difference = -0.8, 95%CI: -1.1,-0.6).

In conclusion, these are the first studies to examine changes in mortality patterns among HIV-infected individuals in the HAART era in Brazil. Conditions that potentially

may be prevented and/or treated, such as CVD and DM appear to have become more common causes of death over time among HIV-infected individuals than in the general population when death certificates data was studied. The major driving force behind the observed increase in deaths from non-AIDS related causes appears to be longer survival due to the substantial decrease in deaths due to AIDS-related causes.

These findings have major public health and programmatic implications for developing countries that are providing access to HAART. The clinical management of persons living with HIV/AIDS in Brazil needs to be expanded to include prevention, diagnosis, and treatment of chronic conditions such as CVD and DM. Successfully managing non-AIDS related conditions will require the integration of HIV/AIDS programs with other public health programs, such as smoking cessation and control of hypertension [57]. Finally, for other developing countries that are just starting to scale-up their programs, lessons learned from the Brazilian experience might help to prepare in advance to changes in morbidity and mortality patterns that will likely occur once HAART becomes widely available.

APPENDIX A. Linkage algorithm code (PYTHON)

```python
#-----------------------------------------------------------------------------
# Name:         supplemental.py
# Purpose:      Supplemental material for the manuscript 'Validation of a
#               hierarchical deterministic record linkage algorithm using
#               data from two different cohorts of HIV-infected individuals
#               and mortality databases in Brazil' from Pacheco, AG et al.
#               Presents functions, lists and blocks written in Python that
#               implement the algorithm validated in the manuscript.
#
#
# Author:       Antonio Guilherme Pacheco
#
# Created:      2008/02/19
# Copyright:    (c) 2008
# Licence:      GPL
#-----------------------------------------------------------------------------


# -*- coding: cp1252 -*-
import traceback
from difflib import *
from datetime import *


"""
This file contains a collection of functions and lists used in the linkage
algorrithm.
Actual implementation needs further programming, but it is readily extensible
in Python.
"""

"""
Part 1: Data pre-processing. Lists for conversion of non-ASCII characters
into ASCII characters
considering both cp1252 and cp850 coding pages.
"""

#cp1252
caracteres_cp1252={
u'Ç':'C',
u'Á':'A',
u'À':'A',
u'Ã':'A',
u'Ä':'A',
```

```
u'Â':'A',
u'É':'E',
u'È':'E',
u'Ë':'E',
u'Ê':'E',
u'Í':'I',
u'Ì':'I',
u'Ï':'I',
u'Ó':'O',
u'Ò':'O',
u'Õ':'O',
u'Ö':'O',
u'Ô':'O',
u'Ú':'U',
u'Ù':'U',
u'Ü':'U',
u'Ñ':'N',
u"'":'',
u'"':'',
u'`':'',
u'.':'',
u'=':'',
u';':'',
u'/':'',
u'\\':'',
u'+':'',
u'-':'',
u'(':'',
u')':'',
u'*':'',
u']':'',
u'[':'',
u'#':'',
u'~':'',
u':':'',
u'\x8a': 'ARIA',
u'\xc1': 'A',
u'\xc0': 'A',
u'\xc3': 'A',
u'\xc2': 'A',
u'\xc4': 'A',
u'\xc9': 'E',
u'\xc8': 'E',
u'\xcb': 'E',
u'\xca': 'E',
u'\xcd': 'I',
u'\xcc': 'I',
u'\xcf': 'I',
u'\xce': 'I',
u'\xd3': 'O',
u'\xd2': 'O',
u'\xd5': 'O',
u'\xd4': 'O',
u'\xd6': 'O',
u'\xd9': 'U',
u'\xdb': 'U',
u'\xda': 'U',
```

```python
u'\xb4':''}

#cp850
caracteres_cp850={
u'Ç':'C',
u'Á':'A',
u'À':'A',
u'Ã':'A',
u'Ä':'A',
u'Â':'A',
u'É':'E',
u'È':'E',
u'Ë':'E',
u'Ê':'E',
u'Í':'I',
u'Ì':'I',
u'Ï':'I',
u'Ó':'O',
u'Ò':'O',
u'Õ':'O',
u'Ö':'O',
u'Ô':'O',
u'Ú':'U',
u'Ù':'U',
u'Ü':'U',
u'Ñ':'N',
u"'":'',
u'"':'',
u'`':'',
u'.':'',
u'=':'',
u';':'',
u'/':'',
u'\\':'',
u'+':'',
u'-':'',
u'(':'',
u')':'',
u'*':'',
u']':'',
u'[':'',
u'#':'',
u'~':'',
u':':'',
u'!':'',
u'|':'',
u'\xd8': 'U',
u'\u0131': 'O',
u'\u2550': 'I',
u'\u0131': 'O',
u'\u2566': 'E',
u'\u250c': 'A',
u'\u2569': 'E',
u'\u0131': 'ARIA ',
u'\u2534': 'A',
u'\xe3': 'A',
u'\xe9': 'U',
```

```
u'\u2560': 'A',
u'\xe0': 'O',
u'\xef': '',
u'\u252c': 'O',
u'\u253c': 'O',
u'\u251c': 'O',
u'\xa4': '',
u'\u2514': 'O',
u'\u2554': 'U',
u'\xc5':'O',
u'\xf2':'A',
u'\xb4':'',
u'\xe5':'O',
u'\u2563':'A',
u'\u2524':'',
u'\u2557':'A',
u'\xa6':'ARIA ',
u'\xba':'ARIA ',
u'\xaa':'ARIA'}

def limpa_caracter(string, code='latin1'):
    """
    Turns non-ASCII characters into ASCII characters. Should be used
    BEFORE applying soundex algorithm
    """
    string1=''
    if string==None or string=='':
        return ''
    elif code=='latin1':
        string=unicode(string.upper(), 'latin1')
        for s in string:
            if s in caracteres_cp1252.keys():
                string1=string1+caracteres_cp1252[s]
            else:
                string1=string1+s
    elif code=='dos':
        string=unicode(string.upper(), 'cp850')
        for s in string:
            if s in caracteres_cp850.keys():
                string1=string1+caracteres_cp850[s]
            else:
                string1=string1+s
    return string1


"""
Part 2: Modified Soudex: Soundex algorithm modified for Brazilian Protuguese
spelling
"""

#Last names to be excluded:
nome_exclui=[
'JUNIOR'
'JR',
'FILHO',
'I',
'II',
```

```
'III',
'F']

#Modified Soundex coding system:
letra_cod={
u'A':'0',
u'E':'0',
u'H':'0',
u'I':'0',
u'O':'0',
u'U':'0',
u'Y':'0',
u'S':'0',
u'0':'0',
u'B':'1',
u'F':'1',
u'P':'1',
u'V':'1',
u'W':'1',
u'C':'2',
u'G':'2',
u'J':'2',
u'K':'2',
u'Q':'2',
u'S':'2',
u'X':'2',
u'Z':'2',
u'D':'3',
u'T':'3',
u'L':'4',
u'M':'5',
u'N':'5',
u'R':'6'}

#Conversion of (possible) non-ASCII characters (not used if pre-processed
names are used)
conv_simb={
u'Ç':'2',
u'Á':'0',
u'À':'0',
u'Ã':'0',
u'Ä':'0',
u'Â':'0',
u'É':'0',
u'È':'0',
u'Ë':'0',
u'Ê':'0',
u'Í':'0',
u'Ì':'0',
u'Ï':'0',
u'Ó':'0',
u'Ò':'0',
u'Õ':'0',
u'Ö':'0',
u'Ô':'0',
u'Ú':'0',
u'Ù':'0',
```

```
u'Ü':'0'}

#Exclusion of (possible) symbols and letter H (not really a consonant in
Portuguese)

simb_exclui=[
u'H',
u'',
u' ',
u',',
u"'",
u'"',
u'\xb4',
u'\xa4',
u'`',
u'-',
u'_',
u'\\',
u'0',
u'1',
u'2',
u'3',
u'4',
u'5',
u'6',
u'7',
u'8',
u'9',
u'#']

#Function to find first name in a string
def find_pnome(nome):
    """
    Finds first name
    """
    if ' ' not in nome:
        return '0000'
    else:
        pnome=nome[0]
        i=1
        while nome[i] != ' ':
            pnome=pnome+nome[i]
            i+=1
        return pnome.upper()


#Function to find last name in a string
def find_snome(nome):
    """
    Finds last name
    """
    if ' ' not in nome:
        return '0000'
    else:
        i=len(nome)-1
        snome=''
        while nome[i] != ' ':
```

```python
            snome=nome[i]+snome
            i-=1
        return snome.upper()


def nome_2_code(nome, tam=4):
    """
    Turns a single name (first or last) into modified soundex code
    """
    try:
        nome=unicode(nome.upper(),'cp1252')
    except TypeError:
        nome=nome.upper()
    if nome[0] not in simb_exclui:
        codigo=nome[0]
        i=1
    else:
        codigo=nome[1]
        i=2
    if codigo=='W':
        codigo='V'
    elif codigo=='K':
        codigo='C'
    elif codigo=='Y':
        codigo='I'
    elif codigo=='J' and (nome[i]=='I' or nome[i]=='E'):
        codigo='G'
    nome=nome[i:]
    for x in range(1,len(nome)):
        if nome[x-1] not in simb_exclui and nome[x-1]!=nome[x]:
            if nome[x-1] in conv_simb.keys():
                codigo+=conv_simb[nome[x-1]]
            else:
                codigo+=letra_cod[nome[x-1]]
    codigo=codigo.replace('0','')
    return (codigo + (tam*'0'))[:tam]


def acha_soundex(nome):
    """
    This binding function finds first and last names, get their modified
    soundex codes and combines them in a single string, e.g u'N420S410'.
    """
    if ' ' not in nome:
        return '00000000'
    else:
        pnome=find_pnome(nome)
        snome=find_snome(nome)
        if len(pnome)==1 or len(snome)==1:
            return '00000000'
        else:
            while nome_exclui.__contains__(snome) or
simb_exclui.__contains__(snome):
                nome=nome[:len(nome)-len(snome)-1]
                snome=find_snome(nome)
            return nome_2_code(pnome) + nome_2_code(snome)
```

82

```
"""
Part 3: Main functions of the algorithm (uses Part 2 functions as well)
"""


def cria_soundex(match):
    """
    Takes a list of lists in which mother's name is in the last position
    and patient's name in the third position from the last. Allows other
    variables to be in any other positions (in general second position from
    the last is date of birth, but no necessarily here). Returns a list of
    lists with 3 elements: soundex of first and last patient's name, soundex
    of first and last patient's mother's name and soundex of first names of
    both patient and mother, combined in a single string,
    e.g. [u'N420S410', u'L500S410', u'N420L500']. The resulting list can be
    used for blocking and is also used in the main function below.
    """
    recs=[]
    for m1 in match:
        try:
            recs.append([acha_soundex(m1[-3].strip()), acha_soundex(m1[-
1].strip()), acha_soundex(find_pnome(m1[-3].strip())+' '+find_pnome(m1[-
1].strip()))])
        except:
            recs.append([])
            print m1[-3].strip(), m1[-1].strip()
            print traceback.print_exc( )

    return recs


def monta_data(dia, mes, ano):
    """
    Auxiliary function to set up a date when comparing number by number
    """
    dia=str(dia)
    mes=str(mes)
    ano=str(ano)
    if ano=='9999' or ano=='None' or dia=='99' or mes=='99':
        return None
    if len(mes)==1:
        mes='0'+mes
    if len(dia)==1:
        dia='0'+dia
    return str(ano) +  str(mes) +  str(dia)


def compara_dob(dob1, dob2):
    """
    Compares two dates in datetime format and returns scores, depending on
    the similarity between them - 0 no match; 1 perfect match; 2 day/month
    swap; 3 at most 1 error in any position
    """
    if dob1==dob2:
        return 1
    else:
```

```python
        if dob1.year==dob2.year and dob1.month==dob2.day and
dob2.month==dob1.day:
            return 2
        else:
            score=0
            dob11=monta_data(dob1.day,dob1.month,dob1.year)
            dob21=monta_data(dob2.day,dob2.month,dob2.year)
            for i in range(len(dob11)):
                if dob11[i]==dob21[i]:
                    score=score+1
            if score>=7:
                return 3
            else:
                return 0


def cria_level(m):
    """
    Main algorithm function. Takes a single list of (possibly) pre-blocked
    records from 2 patients in the databases being linked and returns 3
pieces
    of information: the linkage level, the string comparison score for
patient's
    name and the string comparison score for patient's mother's name. The
list
    must be of length 12, being the first 6 from one patient and the last 6
from
    the one being compared. These 6 fields are a combination of: full name,
    date of birth, full mother's name and the list generated with the
    cria_soundex() function above.
    """
    limiar=0.8
    level=''
    sc=SequenceMatcher(lambda x: x == " ", m[0], m[6]).ratio()
    if m[2]!='' and m[8]!='':
        sc_mae=SequenceMatcher(lambda x: x == " ", m[2], m[8]).ratio()
    else:
        sc_mae=0
    if m[1]!=None and m[2]!='' and m[7]!=None and m[8]!='':
        if m[0]==m[6] and m[1]==m[7] and m[2]==m[8]:
            level=0
        elif m[0]==m[6] and m[1]==m[7] and m[4]==m[10]:
            level=1
        elif m[0]==m[6] and compara_dob(m[1],m[7])!=0 and m[2]==m[8]:
            level=2
        elif m[0]==m[6] and compara_dob(m[1],m[7])!=0 and m[4]==m[10]:
            level=3
        elif m[0]!=m[6] and m[1]==m[7] and m[2]==m[8] and sc>0.75:
            level=4
        elif m[0]!=m[6] and compara_dob(m[1],m[7])!=0 and m[2]==m[8] and
sc>0.75:
            level=5
        elif m[0]!=m[6] and m[1]==m[7] and m[4]==m[10] and sc>0.75 and
sc_mae>0.75:
            level=6
        elif compara_dob(m[1],m[7])==0 and m[4]!=m[10]:
            level=999
```

```
        elif compara_dob(m[1],m[7])!=0 and sc>0.9 and sc_mae>0.8:
            level=7
        elif compara_dob(m[1],m[7])==0 and sc<=0.9 and sc_mae<=0.8:
            level=999
        elif compara_dob(m[1],m[7])==0 and sc_mae<=0.7:
            level=999
        else:
            if sc<limiar:
                level=999
            else:
                level=91
    elif m[1]!=None and m[7]!=None:
        if m[0]==m[6] and m[1]==m[7]:
            level=8
        elif m[0]==m[6] and compara_dob(m[1],m[7])!=0:
            level=9
        elif m[1]==m[7] and sc>0.9:
            level=10
        elif compara_dob(m[1],m[7])==0 and m[1].year!=m[7].year and
m[1].day!=m[7].day and m[1].month!=m[7].month:
            level=999
        else:
            if sc<limiar:
                level=999
            else:
                level=92
    elif m[2]!='' and m[8]!='':
        if m[0]==m[6] and m[2]==m[8]:
            level=11
        elif m[0]==m[6] and m[4]==m[10]:
            level=12
        else:
            level=999
    else:
        if sc<limiar:
            level=999
        else:
            level=99
    return level, sc, sc_mae



"""
Part 4: Blocking strategy. If blocking is implemented to speed up
computational
time (as it was done in this algorithm) it is done in a separate block, that
call the above functions. In our case, the 'match1' and 'match2' objects are
the lists used by the cria_soundex() function and 'sound1' and 'sound2' its
output for records to be compared in database 1 and database 2, repectively,
as seen below.
"""

result=[]
for i in xrange(len(sound1)):
    for j in xrange(len(sound2)):
        try:
            #Blocking
```

```
            if sound1[i]!=[] and sound2[j]!=[] and
(sound1[i][0]==sound2[j][0] or ((sound1[i][1]!='00000000' and
sound2[j][1]!='00000000') and (sound1[i][1]==sound2[j][1] or
sound1[i][2]==sound2[j][2]))):
                #Setting up for cria_level()
                res=match1[i][-3:]+sound1[i]+match2[j][-3:]+sound2[j]
                try:
                    level, score, score_mae=cria_level(res)
                except:
                    print res
                #Discard unmatched pairs
                if level!=999:
                    #Setup a combination of variables for output
                    result.append(match1[i]+match2[j]+[str(level),
str(score), str(score_mae)])
        except:
            print sound1[i], sound2[j]
```

BIBLIOGRAPHY

1. Braitstein, P., et al., *Mortality of HIV-1-infected patients in the first year of antiretroviral therapy: comparison between low-income and high-income countries.* Lancet, 2006. **367**(9513): p. 817-24.

2. Palella, F.J., Jr., et al., *Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators.* N Engl J Med, 1998. **338**(13): p. 853-60.

3. Brito, A.M., E.A. Castilho, and C.L. Szwarcwald, *Regional patterns of the temporal evolution of the AIDS epidemic in Brazil following the introduction of antiretroviral therapy.* Braz J Infect Dis, 2005. **9**(1): p. 9-19.

4. Marins, J.R., et al., *Dramatic improvement in survival among adult Brazilian AIDS patients.* Aids, 2003. **17**(11): p. 1675-82.

5. Crum, N.F., et al., *Comparisons of causes of death and mortality rates among HIV-infected persons: analysis of the pre-, early, and late HAART (highly active antiretroviral therapy) eras.* J Acquir Immune Defic Syndr, 2006. **41**(2): p. 194-200.

6. Sackoff, J.E., et al., *Causes of death among persons with AIDS in the era of highly active antiretroviral therapy: New York City.* Ann Intern Med, 2006. **145**(6): p. 397-406.

7. Smit, C., et al., *Effective therapy has altered the spectrum of cause-specific mortality following HIV seroconversion.* Aids, 2006. **20**(5): p. 741-9.

8. Palella, F.J., Jr., et al., *Mortality in the Highly Active Antiretroviral Therapy Era: Changing Causes of Death and Disease in the HIV Outpatient Study.* J Acquir Immune Defic Syndr, 2006. **43**(1): p. 27-34.

9. Selik, R.M., R.H. Byers, Jr., and M.S. Dworkin, *Trends in diseases reported on U.S. death certificates that mentioned HIV infection, 1987-1999.* J Acquir Immune Defic Syndr, 2002. **29**(4): p. 378-87.

10. Lau, B., S.J. Gange, and R.D. Moore, *Risk of Non-AIDS-Related Mortality May Exceed Risk of AIDS-Related Mortality Among Individuals Enrolling Into Care With CD4+ Counts Greater Than 200 Cells/mm3.* J Acquir Immune Defic Syndr, 2006.

11. Bergersen, B.M., *Cardiovascular Risk in Patients with HIV Infection : Impact of Antiretroviral Therapy.* Drugs, 2006. **66**(15): p. 1971-87.

12. El-Sadr, W.M., et al., *CD4+ count-guided interruption of antiretroviral treatment.* N Engl J Med, 2006. **355**(22): p. 2283-96.

13. Mocroft, A., et al., *Is there evidence for an increase in the death rate from liver-related disease in patients with HIV?* Aids, 2005. **19**(18): p. 2117-25.

14. Qurishi, N., et al., *Effect of antiretroviral therapy on liver-related mortality in patients with HIV and hepatitis C virus coinfection.* Lancet, 2003. **362**(9397): p. 1708-13.

15. Carr, A., et al., *A syndrome of peripheral lipodystrophy, hyperlipidaemia and insulin resistance in patients receiving HIV protease inhibitors.* Aids, 1998. **12**(7): p. F51-8.

16. Friis-Moller, N., et al., *Cardiovascular disease risk factors in HIV patients--association with antiretroviral therapy. Results from the DAD study.* Aids, 2003. **17**(8): p. 1179-93.

17. Manfredi, R. and F. Chiodo, *Disorders of lipid metabolism in patients with HIV disease treated with antiretroviral agents: frequency, relationship with administered drugs, and role of hypolipidaemic therapy with bezafibrate.* J Infect, 2001. **42**(3): p. 181-8.

18. Riddler, S.A., et al., *Impact of HIV infection and HAART on serum lipids in men.* Jama, 2003. **289**(22): p. 2978-82.

19. Bergersen, B.M., et al., *Elevated Framingham risk score in HIV-positive patients on highly active antiretroviral therapy: results from a Norwegian study of 721 subjects.* Eur J Clin Microbiol Infect Dis, 2004. **23**(8): p. 625-30.

20. Bergersen, B.M., et al., *Important differences in components of the metabolic syndrome between HIV-patients with and without highly active antiretroviral therapy and healthy controls.* Scand J Infect Dis, 2006. **38**(8): p. 682-9.

21. Cheseaux, J.J., et al., *Hyperlipidemia in HIV-infected children treated with protease inhibitors: relevance for cardiovascular diseases.* J Acquir Immune Defic Syndr, 2002. **30**(3): p. 288-93.

22. Brown, T.T., et al., *Antiretroviral therapy and the prevalence and incidence of diabetes mellitus in the multicenter AIDS cohort study.* Arch Intern Med, 2005. **165**(10): p. 1179-84.

23. Sattler, F.R., et al., *Elevated blood pressure in subjects with lipodystrophy.* Aids, 2001. **15**(15): p. 2001-10.

24. Bozzette, S.A., et al., *Cardiovascular and cerebrovascular events in patients treated for human immunodeficiency virus infection.* N Engl J Med, 2003. **348**(8): p. 702-10.

25. Klein, D., et al., *Do protease inhibitors increase the risk for coronary heart disease in patients with HIV-1 infection?* J Acquir Immune Defic Syndr, 2002. **30**(5): p. 471-7.

26. Coplan, P.M., et al., *Incidence of myocardial infarction in randomized clinical trials of protease inhibitor-based antiretroviral therapy: an analysis of four different protease inhibitors.* AIDS Res Hum Retroviruses, 2003. **19**(6): p. 449-55.

27. d'Arminio, A., et al., *Cardio- and cerebrovascular events in HIV-infected persons.* Aids, 2004. **18**(13): p. 1811-7.

28. Friis-Moller, N., et al., *Combination antiretroviral therapy and the risk of myocardial infarction.* N Engl J Med, 2003. **349**(21): p. 1993-2003.

29. Mary-Krause, M., et al., *Increased risk of myocardial infarction with duration of protease inhibitor therapy in HIV-infected men.* Aids, 2003. **17**(17): p. 2479-86.

30. Currier, J.S., et al., *Coronary heart disease in HIV-infected individuals.* J Acquir Immune Defic Syndr, 2003. **33**(4): p. 506-12.

31. Barbaro, G., et al., *An open-label, prospective, observational study of the incidence of coronary artery disease in patients with HIV infection receiving highly active antiretroviral therapy.* Clin Ther, 2003. **25**(9): p. 2405-18.

32. Iloeje, U.H., et al., *Protease inhibitor exposure and increased risk of cardiovascular disease in HIV-infected patients.* HIV Med, 2005. **6**(1): p. 37-44.

33. Kwong, G.P., et al., *Comparison of the risks of atherosclerotic events versus death from other causes associated with antiretroviral use.* Aids, 2006. **20**(15): p. 1941-50.

34. Holmberg, S.D., et al., *Protease inhibitors and cardiovascular outcomes in patients with HIV-1.* Lancet, 2002. **360**(9347): p. 1747-8.

35. Department of Health and Human Services, *Panel on Antiretroviral Guidelines for Adult and Adolescents. Guidelines for the use of antiretroviral agents in HIV-infected adults and adolescents.* 2006.

36. Ministério da Saúde - Secretaria de Vigilância em Saúde - Programa Nacional de DST e Aids, *Recomendações para Terapia Anti-Retroviral em Adultos e Adolescentes Infectados pelo HIV, 2006.* 2006.

37. Hammer, S.M., et al., *Treatment for adult HIV infection: 2006 recommendations of the International AIDS Society--USA panel.* Jama, 2006. **296**(7): p. 827-43.

38. Walmsley, S., et al., *Lopinavir-ritonavir versus nelfinavir for the initial treatment of HIV infection.* N Engl J Med, 2002. **346**(26): p. 2039-46.

39. Jemsek, J.G., et al., *Body fat and other metabolic effects of atazanavir and efavirenz, each administered in combination with zidovudine plus lamivudine, in antiretroviral-naive HIV-infected patients.* Clin Infect Dis, 2006. **42**(2): p. 273-80.

40. Malan, N., et al., *Efficacy and safety of atazanavir-based therapy in antiretroviral naïve HIV-1 infected subjects, both with and without ritonavir: 96-week results from AI424089*, in *4th IAS Conference on HIV Pathogenesis, Treatment and Prevention*. 2007: Sydney, Australia. Abstract WEPEB024.

41. Horberg, M., et al., *Efficacy and safety of non-boosted to ritonavir-boosted atazanavir among antiretroviral naïve patients*, in *4th IAS Conference on HIV Pathogenesis, Treatment and Prevention*. 2007: Sydney, Australia. Abstract WEPEB025.

42. Mondy, K. and P. Tebas, *Cardiovascular risks of antiretroviral therapies.* Annu Rev Med, 2007. **58**: p. 141-55.

43. Lohse, N., et al., *Improved survival in HIV-infected persons: consequences and perspectives.* J Antimicrob Chemother, 2007.

44. Braithwaite, R.S., et al., *Estimating the proportion of patients infected with HIV who will die of comorbid diseases.* Am J Med, 2005. **118**(8): p. 890-8.

45. Lewden, C., et al., *Causes of death among human immunodeficiency virus (HIV)-infected adults in the era of potent antiretroviral therapy: emerging role of hepatitis and cancers, persistent role of AIDS.* Int J Epidemiol, 2005. **34**(1): p. 121-30.

46. Grangeiro, A., et al., *[Sustainability of Brazilian policy for access to antiretroviral drugs].* Rev Saude Publica, 2006. **40 Suppl**: p. 60-9.

47. WHO, *Progress on global access to HIV antiretroviral therapy: a report on "3 by 5" and beyond.* 2006: Geneva.

48. Ministério da Saúde - Datasus, *Mortalidade - Brasil.* 2007.

49. Centers for Disease Control and Prevention, *HIV/AIDS Surveillance Report, 2005. Vol. 17. Rev ed.* 2007, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention: Atlanta. p. 16-17.

50. Centers for Disease Control and Prevention, *HIV Mortality (through 2004).* 2007.

51. Tuboi, S.H., et al., *Predictors of virologic failure in HIV-1-infected patients starting highly active antiretroviral therapy in Porto Alegre, Brazil.* J Acquir Immune Defic Syndr, 2005. **40**(3): p. 324-8.

52. Friis-Moller, N., et al., *Class of antiretroviral drugs and the risk of myocardial infarction.* N Engl J Med, 2007. **356**(17): p. 1723-35.

53. Haraki, C.A.C., S.L.D. Gotileb, and R. Laurenti, *Accuracy of the Mortality Data System in a city in the south of the State of São Paulo, Brazil.* Rev Bras Epidemiol, 2005. **8**(1): p. 19-24.

54. Jorge, M.H.P.M., S.L.D. Gotileb, and R. Laurenti, *The national mortality information system: problems and proposals for solving them I - Deaths by natural causes.* Rev Bras Epidemiol, 2002. **5**(2): p. 197-211.

55. Johansson, L.A. and R. Westerling, *Comparing hospital discharge records with death certificates: can the differences be explained?* J Epidemiol Community Health, 2002. **56**(4): p. 301-8.

56. Miettinen, O.S. and J.D. Wang, *An alternative to the proportionate mortality ratio.* Am J Epidemiol, 1981. **114**(1): p. 144-8.

57. Schechter, M., *Treatment at scale in Brazil: a physician's perspective.* Aids, 2007. **21**(suppl 4): p. S31–S35.

58. Instituto Brasileiro de Geografia e Estatística, *Projeção da População.* 2006.

59. WHO, *Health statistics and health information systems.* 2007, WHO.

60. WHO, *International Statistical Classification of Diseases and Health Related Problems ICD-10, Volume 2: Instruction Manual.* 2nd ed. 2004. 184.

61. R Development Core Team, *R: A language and environment for statistical computing.* 2006, R Foundation for Statistical Computing: Vienna, Austria.

62. Yan, J. and J. Fine, *Estimating equations for association structures.* Stat Med, 2004. **23**(6): p. 859-74; discussion 875-7,879-80.

63. Casau, N.C., *Perspective on HIV infection and aging: emerging research on the horizon.* Clin Infect Dis, 2005. **41**(6): p. 855-63.

64. Pacheco, A.G., et al., *Increase in non-AIDS related conditions as causes of death among HIV-infected individuals in the HAART era in Brazil.* PLoS ONE, 2008. **3**(1): p. e1531.

65. Dabis, F., et al., *Cohort Profile: Antiretroviral Therapy in Lower Income Countries (ART-LINC): international collaboration of treatment cohorts.* Int J Epidemiol, 2005. **34**(5): p. 979-86.

66. Pacheco, A.G., et al., *Validation of a linkage algorithm to recover patients lost to follow-up in cohort studies: Results from two large HIV cohorts in Rio de Janeiro, Brazil*, in *4th Conference on HIV Pathogenesis, Treatment and Prevention.* 2007: Sydney, Australia. Abstract CDB019.

67. CoDe, *Coding Causes of Death in HIV.* 2006.

68. CDC, *From the Centers for Disease Control and prevention. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults.* Jama, 1993. **269**(4): p. 460.

69. Agresti, A., *Categorical data analysis.* 2nd ed. Wiley series in probability and statistics. 2002, New York: Wiley-Interscience. xv, 710 p.

70. Glynn, R.J. and B. Rosner, *Comparison of risk factors for the competing risks of coronary heart disease, stroke, and venous thromboembolism.* Am J Epidemiol, 2005. **162**(10): p. 975-82.

71. Therneau, T.M. and P.M. Grambsch, *Modeling survival data : extending the Cox model.* Statistics for biology and health. 2000, New York: Springer. xiii, 350 p.

72. Putter, H., M. Fiocco, and R.B. Geskus, *Tutorial in biostatistics: competing risks and multi-state models.* Stat Med, 2007. **26**(11): p. 2389-430.

73. Gaynor, J.J., et al., *On the Use of Cause-Specific Failure and Conditional Failure Probabilities: Examples From Clinical Oncology Data.* J Am Stat Assoc, 1993. **88**(422): p. 400-409.

74. Fine, J.P. and R.J. Gray, *A proportional hazards model for the subdistribution of a competing risk.* J Am Stat Assoc, 1999. **94**(446): p. 496-509.

75. Gray, R.J., *A class of K-sample tests for comparing the cumulative incidence of a competing risk.* Ann Stat, 1988. **16**(3): p. 1141-1154.

76. Gray, R.J., *cmprsk: Subdistribution Analysis of Competing Risks.* 2006.

77. Howe, G.R., *Use of computerized record linkage in cohort studies.* Epidemiol Rev, 1998. **20**(1): p. 112-21.

78. Li, B., et al., *Assessing record linkage between health care and Vital Statistics databases using deterministic methods.* BMC Health Serv Res, 2006. **6**: p. 48.

79. Gomatam, S., et al., *An empirical comparison of record linkage procedures.* Stat Med, 2002. **21**(10): p. 1485-96.

80. Clark, D.E. and D.R. Hahn, *Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry.* Proc Annu Symp Comput Appl Med Care, 1995: p. 397-401.

81. Fellegi, I.P. and A.B. Sunter, *A Theory for Record Linkage.* Journal of the American Statistical Association, 1969. **64**(328): p. 1183-1210.

82. Newcombe, H.B., et al., *Automatic linkage of vital records.* Science, 1959. **130**: p. 954-9.

83. Clark, D.E., *Practical introduction to record linkage for injury research.* Inj Prev, 2004. **10**(3): p. 186-91.

84. The West of Scotland Coronary Prevention Study Group, *Computerised record linkage: compared with traditional patient follow-up methods in clinical trials and illustrated in a prospective epidemiological study. .* J Clin Epidemiol, 1995. **48**(12): p. 1441-52.

85. Roos, L.L., et al., *Record linkage strategies, outpatient procedures, and administrative data.* Med Care, 1996. **34**(6): p. 570-82.

86. Whiteman, D., et al., *Reproductive factors, subfertility, and risk of neural tube defects: a case-control study based on the Oxford Record Linkage Study Register.* Am J Epidemiol, 2000. **152**(9): p. 823-8.

87. Ahmed, A.B., et al., *The growing impact of HIV infection on the epidemiology of tuberculosis in England and Wales: 1999 2003.* Thorax, 2007. **62**(8): p. 672-6.

88. Frisch, M., et al., *Association of cancer with AIDS-related immunosuppression in adults.* Jama, 2001. **285**(13): p. 1736-45.

89. Polesel, J., et al., *Non-Hodgkin lymphoma incidence in the Swiss HIV Cohort Study before and after highly active antiretroviral therapy.* Aids, 2008. **22**(2): p. 301-6.

90. Schechter, M., et al., *Predicting CD4 counts in HIV-infected Brazilian individuals: a model based on the World Health Organization staging system.* J Acquir Immune Defic Syndr, 1994. **7**(2): p. 163-8.

91. Golub, J.E., et al., *The impact of antiretroviral therapy and isoniazid preventive therapy on tuberculosis incidence in HIV-infected patients in Rio de Janeiro, Brazil.* AIDS, 2007. **21**(11): p. 1441-1448.

92. Moulton, L.H., et al., *Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention.* Clin Trials, 2007. **4**(2): p. 190-9.

93. Ministério da Saúde. Secretaria de Vigilância em Saúde, *Sistema Nacional de Vigilância em Saúde : relatório de situação : Rio de Janeiro 3. ed.* 2007, Ministério da Saúde: Brasilia. p. 24.

94. The National Archives, *The Soundex Indexing System.* 2007.

95. Lutz, M., *Programming Python.* 3rd ed. 2006, Sebastopol, CA: O'Reilly. xlii, 1551 p.

96. Zhou, X.-h., N.A. Obuchowski, and D.K. McClish, *Statistical methods in diagnostic medicine.* Wiley series in probability and statistics. 2002, New York: Wiley-Interscience. xv, 437 p.

97. Quantin, C., et al., *Which are the best identifiers for record linkage?* Med Inform Internet Med, 2004. **29**(3-4): p. 221-7.

98. Camargo, K.R., Jr. and C.M. Coeli, *[Reclink: an application for database linkage implementing the probabilistic record linkage method].* Cad Saude Publica, 2000. **16**(2): p. 439-47.

99. Coutinho, E.S. and C.M. Coeli, *[Accuracy of the probabilistic record linkage methodology to ascertain deaths in survival studies].* Cad Saude Publica, 2006. **22**(10): p. 2249-52.

100. Brenner, H., I. Schmidtmann, and C. Stegmaier, *Effects of record linkage errors on registry-based follow-up studies.* Stat Med, 1997. **16**(23): p. 2633-43.

101. Anglaret, X., et al., *Impact of vital status investigation procedures on estimates of survival in cohorts of HIV-infected patients from Sub-Saharan Africa.* J Acquir Immune Defic Syndr, 2004. **35**(3): p. 320-3.