HIGH STAKES TESTING: CHRONIC DISEASE MANAGEMENT IN LOW RESOURCE SETTINGS

JACOB BOR*
Boston University School of Public Health, USA
Africa Centre for Health and Population Studies, South Africa

TILL BÄRNIGHAUSEN
Harvard School of Public Health, USA
Africa Centre for Health and Population Studies, South Africa

April 10, 2015

ABSTRACT: Management of chronic diseases in resource-poor settings is based on standardized treatment guidelines. For example, HIV patients are eligible for antiretroviral therapy (ART) when their CD4+ (white blood) cell count is observed to have fallen below a threshold value. Little is known about the effect of such decision rules on patient behavior and health outcomes in low-resource settings. Using data on 4391 public sector HIV patients in rural South Africa, we estimate the effect of immediate vis-à-vis deferred ART eligibility on survival, immune recovery, and clinical retention. Treatment effects are estimated in a regression discontinuity framework using flexible parametric survival models, which are robust to unobserved heterogeneity, treatment effect heterogeneity, and time-varying effects of the treatment. Patients presenting for care with CD4+ counts just below 200 cells/μL were 4.8% points (95% CI 0.6, 9.0) more likely to be alive at three years compared to patients presenting with ineligible CD4+ counts just above the cut-off. Among patients whose treatment status was determined by the threshold rule, testing on the wrong side of the cut-off reduced three-year survival from 100% to 85%. Our data suggest that behavioral responses mediate the effect of treatment eligibility/ineligibility on outcomes and thus raise the stakes for diagnostic testing: patients who were ineligible for ART at baseline were much less likely to be retained in clinical care, leading to further delays in treatment and worse health outcomes. Immediate ART eligibility saved 0.18 years of life over five years at a cost of $1967 per year of life saved. *JEL Codes: I12, O15, C41*

* Department of Global Health and Center for Global Health and Development, 801 Massachusetts Ave, Boston MA, 02118, +1.617.414.1444, jbor@bu.edu. We thank the staff of the Africa Centre for Health and Population Studies and the study participants. For thoughtful feedback, we thank Marcella Alsan, Joshua Angrist, Jeremy Barofsky, David Canning, Gabriel Chodorow-Reich, Matthew Fox, Miguel Hernan, Brigham Frandsen, Willa Friedman, Guido Imbens, Bruce Larson, Zoe McLaren, Portia Mutevedzi, Kevi Naidu, Marie-Louise Newell, Deenan Pillay, Heidi Williams, and seminar participants at the International Workshop on HIV/AIDS Observational Databases, NEUDC, Atlantic Causal Inference Conference, University of Witwatersrand, University of North Carolina, Harvard School of Public Health, and Boston University. All errors and omissions are our own.

I. INTRODUCTION

Over the last 30 years, the world has undergone an epidemiological transition (Omran 1971): chronic diseases such as heart disease, diabetes, and cancer are now the leading causes of death and disability in low- and middle-income countries (Lozano et al. 2013). Unlike acute infections and injuries, chronic diseases require long-term clinical management. People at risk for chronic disease must undergo regular screening; once diagnosed, patients must be monitored routinely; and treatment regimens must prescribed and adapted iteratively based on patient response. Chronic disease management thus requires a long-term relationship between patient and provider, a new challenge for health systems that were set up to provide maternal services, vaccinations, and acute health care.

Standardized diagnostic and treatment guidelines greatly simplify chronic disease management. A common approach is the use of threshold rules on routinely monitored biomarkers to inform clinical decisions. E.g., patients with systolic blood pressure ≥ 140 mmHg are considered hypertensive (WHO/ISH 2003); patients with fasting blood glucose ≥ 7.0mmol/l are diabetic (WHO 2006); HIV patients with a CD4+ (white blood) cell count below a threshold are eligible for antiretroviral therapy (ART) (WHO 2003, 2010, 2013). Standardized guidelines simplify decision-making, enabling rapid diffusion of disease management technologies by lower-skilled cadres of health workers. Threshold rules also have benefits with respect to targeting, providing a transparent device to ration care to patients in greatest need. Additionally, evidence on "checklists" suggests that guidelines may yield superior outcomes to clinical judgment even in well-resourced settings (Haynes et al. 2009).

One benefit of threshold-based treatment guidelines is the ability to evaluate the real world impacts of medical interventions. Clinical and health policy decisions require evidence on the real world

effectiveness – and cost-effectiveness – of different therapeutic inputs (Drummond & McGuire 2001). However, causal evidence on the real-world effectiveness of specific medical interventions is scarce. Results of clinical trials often have not been replicated in real world settings (Weiss et al. 2008), and economists have been interested in natural experiments that identify the effectiveness of medical care as a health input (e.g., Doyle 2005; Almond et al. 2009; Card & Dobkin 2009; Finkelstein et al. 2012; Baicker et al. 2013). Threshold rules offer a unique opportunity for identification (Campbell & Thistlethwaite 1960). For example, Almond et al. (2009) evaluate the costs and survival benefits of neonatal intensive care provided to infants with "very low birth weight" in the United States. Zhao, Konishi, and Glewwe (2013) assess the effect of a hypertension diagnosis on health behaviors in China. To our knowledge, no previous study has evaluated the effect of clinical thresholds on survival in a low- or middle-income country, where threshold rules are an increasingly important part of patient care (Bor et al. 2014; Moscoe, Bor, Bärnighausen 2015).

In this paper, we ask: are the routine diagnostic and screening tests used for chronic disease management in low resource settings "high stakes"? Do test results at a particular point in time affect long run health outcomes? Health consequences of deferred eligibility reflect not only the potential biological impact of delayed treatment, but also possible effects of patient retention and provider rationing behaviors. Providers may overrule threshold decisions and grant patients eligibility if patients have other high-risk symptoms or high demand for treatment. Patients can mitigate the consequences of non-eligibility by returning to the clinic for regularly-scheduled screening tests or if they have additional symptoms. Patient retention in clinical monitoring is a *sine qua non* of chronic disease management, and yet, this particular health behavior has received scant attention in the economics literature (e.g., Dupas 2011). How do at-risk patients respond to being turned away from the clinic without treatment?

As a case study, we assess the effect of immediate vis-à-vis deferred treatment eligibility for HIV patients in South Africa's public sector ART program, the world's largest HIV treatment program. Like other chronic diseases, HIV requires lifelong clinical monitoring and management. People infected with HIV can live long and healthy lives if they start ART before they become very sick (Mills et al. 2011). Since the mid-2000s, low- and middle-income countries have provided highly-subsidized ART through public sector treatment programs, often with foreign assistance. Mass provision of ART has raised adult life expectancy by over a decade in some high prevalence populations in southern Africa (Bor et al. 2013). Still, HIV remains the leading cause of death in the region (Lozano et al. 2013) and there is substantial scope to improve screening and long-term management of HIV. In addition to improving survival, higher uptake of ART would have large social benefits due to reduced HIV transmission (Cohen et al. 2011; Friedman 2012; Tanser et al. 2013).

Under South African guidelines, once patients are diagnosed with HIV, the health of their immune system is monitored through regular CD4+ count blood tests. From 2004 through 2011, patients were eligible to start ART if their CD4+ count was less than 200 cells/μL or if they had a WHO Stage IV AIDS-defining illness – an opportunistic infection or cancer indicative of advanced HIV disease (NDOH 2004). Exploiting this threshold rule, we assess the causal effect of immediate vis-à-vis deferred ART eligibility on patient retention, immune health, and survival. Data come from a large health and demographic surveillance site in Umkhanyakude District, KwaZulu-Natal (Tanser et al. 2008). The area is poor and largely rural. Demographic data from a semi-annual household census were linked at the individual level with clinical records from the public sector ART program that serves the area. We follow patients from their first CD4 count in care and observe patient survival regardless of clinical retention.

There is reason to suspect that an eligible CD4 count might have little effect on health. First, the sickest

patients should be initiated because of clinical symptoms, regardless of CD4 count. Patients with Stage IV illness were eligible under national guidelines, and providers have discretion to ignore the guidelines if patient health is perceived to be at stake. Second, patients with non-eligible CD4 counts were instructed to return for follow-up CD4 monitoring at six months (Houlihan et al. 2010). Since the threshold only binds for patients presenting in reasonably good health, a six-month delay in eligibility would not be expected to have a large effect on survival for these patients. Third, to the extent that non-eligibility discourages some patients from returning for care, it may disproportionately discourage those patients who would have had poor outcomes had they initiated treatment (e.g., due to low adherence), and thus would have benefited little from treatment eligibility. It is plausible that pre-ART CD4 monitoring visits serve as an efficient "ordeal" (Nichols & Zeckhauser 1982).

On the other hand, if deferred eligibility leads to lower retention among patients who would have had successful experiences on therapy – and if providers do not identify and treat those patients – then the effects of non-eligibility may be large. Existing evidence suggests very low patient retention in pre-ART monitoring in the population we study: just 45% (Lessells et al. 2011) of patients not yet eligible for treatment are retained in care at one year. In contrast, 12-month retention after ART initiation is 85% (Mutevedzi et al, 2010). These are selected samples, however, and it is unknown whether these patterns reflect a causal effect of treatment eligibility on retention in care. Though providers have some discretion to overrule the threshold decision, they may lack the skill to improve treatment allocation or may choose not to; e.g., they may perceive the guidelines as a defense against accusations that they are rationing care unfairly.

To preview our results, first we find that diagnostic CD4 testing to determine HIV treatment eligibility is indeed "high stakes": we observe large reduced form effects on survival and immune health, implying very large effects for the approximately one third of patients for whom the threshold rule

binds. Among patients induced to initiate ART because they had an eligible baseline CD4 count (treated compliers), 100% were alive at 3 years; among patients barred from initiating ART because they had a non-eligible baseline CD4 count (control compliers), just 85% were alive at 3 years. Differences in survival and immune health persisted at 5 years.

Second, we find suggestive evidence that these effects are driven by behavioral – rather than biological – mechanisms. Baseline non-eligibility lowers 18-month retention in care from 100% to 63% among patients for whom the threshold binds. The divergence in survival curves occurs between 6 and 30 months, extending after the period when we would have expected the biological impacts of a six-month delay in therapy to manifest. And survival impacts are observed only among patients facing physical or psychological barriers to retention: those living far from clinics and living in households without another HIV patient. Together, these results suggest that initiating patients on therapy improves health by increasing patient retention in care. Among possible mechanisms, ART patients are required to attend counseling and education sessions which may provide information to patients about the importance of retention; ART patients are encouraged to disclose their status to household members, removing a barrier to future care-seeking; and the weekly, and then monthly visits associated with starting therapy may assist with habit formation. These non-pharmaceutical aspects of ART care may enable patients to overcome the barriers faced by patients who are not yet eligible for therapy. We also find evidence consistent with provider-rationing, in which providers have the skill to improve on the threshold-based treatment allocation, but do so preferentially for men and older patients, persons of social esteem in this community. Third, we establish the cost-effectiveness of immediate eligibility at the threshold studied, providing evidence to policy makers on the marginal returns to medical care. Over a five-year horizon, immediate eligibility saved 0.18 years of life at a cost-effectiveness ratio of $1967 per year of life saved.

Our paper proceeds as follows. Section II provides background on the history of clinical guidelines for ART treatment and a rationale for the study design. Section III describes the data sources. Section IV describes the empirical strategy and introduces our approach to analyze survival time data in an RD study. Main results are presented in Section V, with robustness checks in Section VI. Section VII investigates mechanisms through an assessment of treatment effect heterogeneity. Section VIII presents results on cost-effectiveness. Section IX concludes.

II. BACKGROUND

II.A. Clinical Guidelines and Existing Evidence on When to Start ART

HIV attacks the immune system, making HIV-infected people vulnerable to a wide variety of opportunistic infections and cancers usually avoided by people with health immune systems. In clinical settings across the world, ART traditionally has been allocated according to a simple decision rule: if the concentration of CD4+ white blood cells in a patient's blood – known as a "CD4 count" – falls below a threshold, then that patient is deemed eligible to initiate therapy. (Patients may also be initiated at higher CD4 counts due to clinical symptoms such as the presence of AIDS-defining opportunistic infections.) The question about "when to start" ART has largely been a question about the appropriate CD4 count threshold.

When ART first became available, early recommendations were to "hit HIV early and hard" (Ho, 1995). However, due to the harmful side effects of the earliest drugs, the perception that ART would be effective for a given patient only for a certain number of years, and the need to triage the sickest patients for immediate ART, initiation of therapy was commonly delayed until patients were quite sick. The first World Health Organization (WHO) guidelines for ART recommended initiating therapy when patients' CD4 counts had fallen below 200 cells/μL or when the patient was diagnosed with advanced clinical symptoms (Stage IV AIDS defining illness) (WHO 2002). In 2010, WHO amended these

guidelines, recommending initiation at CD4+ counts < 350 cells/μL or moderate-to-advanced HIV disease (Stage III or IV). In guidelines revised June 2013, WHO recommended initiating antiretroviral therapy for all HIV-infected people with CD4+ lymphocyte counts < 500 cells/μL. In spite of these changes, evidence on the clinical benefits to patients from immediate vis-à-vis deferred treatment is limited (WHO 2013). WHO itself cited only "moderate-quality" evidence regarding the clinical benefits of initiation at CD4+ counts above 200 cells/μL (WHO 2010, 2013 p95), which such evidence deriving from observational clinical cohort studies.

Existing experimental evidence comes from a single RCT in Haiti, which randomly assigned patients with CD4+ counts between 200 and 350 cells/μL to receive immediate ART or to wait until their CD4+ count fell below 200 cells/μL. Patients in the delayed treatment group had mortality rates four times higher than those receiving immediate therapy, and the study was terminated early (Severe et al. 2010). Two other RCTs found reductions in adverse clinical events, but were under-powered to detect differences in mortality (Emery, et al. 2008, Cohen et al. 2011, Grintzstejn et al. 2014). An ongoing multi-site RCT will assess outcomes for people initiating ART between 350 and 500 cells/μL, but includes very few participants from sub-Saharan Africa (NIAID 2009). No RCT has evaluated the effect of different CD4+ count thresholds on survival in sub-Saharan Africa, where the majority of ART patients reside, and where migration, clinical loss-to-follow-up, and specific burdens of opportunistic infections present challenges (De Cock & El-Sadr 2013).

Observational cohort studies have compared survival for patients initiating ART at different CD4 counts (Ford et al. 2010; Sterne et al. 2009; Kitahata et al. 2009). But these studies may be biased due to unobserved patient characteristics that are correlated with both survival and the timing of ART initiation. Further, these studies systematically exclude patients who presented for care but never initiated ART – perhaps because they were ineligible. Excluding non-initiators likely biases estimates

of causal effects towards the null since the sample of late initiators excludes patients who did not initiate by the end of follow-up. Further, this approach precludes analysis of the group most negatively affected by ineligibility for treatment – namely those who never make it back to initiate at a later date (Rosen & Fox 2011; Fox, Larson, Rosen 2012). Due to the limits of existing evidence, there have been recent calls for a randomized trial on "when to start" in sub-Saharan Africa (De Cock & El Sadr 2013). However, given current WHO guidelines, it would be difficult to argue for equipoise, the ethical requirement for an RCT.

II.B. Rational For Study Design

RD can be implemented when treatment assignment is determined by a threshold rule: patients are eligible if they are below (or above) some cut-off value on a continuously measured pre-treatment covariate. Random error in measurements of this assignment variable implies that patients with a true, underlying value close to the cut-off are quasi-randomized to being above or below the cut-off. Although treatment assignment is discontinuous at the threshold, continuity is guaranteed in all measured *and unmeasured* covariates so long as patients (or providers) cannot precisely manipulate the value of the assignment variable. Causal effects can be estimated by comparing outcomes immediately above vs. below the cut-off (Thistlethwaite & Campbell 1960; Imbens & Lemieux 2008; Lee & Lemieux 2010; Bor, et al. 2013).

We implemented an RD design using data on first CD4+ counts for patients presenting to a public sector HIV care and treatment program in rural South Africa. Previous studies have found very high within-subject variability in CD4+ counts (Hughes 1994), which we confirm for our sample. This random variability results from classical measurement error, from sampling variability in blood draws, and from random factors such as ambient temperature at the time of the blood draw. The imprecision of measured CD4 counts leads to a "strong" RD design with a "local randomization" interpretation at the

threshold (Lee 2008, Lee and Lemieux 2010, Bor et al. 2015): for patients with true CD4 counts close to the threshold, measurement error randomizes patients to have observed CD4 counts above or below the threshold. Opportunities for patient and provider manipulation of CD4 counts are slim: whereas height and weight are measured by clinicians in the context of providing patient care, diagnostics such as CD4 counts and blood lipid levels are analyzed by laboratory technicians (often off-site) and are less vulnerable to manipulation and/or heaping (Almond et al. 2010; Barreca et al. 2011; Shigeoka & Fushimi 2014). Finally, CD4 test results are revealed immediately prior to the determination of treatment eligibility, in contrast to other RD examples such as age (Card & Dobkin 2009) and distance from an administrative boundary (Chen et al. 2013), which may affect outcomes via pathways unrelated to treatment assignment.

Although patients are similar on either side of the threshold, we find that eligibility led to a large, discontinuous increase in the probability that a patient initiated ART within six months. Since patients are nearly identical within any small range of CD4+ counts, the causal effect of treatment eligibility can be estimated by comparing mortality rates among those presenting for care on either side of the CD4+ count initiation threshold. Under plausible assumptions, the causal effect of the treatment on those induced to take up by the threshold can be estimated by dividing the intent-to-treat estimate by the difference in the probability of rapid initiation at the threshold. In spite of the rigor of the design and many potential applications, few studies have evaluated the effects of clinical interventions using RD designs (see Moscoe et al. 2015 for a review, and Bor et al. 2014 for a recent example).

In addition to its substantive contributions, our paper makes a methodological contribution to the literature on regression discontinuity designs. RD designs have typically used linear models, even for non-continuous outcomes; in a departure that links RD designs more closely to the clinical literature, we show that regression discontinuity designs are amenable to non-linear and survival models, which

accommodate censoring. These models make more efficient use of the data, avoid biases that can result from ignoring censoring, and are more appropriate for modeling low-probability events than linear probability models. We generalize the RD design to model the time path of treatment effects using regression splines. And, we show how valid complier causal effects can be estimated in a fuzzy RD design using survival data. Specifically, we estimate the complier population survival curves using flexible parametric survival models, which are robust to unobserved heterogeneity in the underlying hazards, treatment effect heterogeneity, and time-varying effects of the treatment on any scale. This approach suggests a way forward for fuzzy RD studies of censored time-to-event data, e.g. in health and labor economics.

Preliminary analyses of these data were presented as a proof of concept in Bor et al. (2014). This paper substantially extends that analysis, presenting evidence on additional outcome measures and cost-effectiveness, using novel methods to model survival times in the context of fuzzy RD, and exploring behavioral mechanisms for the results.

III. DATA

III.A. Data Sources

Data were obtained from the Hlabsia HIV Treatment and Care Programme, the public sector ART program serving Hlabsia sub-district, in northern KwaZulu-Natal, South Africa (Houlihan et al. 2010). The Hlabisa program is decentralized, nurse-led, and is implemented through 17 clinics and one subdistrict hospital. The program follows South Africa's National Treatment Guidelines: From 2004-2010, patients with CD4+ counts under 200 cells/μL, or with Stage IV AIDS-defining illness were eligible for treatment. TB patients and pregnant women were eligible with CD4+ counts < 350 cells/μL. On 12 August 2011, the Ministry of Health announced updated treatment guidelines: all patients with CD4+ counts below 350 cells/μL would be eligible for treatment in the government ART

program (Bor et al. 2013).

Since its inception, the Hlabisa program has received technical assistance from the Africa Centre for Health and Population Studies (Africa Centre), a health and demographic surveillance site (HDSS) affiliated with the University of KwaZulu-Natal and funded by the Wellcome Trust. The Africa Centre has collected longitudinal demographic data since 2000 on a large population cohort residing in the Hlabisa health catchment area. The population cohort includes all members (resident and non-resident) of households residing in a 438 km^2 demographic surveillance area (DSA). The cohort is described extensively elsewhere (Tanser et al. 2008). In an agreement with the Department of Health, clinical records from patients in the Hlabisa HIV Care and Treatment Programme were matched to the Africa Centre's population surveillance data, with patients matched on national ID number, or full name, sex, and date of birth (Bor et al. 2010). Population-based surveillance data enables longitudinal follow-up of patient survival regardless of whether they are still in clinical care.

III.B. Study Population

The study population included all patients in the Hlabisa HIV Treatment and Care Programme who sought clinical care for HIV between 1 January 2007 and 11 August 2011, who were members of a household in the Africa Centre DSA at the time of their first CD4+ count in care, and whose first CD4+ count was less than 350 cells/μL. Pre-ART CD4+ counts were collected by the Africa Centre's database only after 1 January 2007; thus, patients who initiated ART or were reported to have their first CD4+ count prior to 1 January 2007 were excluded. Upon entry into the study all patients had yet to initiate ART in the Hlabisa program, although treatment naiveté could not be verified as some patients may have initiated therapy elsewhere. Women who were pregnant at the time of their first CD4+ count were excluded from the analysis. No sampling was conducted; all members of the study population were included in the analysis.

III.C. Treatment Assignment

Data on patients' CD4+ counts (number of cells/μL and date of CD4 test) were obtained upon enrollment into clinical care and at subsequent clinic visits. Patient CD4+ counts were assessed through a blood test, analyzed at an off-site laboratory, and reported directly by the lab to the Africa Centre's database. Patients entered the study on the date of their first CD4 count (blood test), taken upon testing positive for HIV in the public sector health system. At entry, some patients were very sick and eligible for ART based on clinical symptoms (Stage IV AIDS-defining illness). Most patients were instructed to return in a week for their CD4 results and at this subsequent visit, the decision to initiate was typically made. If the patient and provider decided to initiate ART, the patient was required to attend a series of weekly treatment literacy and adherence counseling sessions prior to initiating ART, except in cases of medical emergency when initiation was fast-tracked.

Dates of ART initiation were obtained from clinical records. We analyzed time from first CD4 count to date of initiation on a continuous time scale. We also created an indicator variable for rapid ART initiation, taking the value 1 if the patient initiated treatment within six months of her first CD4 count and zero if the patient still had not initiated treatment at six months. Based on standard of care, all patients that were assigned to initiate ART based on their initial CD4 count would have initiated within six months; and if six months had passed without initiating therapy, another CD4 count would have been taken to determine eligibility. After initiating ART, patients were scheduled for monthly visits to pick up medicines and to return every six months for laboratory tests (CD4 count and HIV viral load).

Patients who were determined not to be eligible for ART at their first CD4 count (CD4 of 200 or greater and no Stage IV illness) were referred to pre-ART care and monitoring, consisting of counseling and instructions to return in six months (and every six months thereafter) for another CD4

count to determine whether they had become eligible to initiate. Although CD4 counts are highly variable, they are – in expectation – monotonically decreasing for people with HIV not on treatment.

III.D. Outcome Measures

Vital status of study participants was ascertained through semi-annual household interviews conducted by Africa Centre staff. Household response rates in the demographic surveillance are very high (>99%) (Tanser, et al. 2008). Dates of death were recorded for all fatalities. Cause of death was determined by verbal autopsy, and deaths were categorized as HIV/TB-related or other (Herbst et al. 2011). Patients were followed up from the date of their first CD4+ count to their date of death, or the date when their vital status was last observed in the population surveillance.

The primary endpoint was time from first CD4+ count to death from any cause. As secondary endpoints, we assessed time to HIV/TB-related death and time to non-HIV/TB-related death. We also assessed trends in CD4+ counts, a measure of immunological health, as captured in routine clinical monitoring of patients retained in care. Finally, as a potential mediator, we assessed the effect of eligibility on retention in care in six-month intervals following first CD4 count. Both ART initiators and pre-ART patients are instructed to return every six months for a CD4 count (and viral load for patients on ART). According to guidelines, the only time a gap greater than six months should occur is at initiation of ART, when six months are counted from initiation rather than the last CD4 count. We defined retention in care as the presence of any laboratory test (CD4 or viral load) or ART initiation in six-month intervals (0-6, 6-12, 12-18, 18-24). We also defined retention at 6-18 months, giving patients the opportunity to be "retained" even if they were late for their lab tests, so long as they returned to the clinic for tests at least every 12 months.

Costs were calculated by estimating the expected number of "years on ART", "years in pre-ART care",

and "years not in care (deceased)" over a five-year horizon, and assigning clinic-based costs of $621, $104, and $0 respectively for per-patient-per-year costs of care in South Africa in 2011 (2011 US dollars) (Bor et al. 2013). As a conservative estimate of time in each state, patients were assumed to be in pre-ART care until their date of ART initiation and to be on ART until date of death or five years, whichever came first.

IV. EMPIRICAL APPROACH

IV.A. Empirical Approach

Regression discontinuity studies traditionally have used linear models even when modeling discrete-outcomes and rare events such as mortality (e.g., Almond et al. 2010; Card et al. 2009). However, linear models have some limitations. First, for binary, count, and survival data, linear regression models are less efficient than likelihood methods that correctly model the data-generating process. Second, when the underlying probability of the event is low (or high), effects of covariates may be approximately linear in Logits or Probits but non-linear in the expectation. In many RD applications, the treatment is assigned based on a continuous measure of risk, which is correlated with outcomes. The conditional expectation function thus may be nonlinear about the threshold, increasing the possibility that an RD study would falsely identify a treatment effect in finite samples. Third, when survival times are censored prior to the end of follow-up, models that accommodate censoring are required in order for these observations to be included – an important consideration when long run outcomes are of interest. The absence of RD applications with non-linear and survival outcomes may be a barrier to uptake in the clinical literature (Moscoe et al. 2014).

For all analyses, we compared predicted outcomes for patients presenting with CD4 counts just above vs. just below the 200-cell threshold. We modeled outcomes as follows: first, we assessed the effect of treatment eligibility on take-up, i.e., the probability of rapid ART initiation (within six months). Rapid

ART initiation was estimated on the risk difference scale, using linear probability (OLS) models to estimate Equation 1, which models the conditional expectation function (CEF) as a continuous function earliest CD4 count, except for an intercept shift at the threshold. We allowed for different slopes on either side of the threshold, which would arise in the case of effect heterogeneity. The intercept shift, β_2, is the effect of being CD4-count eligible on the probability of rapid ART initiation, for observations presenting with CD4 counts close to 200 cells. Similar linear probability models were estimated for our retention outcomes, which, like treatment initiation, was not a rare event. Models were estimated for different ranges (bandwidths) of CD4 counts on either side of the threshold, which is identical to a non-parametric, local linear regression with a rectangular kernel.

Equation 1

$$E\left[Y_i \mid CD4_i\right] = \beta_1(CD4_i - 200) + \beta_2 1\left[CD4_i < 200\right] + \beta_3(CD4_i - 200) * 1\left[CD4_i < 200\right]$$

For our health outcome variables, we extended this basic RD model in two novel directions. First, for our analysis of mortality (a rare event), we embedded the RD model in a generalized linear models framework, in which a continuous, possibly non-linear "link" function of the CEF is modeled as a linear function of predictors (Bor et al. 2014). Second, for both mortality and follow-up CD4 counts, we interacted the right hand side of the equation with parametric spline functions in follow-up time, to allow the effect of baseline treatment eligibility on outcomes to evolve flexibly over time. We estimated time-varying, generalized RD models of the form:

Equation 2

$$\begin{aligned} g\left(E\left[Y_i \mid t, CD4_i\right]\right) = & f\left(t \mid \mathbf{b}_0, k_0\right) \\ & + f\left(t \mid \mathbf{b_1}, k_1\right) * (CD4_i - 200) \\ & + f\left(t \mid \mathbf{b}_2, k_2\right) * 1\left[CD4_i < 200\right] \\ & + f\left(t \mid \mathbf{b}_3, k_3\right) * (CD4_i - 200) * 1\left[CD4_i < 200\right] \end{aligned}$$

where $f\left(t \mid \mathbf{b_j}, k_j\right)$ is a restricted cubic spline function of time (or log-time), with k_j knots with data-driven locations, and parameter vector $\mathbf{b_j}$ of length k_j. In all analyses, the number of knots for each of the interaction terms was identical, $k_1 = k_2 = k_3$, and less than or equal to the number of knots for the spline describing the "baseline" trend in outcomes, k_0. The spline function of time was interacted with each of the terms from the regression discontinuity model on the right-hand side of Equation 1. As with our prior models, we assessed robustness to a wide range of bandwidths of CD4 counts on either side of the threshold. The causal effect of treatment eligibility on the difference scale is equal to:

Equation 3

$$E\left[Y_i \mid t, CD4_i \uparrow 200\right] - E\left[Y_i \mid t, CD4_i \downarrow 200\right] = g^{-1}\left(f\left(t \mid \mathbf{b}_0, k_0\right) + f\left(t \mid \mathbf{b}_2, k_2\right)\right) - g^{-1}\left(f\left(t \mid \mathbf{b}_0, k_0\right)\right)$$

This effect may vary over time, e.g., as in the reduced form difference in survival curves. The ratio of means $E\left[Y_i \mid t, CD4_i \uparrow 200\right] / E\left[Y_i \mid t, CD4_i \downarrow 200\right]$ is also identified due to Slutsky's Theorem so long as the denominator is nonzero.

IV.B. Flexible Parametric Survival Models

For survival times – time to death, time to HIV-related death, and time to HIV-unrelated death – let $Y_i|t$ be an indicator for whether the event has still not occurred by time t, i.e., $Y_i = 1[T_i > t]$, where

$E[Y_i \mid t] = S(t)$, the survivorship function. We modeled survival probabilities using a complementary log-log link function, which implies a linear model for the log-integrated hazard, $\log\left(-\log\left[S(t \mid CD4_i)\right]\right) = \log\left(H(t \mid CD4_i\right)$. This model is the *flexible parametric survival model* (FPSM) developed by Royston and colleagues (Royston & Parmar 2002; Lambert & Royston 2009). The conditional survivorship function for a given CD4 count is obtained by inverting the link function as in Equation 3; and the time varying population hazard is obtained by taking derivatives of the survival function (Lambert & Royston 2009).

An alternative approach to modeling survival times would be to define binary indicators for survival to one year, survival to two years, etc., and estimate linear probability models (as in Almond et al. 2010). However, if some units are not followed up for the full interval, e.g. due to random censoring times, then this approach discards these censored observations. Worse than the loss of efficiency, if the event of interest is an absorbing state (e.g., death), then units that experience the event during follow-up are less likely to be censored and will be overrepresented in the data. (This is not an issue if follow-up is complete and censoring is an administrative end-of-study date, as in Almond et al. 2010; however, there are many applications in which censoring times are random, e.g. non-selective clinical attrition.) Like other survival methods based on the hazard, FPSM is designed to accommodate censoring, so long as it is non-informative (i.e., not correlated with failure times).

FPSM has the benefits of a fully parametric model: computation is quick and prediction is simple. However, with flexible functions for the baseline log-integrated hazard and time-varying effects of the treatment, FPSM has a distinctly non-parametric flavor. Similar to the Cox proportional hazards model, FPSM allows the analyst to be agnostic about how the population baseline hazard function varies over time. Importantly, however, the FPSM also allows for arbitrary non-proportionality over time in the

treated vs. control population hazards, which may arise due to frailty effects, heterogeneity in hazard ratios, or time-varying effects of the treatment. In conventional hazard models, choices of frailty distributions (e.g., gamma, inverse Gaussian) and functional assumptions for time-varying treatment effects (e.g. linear, piecewise constant) are often arbitrary and may lead to different results. This results from the fundamental non-identifiability of the underlying (structural) hazard model in the absence of arbitrary assumptions.[†]

Fortunately, the *population survival curve* is identified, and can be estimated consistently using the non-parametric Kaplan-Meier estimator for different covariate combinations (Kaplan & Meier, 1958). (Its scaled derivative, the population hazard curve is also identified, although the ratio and difference in population hazards are not causal parameters at t>0, since survivorship bias is built in to population hazard estimates (Lancaster 1979; Abbring & Van den Berg 2005).) Abbring & Van den Berg (2005) show that Kaplan-Meier estimates of the population survival curves for treatment eligible and non-eligible subjects can be plugged into the Wald estimator to obtain a time-varying LATE(t) parameter, the "complier difference-in-survival at time *t*".

A limitation for RD is that Kaplan-Meier cannot accommodate continuous covariates, and estimation relies on local linear regression predictions at the threshold. FPSM offers a flexible parametric alternative, using restricted cubic regression splines to describe the evolution of the regression discontinuity CEF over time. Spline functions are approximations to the underlying true functional

[†] Consider the individual-specific hazard model $h_i \mid t, D_i, V_i, W_i, \theta(t) = h_0(t) V_i \exp(\theta(t) W_i D_i)$, where V_i reflects heterogeneity in the baseline hazard, $h_0(t)$ reflects the average baseline hazard, which varies over time, W_i reflects heterogeneity in the proportional effect of the treatment D_i, and $\theta(t)$ describes how the treatment effect (average log hazard ratio) changes over time. In population data, time-varying hazards cannot be disentangled from frailty effects without untestable assumptions (Elbers & Ridder 1982; Heckman & Singer 1984); similarly, time-varying (proportional) treatment effects cannot be disentangled from (proportional) treatment effect heterogeneity without untestable assumptions. Non-identifiability suggests that a focus on the underlying structure of individual-specific hazard functions may be misplaced.

form; however, they can be arbitrarily good approximations with increased numbers of knots as the sample size grows. With finite knots, restricted cubic spline can fit all continuous functions subject to: i) linearity outside the outermost knots; ii) continuous first and second derivatives at the knots; and iii) number and placement of knots (Hastie & Tibshirani 1990). In practice, 3-5 knots placed at quantiles of the data are generally enough to describe most functions (Harrell 2001), and in particular, functions that are monotonically increasing (e.g., log cumulative hazard as a function of time) or decreasing (e.g., survival as a function of time) as shown in simulations (Lambert & Royston 2009). Further, regression splines can be expressed as simple transformations of the continuous predictor, such that they can be included in any regression model and inherit the consistency properties of that model.

In applying FPSM to RD, the non-parametric flavor of FPSM is further enhanced by the use of non-parametric local linear regression to model the relationship between CD4 count and log integrated hazards, i.e., by limiting the analysis to different bandwidths of CD4 counts around the 200-cell threshold. Thus, our analysis describes the "mortality risk surface" across CD4 counts and over time using flexible semi-parametric methods; the effect of interest is the time-varying gap in survival at the threshold. In our analysis, time since first CD4 count was modeled as a restricted cubic spline with four data-driven knots (at 0, 211, 653, and 1490 days). Findings were robust to different numbers of knots and knot location (not shown). The Stata command `stpm2` was used for all flexible parametric survival analysis (Lambert & Royston 2009).

We report population mortality hazards and cumulative survival probabilities at annual intervals, up to five years follow-up. We summarize the effect of treatment eligibility over a five-year horizon by calculating and comparing the expected years of life lost for observations just above vs. just below the threshold. As a point of comparison, we also present hazard ratios from more conventional exponential and Weibull regression models, which assume constant or monotonically-increasing (decreasing)

hazards over time and a proportional treatment effect. We also present models adjusting for Gamma (or inverse-Gaussian) distributed, individual-level, random frailty effects, in which case hazard ratios are interpreted as individual-level (rather than population-level) measures, i.e., conditional on the frailties. These results are presented for completeness, but since the proportional hazards assumption is violated in our data and because causal effects are difficult to interpret in terms of hazards, we emphasize the comparison of population survival curves as our main result.

IV.C. Modeling the Distribution of CD4 Counts

To model the effect of treatment eligibility on follow-up CD4 counts, we used OLS regression (where *g(.)* in Equation 2 is the identity link function) with the continuous variable "time since first CD4 count" modeled as a restricted cubic spline with four data-driven knots (at 0, 211, 653, and 1490 days). Unlike the mortality data, which were collected through population surveillance, follow-up CD4 counts were observed only if the patient was retained in care; thus, there is potential for bias from clinical attrition. To reduce the influence of selective attrition, we estimated a linear mixed effects models by maximum likelihood, which account for attrition under the assumption that missingness is random conditional on CD4 count history (Laird & Ware 1982; Verbeke and Molenberghs 2000; Molenberghs and Kenward 2007; Allison 2012). Specifically, we used the Stata command `xtmixed`, allowing for individual-specific random intercepts and random slopes for all terms in the spline (a so-called "growth curve" model). We compared the predicted mean CD4 count growth path for patients presenting just above vs. just below the threshold. In addition to mean CD4 counts, we also assessed the effect of treatment eligibility on the distribution of CD4 counts. We estimated RD models using quantile regression (Fransden et al. 2012), with the identical specification as the OLS model, and obtained predictions for the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentile CD4 counts at annual intervals.

IV.D. Treatment Effects Among Compliers

Thus far, we have focused on the intent-to-treat ITT_{RDD} effect of treatment eligibility (as determined by CD4 count) on outcomes. This effect is of interest to policy makers because it is the causal effect of treatment eligibility on the complete population, including those induced to take-up by the threshold and those who would have (or would not have) initiated ART regardless of CD4 count. Clinicians and patients may be interested in another causal effect: the effect of early vs. deferred ART initiation on patients that initiated ART rapidly because their CD4 count was below 200, i.e. compliers in the terminology of Imbens & Angrist (1994). Under the plausible assumptions that (a) rapid ART initiation is indeed the only pathway through which earlier ART eligibility would affect health (*exclusion restriction*), and that (b) no patient who would have been treated if ineligible would have rejected treatment if eligible and vice-versa (*monotonicity*), we can divide the intent-to-treat RD estimates by the "first stage" RD estimates (the Wald IV estimator) to obtain a complier average causal effect (CACE) or local average treatment effect (LATE). To avoid confusion, we use the terminology $CACE_{RDD}$ in lieu of "LATE" to specify that in an RD design, the complier treatment effect is in fact "local" across two dimensions: local to the population at the threshold and local to compliers. $CACE_{RDD}$ is denoted by the following equation:

Equation 4

$$\begin{aligned} CACE_{RDD} &= E[Y_i(1) - Y_i(0) \mid complier, CD4_i = 200] \\ &= \frac{\text{Intent-to-treat effect}}{\Pr(complier)}, \text{ by IV assumptions,} \\ &= \frac{E[Y_i \mid CD4_i \uparrow 200] - E[Y_i \mid CD4_i \downarrow 200]}{E[T_i \mid CD4_i \uparrow 200] - E[T_i \mid CD4_i \downarrow 200]}, \text{ by RDD assumptions.} \end{aligned}$$

We obtained $CACE_{RDD}$ estimates for the effect of early vs. deferred ART initiation on the probabilities of all-cause and HIV/TB mortality at annual intervals (on a risk difference scale), years of life lost over

a five-year horizon, and average CD4 count among survivors retained in care.

To further investigate outcomes among compliers, we estimated treated and control complier means for 3-year mortality and retention at 6-18 months. We restricted the analysis to patients surviving at 6 months and who would thus have the opportunity to have initiated ART. To calculate treated and control complier means, we estimated a separate set of RDD models, limiting the sample to patients who did not initiate ART. On the right side of the threshold, this model estimates outcomes for patients who were treatment-ineligible and who did not take up the treatment (a mixture of never-takers and compliers); on the left side, this model estimates outcomes for patients who were eligible, but nevertheless did not take up the treatment (never-takers). These estimates can be used to obtain mortality and retention probabilities for control compliers (Imbens & Rubin 1997; Abadie 2002),‡ and estimates for treated compliers can be obtained by adding $CACE_{RDD}$. (Equivalently, we could have estimated conditional-on-treated RDD models, calculated treated complier means, and subtracted off $CACE_{RDD}$.) As in the full sample, we estimated retention at 6-18 months in linear probability models and 3-year mortality probabilities using flexible parametric survival models.

V. RESULTS

V.A. Study Sample

The study sample included 4391 patients in the Hlabisa HIV Treatment and Care Programme, observed for a total of 13,139 person-years of follow-up. Of these patients, 3150 initiated ART and 820 died during follow-up. The majority of patients (69.2%) were women. The median age at first CD4+ count was 32.5 years, IQR = 26.3, 41.0.

‡ Treated Complier Average: $E[Y|T=1, complier] = \frac{E[Y*T|Z\uparrow c]-E[Y*T|Z\downarrow c]}{E[T|Z\uparrow c]-E[T|Z\downarrow c]}$

Control Complier Average: $E[Y|T=0, complier] = \frac{E[Y*(1-T)|Z\uparrow c]-E[Y*(1-T)|Z\downarrow c]}{E[(1-T)|Z\uparrow c]-E[(1-T)|Z\downarrow c]}$

V.B. Evidence for Validity of the Study Design

Causal inference using a regression discontinuity design is valid if the potential outcomes are continuous at the cut off. Support for this identifying assumption comes from three sources. First, there is a high degree of random noise in CD4+ counts in the study setting. We assessed the correlation between consecutive CD4+ counts among the 146 patients in our sample with repeat CD4+ counts on the same or consecutive days; regressing sqrt(FirstCD4) on sqrt(SecondCD4), the coefficient was close to one, but there was substantial unexplained variability. Our analysis implies that a patient with a "true, underlying" CD4+ count of 200 cells/μL would test within the 95% CI: 120 cells, 300 cells. Random noise in measured CD4+ counts implies that any factors correlated with "true" CD4+ counts will be continuous at the cut-off. It also implies that there is substantial overlap in "true" CD4+ counts among eligible and ineligible patients close to the threshold, such that the analysis does not depend on extrapolation across populations with different underlying immune health.

Second, the validity of the study design would be threatened if health workers or patients were able to manipulate patients' CD4+ count measurements, e.g., in an effort to access treatment earlier. Lab tests were conducted off-site and test results were reported from the lab directly to the Africa Centre database, leaving little opportunity for manipulation. Furthermore, we found no evidence of systematic manipulation in the data. Due to random noise in CD4+ count measurements, the distribution of CD4+ values should be continuous at the threshold; a discontinuity in the density function, with bunching just below the threshold, would suggest the presence of manipulation. Figure 1 displays the density of CD4+ counts upon enrollment in clinical care; there was no evidence of a discontinuity at the threshold ($p=0.79$)[§].

[§] We conducted a statistical test of continuity in the density function of earliest CD4 counts at 200 (McCrary 2008).

Third, support for the validity of the study design can be found by assessing continuity in baseline observables. Figure 2 displays scatter plots of age, sex, and date of first CD4 count against the value of a patient's first CD4 count. We also show mortality hazards predicted as a function of sex, age, age-squared, sex-by-age interactions, and date of first CD4+ count, a summary of the information contained in these covariates relevant to the outcome. This figure is similar to a balance table in an RCT. Random noise in measured CD4+ counts implies that there should be no discontinuity in pre-treatment characteristics, and indeed we found no evidence of systematic differences across the threshold.

V.C. Rapid Initiation of Antiretroviral Therapy

Figure 3 shows the cumulative probability of initiating ART within six months following a patient's first CD4+ count. Cumulative probabilities of initiation were estimated within 10-cell CD4+ count bins using the Kaplan-Meier estimator. (Supplementary Figure S1 shows cumulative probabilities of initiation for 1, 3, 6, 12, 24, and 36 months.) Patients who presented with CD4+ counts below 200 were much more likely than those with CD4+ counts above 200 to initiate ART within the first six months. In linear probability models (Table 1), having a first CD4+ count less than 200 increased the probability of initiation within six months nearly twofold – by 32 percentage points (95% CI 0.26, 0.38). This gap persisted two years later, though it decreased in magnitude as patients who originally presented above 200 went on to initiate therapy (Figure S1).

Specifically, we fit a kernel density function on either side of the threshold (bandwidth=25, rectangular kernel) with a renormalization boundary correction, rescaled so that each density function integrated to the probability of being below (above) the threshold, and calculated the difference in predicted densities at the threshold, which was bootstrapped (1000 replications) to obtain standard errors.

V.D. Treatment Eligibility and Survival: Reduced Form

We examined the effect of having a CD4+ count < 200 on mortality. We begin completely non-parametrically. Figure 4 shows Kaplan-Meier estimates of the cumulative probability of death at three years for 25-cell CD4 count ranges. In general, the higher a patient's CD4 count at baseline, the lower the probability of death. However, there is a discontinuity at 200 cells. Patients presenting with CD4 counts of 200-224 were *more* likely to die than patients presenting with CD4 counts of 175-199, in spite of having marginally better health at baseline.

To obtain predictions at the threshold, we need to put some parametric structure on the relationship between earliest CD4 count and mortality, which we do using FPSM. Figure 5 presents predicted probabilities of death at 1, 2, … and 5 years – the FPSM RD surface – estimated with linear terms on either side of the threshold and including patients presenting with CD4 counts between 50 and 350. Predictions at the threshold are presented in Table 2 and displayed in Figure 6.

In the first six months, survival declined sharply among both treatment-eligible and non-eligible patients, reflecting that some patients present at the clinic when they are already quite sick. Importantly, treatment eligibility did not appear to help these patients. At 6 months, there was no significant difference in the probability of death by treatment eligibility status (Table 2, panel 1: risk difference = 0.3% points, 95% CI -2.0, 2.5).

After 6 months, the survival curves for treatment eligible and ineligible patients diverged sharply. By 2 years after first CD4 count, a statistically significant 4.3% point gap (95% CI 0.6, 8.0) had emerged in the cumulative probability of death between patients who were treatment eligible (6.6%) and patients who were not treatment eligible (10.9%). A gap in survival between 4.0 and 4.8% points persisted between two and five years.

The divergence in survival experiences between six months and two years was driven by a sharp reduction in the hazard of death at 1 year among patients who were treatment eligible relative to those who were not eligible (Table 2, panel 2: HR at 1 year = 0.24, 95% CI 0.10, 0.60). Trends in the time-varying population mortality hazard among patients presenting on either side of the threshold are also presented in Figure 7. There is strong evidence for non-proportionality in the population hazards. The hazard of death was high in the six months after clinical presentation among both eligible and non-eligible patients. After this initial spike in mortality, the hazard of death among ART-eligible patients was approximately constant at about 2 deaths per 100 person-years. Among patients who were not eligible, however, there was substantial excess mortality between about six months and three years. By three years, the mortality hazard among patients who were not eligible had converged to the mortality hazard among patients who were eligible at baseline; some of this convergence in the population hazards may be due to frailty effects. The evolution of population hazard ratios and differences in survival over time are shown in Figures S4 and S5 with 95% CI.

The hazards reported here are descriptive and do not have a causal interpretation since they are estimated based on the population surviving at time t. However, contrasts of the survival curve do have a causal interpretation and are of interest. As a summary measure, we assessed the difference in the expectation of life (mean life years) over the five years of follow-up. Integrating between the survival curves, ART eligibility saved 0.18 years of life over a five-year horizon (95% CI 0.12, 0.26). This implies that a year of life was saved for every 5.6 patients who were eligible for treatment at baseline.

Using verbal autopsy data on cause of death, we were able to assess trends in HIV/TB-related vs. non-HIV/TB-related mortality. Table 3 presents results of separate flexible parametric survival models for HIV/TB-related mortality and non-HIV/TB-related mortality, censoring follow-up at alternate causes

of death. Patients who were not eligible for ART at baseline were 4.8% more likely have died by 2 years (95% CI 1.2, 8.4) than patients who were eligible for ART; these effects persisted at 5 years (Table 3, panel A). There were no differences in mortality due to other causes (Table 3, panel B).

V.E. Treatment Eligibility and Immune Health: Reduced Form

Treatment eligibility had a significant, positive effect on follow-up CD4 counts. Figure 8 displays measured CD4 counts at one year follow-up against baseline CD4 count. There is evidence of a discontinuity at the 200-cell eligibility threshold. The time-varying effect of treatment eligibility on follow-up CD4 counts was assessed in linear regression discontinuity models with a restricted cubic spline in time interacted with the usual RD covariates. For CD4 counts, linear functions across the range 100-300 cells were used. Mean CD4 counts increased over time for both eligible and non-eligible patients. However, CD4 counts increased much faster for ART-eligible patients, leading to an advantage in mean CD4 counts of 52 cells (95% CI 17, 87) at one year and 70 cells (95% CI 25, 115) at three years (Table 4, panel B). Large effects persisted at five years. Results were similar in linear mixed effects models, which are robust to missingness that is correlated with patient-specific CD4 count histories (Table 4, panel B). Trends in mean CD4 count among patients presenting at the threshold who were eligible vs. ineligible at baseline are presented in Figure 9 (linear regression) and Figure S8 (mixed effects). In addition to the effect of treatment eligibility on mean CD4 counts, we also estimated quantile treatment effects. Figure 10 displays the predicted cumulative densities of follow-up CD4 counts for eligible and ineligible patients presenting at the threshold. Baseline predictions placed the full density for both groups at 200 cells, by definition. Figure 10 shows the emergence of a gap in CD4 counts, evident across the full distribution. Over time, the distributions flatten as other sources of variability determine patient trajectories; however, the effect of baseline treatment eligibility persists across the full five years of follow-up. Further, the distribution of follow-up CD4 counts among eligible patients stochastically dominates the distribution among ineligibles.

V.F. Treatment Eligibility and Retention in Care: Reduced Form

Starting HIV treatment may have two conceptually distinct effects on health: first, the direct effect of the drugs on the patient's HIV viral load and immunological functioning in the current period; and second, the effect on retention in HIV care and treatment in future periods. Table 5 and Figure 11 display the reduced form impact of treatment eligibility on retention in six month intervals for two years following the patient's initial CD4 count. A large and persistent gap in retention is evident at the threshold. At 6-12 months, patients whose first CD4 count was ART-eligible were 11.8% points (s.e. = 3.9) more likely to be retained in care than patients who were ineligible at first CD4 count (Table 5, Panel B). Similar results were found when extending the retention window to twelve months: the difference at 6-18 months was 11.3% points (s.e. = 3.7). This gap persisted to two years, with ART-eligible patients 9.4% points (s.e. = 3.9) more likely to be retained at 18-24 months than patients who were ineligible, a 27.3% relative increase. The effects are large – particularly considering that only a third of eligible patients took up ART because they were eligible. It is unlikely that eligibility would have affected retention in care, except through actual ART initiation.

V.G. Clinical Benefits of Rapid Treatment Initiation: Complier Causal Effects

Under plausible assumptions (monotonicity and exclusion restriction) described above, we can assess the causal effect of rapid ART initiation (within six months) on survival, immune health, and retention among patients who initiated based on their CD4+ count. To estimate complier average treatment effects, we scaled the intent-to-treat difference in survival curves at the threshold by the difference in the probability of ART initiation, conditional on survival to six months (Wald estimator). Survival at six months was similar among patients eligible for treatment and among those who were ineligible (Table 2, Figure 6), providing support for our assumption that eligibility only affected survival through treatment itself. Table 6 presents RD results for the first stage (linear probability model) and intent-to-

treat effects (flexible parametric survival model). $CACE_{RDD}$ estimates were formed by dividing the intent-to-treat by the first stage.

The effect of treatment eligibility on rapid treatment initiation was 32.2% points at the threshold. The ITT effect on the cumulative risk of death at three years was -4.8% points. Dividing the two yields a $CACE_{RDD}$ of 14.9% points. This implies that persons who were induced to initiate ART because of an eligible CD4 count were about 15 percentage points more likely to be alive three years later than persons who deferred ART initiation because of an ineligible CD4 count. We calculated $CACE_{RDD}$ for the probability of death and mean CD4 count at annual intervals, as well as total years of life lost over the five years of follow-up. The $CACE_{RDD}$ for survival was in the range of 10-15% points from years one through five. Over this time, patients who initiated ART because they had an eligible CD4 count enjoyed an additional 0.59 years of life, relative to patients who were prevented from initiating because they were ineligible.

Among patients who survived to six months, baseline treatment eligibility had a large, significant, level effect on follow-up CD4 counts: eligible patients had 72 additional CD4 cells/μL at 1 year, and this gap persisted to five years. Dividing by the first stage, patients who initiated ART because they had an eligible CD4 count had about 225 extra CD4 cells/μL. The $CACE_{RDD}$ retention estimate was also large, with rapid ART initiation increasing retention at 6-18 months by 40.1% points (Table 6).

V.H. Investigating Outcomes Among Treated and Control Compliers

In our application, and in chronic disease management generally, the "complier" population affected by the threshold rule is a particular population. This population excludes persons who are so sick that they would be initiated on treatment regardless of their CD4 count (always-takers); and it excludes persons who do not want to initiate treatment, e.g. because they feel relatively healthy, and would not start ART

even if eligible. Compliers are thus those patients who want to initiate treatment but who are not so sick that treatment is an urgent clinical necessity.

To further investigate outcomes among compliers, we estimated treated and control complier means for survival and retention. Table 7 displays our estimated treatment and control complier probabilities of retention at 6-18 months and mortality at 3 years. Among patients prevented from initiating ART because their CD4 count was not eligible, only 63.1% were retained in care at 6-18 months; and 15.6% had died by 3 years. In contrast, patients induced to start ART because they had an eligible CD4 count had approximately perfect retention at 6-18 months (100%) and no mortality at 3 years (0%). Conditional on survival at six months, the complier relative risks of death and attrition were each approximately zero.

Retention and survival were higher among treated compliers than among always-takers, who included patients initiated because they were very sick and may or may not have had a strong preference to initiate ART (Table 7). Retention and survival were also higher among control compliers than among never-takers, who may have included patients that were relatively healthy and uninterested in initiating ART. These patterns are consistent with the joint patient and provider decision-making process that gave rise to this complier population. By definition, *ex ante*, compliers were willing to initiate therapy (and would have if eligible); as our results illustrate, this willingness may have translated into very high retention in care for treated compliers. Similarly, *ex ante*, compliers were not so sick that they would have been initiated on ART regardless of CD4 count; as revealed in the data, this group performed particularly well on treatment, with 100% survival at 3 years. Isolating the complier population illuminates just how high the stakes were for that first CD4 test. Among patients for whom the threshold rule bound, an ineligible CD4 count at baseline led to large clinical attrition and loss of life; these losses were completely avoided for patients with a CD4 count just below the eligibility threshold.

A critical question is why control complier patients performed so poorly relative to never-takers. One possibility is that never-takers were substantially healthier (conditional on CD4 count) at baseline. The difference in outcomes between never-takers and control compliers would suggest that patients have important information about their health that is not observable to clinicians and that outcomes might be improved if providers placed greater stake in patient preferences about whether to start treatment. However, even among compliers, it is a puzzle that this population could do so well if eligible and so poorly if not eligible. We turn to this question in the section on mechanisms, below.

VI. SENSITIVITY ANALYSES

To assess the robustness of our survival results, we estimated FPSMs varying the bandwidth and functional form of earliest CD4 count. The bandwidth was adjusted by reducing the window of data used in the analysis down to +/- 50 cells/μL. This approach coincides with local linear regression with a rectangular kernel. The functional form of earliest CD4 count was varied by including higher order polynomial terms in the model, up to fourth-order (cuartic) terms; these were each interacted with the spline functions in log-time. Results are presented in Tables S1 and S2. Results were consistent with the main results reported in Table 2, though at smaller bandwidths and with higher order polynomials there is some loss of precision.

We also estimated conventional hazard regression models, presented in Table S3. (Given our previous statements about time-varying hazards and time-varying treatment effects, these models make overly restrictive assumptions; however, they provide a point of comparison for other studies that make similar assumptions.) Column (a) presents results for exponential hazard models; column (b) presents Weibull hazard models; similar results are observed for Cox proportional hazard models (not shown).

In models estimated for patients with CD4 counts of 50 – 350, and including linear terms on either side, ART eligibility reduced the hazard of death by about one third (exponential HR: 0.65, 95% CI 0.45, 0.94; Weibull HR: 0.67, 95% CI 0.46, 0.96; Table S3, row 2), a result reported in Bor et al. (2014). Results were robust to smaller bandwidths. If there is unobserved heterogeneity in individual-specific hazards, then the population hazard ratio will under-estimate the individual-specific hazard ratio. Adjusting for random frailty effects – the so-called mixed proportional hazards model – ART eligibility reduced the hazard of death by over 50 percent (exponential with gamma frailties, HR: 0.45, 95%CI 0.24, 0.84; Weibull with gamma frailties, HR: 0.41, 95% CI 0.19, 0.85; Table S3, row 8). The effect of ART eligibility on all-cause mortality was driven entirely by its effect on the hazard of HIV/TB-related mortality (Table S3, rows 10-14).

VII. MECHANISMS

In the setting described, the CD4 tests to determine ART eligibility appear to be "high stakes", with large implications for patient health and survival. Focusing on the complier population for whom the threshold binds, a puzzle remains: how could these patients do so well if eligible and so poorly if not eligible? The poor outcomes among non-eligible compliers are particularly surprising given that this population was relatively healthy at baseline (ie, not eligible to initiate ART on the basis of other clinical symptoms).

Theoretically, initiating ART could have two effects on future health status. First, a direct effect of being on therapy in the current period on future health; and secondly, an indirect effect through increased retention in care (and a higher probability of being on ART) in future periods. The clinical literature has emphasized the direct health benefits. Observational studies have excluded patients who did not initiate therapy, simply comparing patients who initiated ART at different CD4 counts (e.g. Fox

et al. 2010). Existing clinical RCTs followed up patients in the delayed therapy arm with monthly CD4 count monitoring and intervened to reduce clinical attrition (e.g., Severe et al. 2010). Little is known about patients' behavioral responses to non-eligibility, which may drive a wedge between treatment efficacy and effectiveness in a real world setting.

The results above indicate a very large behavioral response to non-eligibility. Among compliers, treatment eligibility reduced 6-18 month retention in care from 100% to 63.1%. This, in spite of the fact that both eligible and non-eligible patients were instructed (as per national guidelines) to return for CD4 or viral load monitoring in six months. To our knowledge, these are the first causal estimates of the effect of eligibility on retention in HIV care.

Which of these two pathways – medical vs. behavioral – predominates in this setting? The timing of the mortality losses suggests that patient behavior mediates the effect of non-eligibility on survival. If the therapeutic benefits of immediate ART were the primary explanation, then we would expect substantial divergence in survival curves within the first six months, during the period when ART-eligible patients were getting onto therapy and non-eligible patients were not. Instead, as reported above, there was no difference in survival at six months. If non-eligible patients returned to the clinic for their scheduled six-month CD4 count, then many of these patients (and certainly the sickest) would have only a six-month delay in ART initiation, so most excess mortality in the baseline non-eligible group should be observed within the first 0-12 months. In fact, the divergence in survival occurs between 6 and 24 months, long after the baseline non-eligible patients would have initiated ART under perfect retention. Thus, non-retention likely mediates the effect of non-eligibility on survival.

We also find support for the retention hypothesis by comparing treatment effects for patients facing different barriers to accessing care. Dividing the sample by distance to clinic (<2km vs. >2km), we find

larger effects of eligibility on survival for patients living further from clinics, where distance may present a barrier to returning for future care (Table 8). Although we cannot reject that the effects are the same, the evidence is suggestive that treatment effects were heterogeneous by distance to the clinic. We also observe larger effects among patients without another HIV patient in the household. When there is another household member on ART or in pre-ART care, baseline treatment eligibility has no effect on patient survival, plausibly because the patient has a support network to remind her to seek care. It was the combination of non-eligibility plus the absence of a household member in HIV care and treatment that led to elevated mortality risk (Table 8).

In unpacking the behavioral pathway, the effect of non-eligibility on outcomes may be a product of three conceptually distinct factors. First, financial, time, or psychological barriers to accessing care may make returning for future clinic visits costly. If patients perceive that pre-ART CD4 monitoring visits are less important than post-ART visits, then such barriers may reduce retention relatively more among non-eligible patients. Heterogeneity in baseline patient characteristics, e.g. distance to a clinic, another HIV patient in the household, provides leverage to assess this pathway as described above.

Second, there may be differences in the clinical care provided if eligible vis-à-vis if not eligible that signal to patients the importance of returning for future visits and assist them in doing so. Patients may interpret the message "not *yet* eligible" as "not *ever* eligible"; they may lose faith in the health system; or in ART itself. Efforts to limit adverse patient learning may be minimal. Patients who are eligible for ART undergo extensive counseling; for patients who are not eligible, counseling is brief if it happens at all. Patients who go on to initiate ART are encouraged to disclose to household members and to identify a treatment "supporter" to help them remember their appointments. They must return to the clinic for weekly and then monthly visits, potentially aiding in habit formation and increasing the probability of attending their six-month visit when the next lab tests are taken. Additionally, substantial

investments have been made in tracing ART patients who have defaulted on their medicines and bringing them back into care. Indeed, although all individuals with a first CD4 count are presenting to the clinic with a life-threatening chronic illness requiring long-term management, many clinics do not even consider individuals to be active patients belonging to that clinic until they are deemed eligible for ART and started on the path towards ART initiation. The responsibility for retention in pre-ART care rests largely on the shoulders of patients. The high retention rate for treated compliers demonstrates that improving retention is possible with existing technologies that target key barriers to accessing care.

Third, dynamics of the patient-provider encounter may shape who the complier population is, and thus the magnitude of reduced form effects of the threshold rule. For example, providers may show favoritism to some patients who were particularly eager to initiate ART by initiating them regardless of eligibility. Such provider behavior would implicitly reclassify those would-be compliers as always-takers, protecting them from the threshold rule, and thus removing them from our natural experiment. In rural KwaZulu-Natal, older patients and men are afforded high esteem. As shown in Table 8, the effect of treatment eligibility on survival was essentially zero for men, and the difference in effects for men and women was statistically significant. This pattern is consistent with a story in which, once the high-performing male patients were removed from the complier population, the remaining compliers did quite poorly on treatment. (Further analysis reveals mortality rates three times higher among treated compliers than among always takers (not shown).) The reduced form effect was zero for older patients as well although due to small sample sizes the estimates are imprecise (Table 8). These results are somewhat speculative, but suggest that not only were CD4 tests “high stakes”, but that providers had substantial discretion about when to protect patients from the consequences of a non-eligible test result.

VIII. COST-EFFECTIVENESS

Studies that obtain causal estimates for both costs and health benefits over a lengthy (5-year) horizon are rare. To estimate costs, we estimated the excess person-years on ART (and/or in pre-ART care) experienced by patients presenting just below the 200-cells/μL threshold. We estimated FPSM models similar to our survival models, but with time to treatment initiation as the outcome, and modeled mortality as a competing risk. We then predicted the time-varying probability of being on ART at each point over a five-year horizon for patients on either side of the threshold. We made the conservative assumption that once a patient initiated ART, they would continue to be on ART for the duration of follow-up. Over five years, patients presenting just below the threshold spent a total of 0.57 more years on ART than patients who presented just above the threshold. We used published estimates for the cost of ART per patient per year in South Africa, which was $621 in 2011 (Bor et al. 2013). Combining our cost estimate with our survival estimates, we calculate that immediate vis-à-vis deferred ART eligibility saved 0.18 years of life over a five year horizon at a cost of $1967 per life year saved. Given the additional reductions in immune function that we identify, and the likely implications for HIV-related morbidity, these are lower-bound estimates on the clinical cost-effectiveness of raising the ART eligibility threshold from 200 cells/μL.

IX. CONCLUSION

Clinical decision-making is often based on standardized guidelines, with treatment indicated if the results of a diagnostic test are above (or below) some threshold value. Threshold decision rules have the advantage that they are easy to implement by lower skilled health workers, enabling more rapid diffusion of medical technologies in low resource settings. For patients close to the threshold, however, the decision is arbitrary and the costs of ineligibility can be high.

We assessed the long run health benefits of immediate vs. deferred ART treatment eligibility for patients in a public sector HIV care and treatment program in rural South Africa. ART eligibility at baseline had a large and statistically significant impact on both survival and immune health. Patients who were initially prevented from starting ART because their CD4 count was above the 200-cell threshold were 15% points less likely to be alive three years later, and lost (on average) 0.59 years of life over the five-year follow-up period, from a baseline of 100% survival. These benefits are local to "compliers", i.e. patients whose initiation decision was made based on CD4 count eligibility. It is likely that the effect of immediate vs. deferred ART would have been even larger for patients who presented clinically with Stage IV AIDS-defining illness and who would have been initiated regardless of CD4 count. In other words, these estimates are likely lower bounds on the effect of the treatment (rapid ART initiation) on the treated.

The gains in survival were attained at a modest cost of $1967 per life year saved, less than a third of South Africa's 2011 per capita GNI, $6960 (World Bank, 2012). By conventional benchmarks, interventions that save a year of life for less than 1x per capita GNI are deemed "very cost effective" (Goldie et al. 2006). In addition to the reductions in mortality, among those who survived, patients prevented from starting ART because they were ineligible went on to have follow-up CD4 counts that were about 225 cells/μL lower than their baseline-eligible counterparts. These differences in immune health have clinically meaningful implications for the incidence of opportunistic infections and for the costs of medical care (Meyer Rath, et al. 2013), not included in our conservative cost-effectiveness estimates. Further, our estimates exclude the prevention benefits of initiating patients on ART earlier.

As with other chronic diseases, management of HIV requires a long-term relationship between patients and the health system. The effect of immediate (vs. deferred) treatment on future health outcomes may

be determined not just by the medical benefit (or harm) of the drugs, but also by the effect of starting therapy on patient retention. Among patients whose treatment decision was made by the threshold rule, non-eligibility reduced clinical retention from 100% to 63% at 6-18 months. Further, the effects of non-eligibility on survival were concentrated among patients who may have faced physical and psychological barriers to retention: i.e., those who lived further from clinics and who did not live with another HIV patient in their household. Behavioral responses to non-eligibility – or conversely, behavioral responses to starting treatment – may drive a wedge between the efficacy of a therapy and its effectiveness in real world settings and need to be considered in designing clinical guidelines.

These are the first quasi-experimental estimates of the survival benefits of immediate vs. deferred ART eligibility in sub-Saharan Africa. To date no experimental evidence exists and new trials are unlikely to be forthcoming given that current WHO guidelines have outpaced the evidence base on the clinical impacts of early initiation (WHO 2013). This study provides causal evidence to clinicians interested in when to start patients on therapy, and for policy makers debating where to direct scarce resources for health. We also illustrate the delicate nature of chronic disease management in resource poor settings, in which the results of a routine screening test have life and death consequences for so many. Designing clinical models of chronic disease management to be more robust to adverse behavioral responses could lower the stakes and save lives.

REFERENCES

Abadie A. Boostrap tests for distributional treatment effects in instrumental variables models. *Journal of the American Statistical Association.* 2002;97(457):284-292.

Allison P. Handling Missing Data by Maximum Likelihood. *SAS Global Forum*. 2012;paper 312.

Almond D, et al. Estimating marginal returns to medical care: Evidence from at-risk newborns. *Q J Econ.* 2010; 125(2):591-634.

Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc.* 1996;91:444-472.

Baicker K, Taubman SL, Allen HL, Bernstein M, Gruber JH, Newhouse JP, ... & Finkelstein AN. The Oregon experiment—effects of Medicaid on clinical outcomes. *New Engl J Med*. 2013; 368(18), 1713-1722.

Barreca AI, Lindo JM, Waddell GR. Heaping-induced bias in regression-discontinuity designs (No. w17408). National Bureau of Economic Research, 2011.

Bor J, Bärnighausen T, Newell C, Tanser F, Newell M-L. Social exposure to an antiretroviral treatment programme in rural KwaZulu-Natal. *Trop Med Int Health*. 2011;16(8):988-94.

Bor J, Herbst AJ, Newell M-L, Bärnighausen T. Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment. *Science*, 2013;339(6122): 961-965.

Bor J, Moscoe E, Mutevedzi P, Newell M-L, Bärnighausen T. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology.* 2014;25:729–737.

Bor J, Moscoe E, Bärnighausen T. Three approaches to causal inference in regression discontinuity designs. *Epidemiology*. 2015;26(2):e28-e30.

Card D, Dobkin C. Does Medicare Save Lives? *Q J Econ*. 2009;124(2): 597-636.

Caughey, Devin and Jasjeet S Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close US House Races, 1942–2008." Political Analysis 19(4):385–408.

Chay K, McEwan P, Urquiola M. The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review.* 2005;95:1237-58.

Chen Y, Ebenstein A, Greenstone M, Li H. Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *Proc Natl Acad Sci*. 2013;110(32):12936-41.

Cohen MS, Chen YQ, McCauley MM, et al. Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med. 2011;365:493-505.

Cutler D, Deaton A, & Lleras-Muney A. The Determinants of Mortality. *Journal of Economic*

Perspectives. 2006; 20(3), 97-120.

Cutler D, Miller G. The role of public health improvements in health advances: the twentieth-century United States. *Demography*. 2005; 42(1), 1-22.

De Cock KM, El-Sadr WM. When to start ART in Africa -- an urgent research priority. *N Engl J Med* 2013;368:886-889.

Doyle JJ. Health Insurance, Treatment and Outcomes: Using Auto Accidents as Health Shocks. *Rev Econ Stat.* 2005;87(2):256-270.

Drummond MF & McGuire A. (Eds.).*Economic evaluation in health care: merging theory with practice*. Oxford University Press, 2001.

Egger M, Hirschel B, Francioli P, et al. Impact of new antiretroviral combination therapies in HIV infected patients in Switzerland: prospective multicentre study. *BMJ.* 1997; 315: 1194–99.

Emery S, Neuhaus JA, Phillips AN, et al. Major clinical outcomes in antiretroviral therapy (ART)-naive participants and in those not receiving ART at baseline in the SMART study. *J Infect Dis.* 2008;197(8):1133-1144.

Finkelstein A, Taubman S, Wright B, Bernstein M, Gruber J, Newhouse JP, ... & Baicker K. The Oregon Health Insurance Experiment: Evidence from the First Year. *Q J Econ*. 2012; 127(3), 1057-1106.

Friedman W. Antiretroviral drug access and behavior change. *University of California, Berkeley Working Paper*, 2012.

Fox MP, Sanne IM, Conradie F, Zeinecker J, Orrell C, Ive P, ... & Wood R. Initiating patients on antiretroviral therapy at CD4 cell counts above 200 cells/microl is associated with improved treatment outcomes in South Africa. *AIDS.* 2010;24(13), 2041-2050.

Frandsen BR, Frölich M, Melly B. (2012). Quantile treatment effects in the regression discontinuity design. *J Econometrics*, *168*(2), 382-395.

Freedman B. Equipoise and the ethics of clinical research. *N Engl J Med* 1987; 317, (3):141–145.

Goldie SJ, Yazdanpanah Y, Losina E,... & Freedberg KA. Cost-effectiveness of HIV treatment in resource-poor settings—the case of Côte d'Ivoire. *N Engl J Med*. 2006; 355(11): 1141-1153.

Grinsztejn B, Hosseinipour MC, Ribaudo HJ, Swindells S, Eron J, Chen YQ... & Cohen MS. Effects of early versus delayed initiation of antiretroviral treatment on clinical outcomes of HIV-1 infection: results from the phase 3 HPTN 052 randomised controlled trial. *Lancet Inf Dis*. 2014; 14(4), 281-290.

Grossman M. The Demand for Health: A Theoretical and Empirical Investigation, National Bureau of Economic Research: New York, 1972.

Harrell F. Regression modeling strategies. Springer, New York, 2001.

Hastie TJ & Tibshirani RJ. Generalized Additive Models. *Monographs on Statistics and Probability, 43*. Chapman & Hall, London, 1990.

Haynes AB, Weiser TG, Berry WR,... & Gawande AA. A surgical safety checklist to reduce morbidity and mortality in a global population. *N Engl J Med* 2009;360(5),491-499.

Herbst AJ, Mafojane T, Newell M-L. Verbal autopsy-based cause-specific mortality trends in rural KwaZulu-Natal, South Africa, 2000-2009. *Population Health Metrics*. 2011;9:47.

Ho DD. Time to hit HIV, early and hard. *N Engl J Med* 1995; 333: 450–51.

Houlihan CF, Bland RM, Mutevedzi P, et al. Cohort profile: Hlabisa HIV Treatment and Care Programme. *Int J Epidemiol.* 2011;40:318.

Hughes MD, Stein DS, Gundacker HM, Valentine FT, Phair JP, Volberding PA. Within-subject variation in CD4 lymphocyte count in asymptomatic Human Immunodeficiency Virus infection: implications for patient monitoring. *J Infect Dis*. 1994;169(1):28-36.

Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. *Econometrica*. 1994;62(2):467-475.

Imbens GW, Lemieux T. Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics*. 2008;142(2):615–35.

Imbens GW, Rubin DB. Estimating outcome distributions for compliers in instrumental variable models. *Rev Econ Stud,*1997;64:555-574.

Kaplan, E. L.; Meier, P. (1958). "Nonparametric estimation from incomplete observations". *J. Amer. Statist. Assn.* **53** (282): 457–481

Kitahata MN, Gange SJ, Abraham AG, et al. Effect of early versus deferred antiretroviral therapy for HIV on survival. N Engl J Med. 2009;360:1815-26.

Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *The Stata Journal*. 2009;9(2):265-90.

Lancaster T. Econometric methods for the duration of unemployment. *Econometrica*, 1979;47:939-956.

Landsberger HA. *Hawthorne Revisited*. Ithaca: New York State School of Industrial and Labor Relations, 1958.

Lane HC, Neaton JD. When to start therapy for HIV infection: a swinging pendulum in search of data. *Ann Intern Med* 2003; 138: 680–81.

Lee DS. Randomized experiments from non-random selection in US House elections. *Journal of Econometrics*. 2008;142(2):675-697.

Lee DS, Lemieux T. Regression discontinuity designs in economics. *J Econ Lit*. 2010;48:281-355.

Lessells RJ, Mutevedzi PC, Cooke GS, Newell ML. Retention in HIV care for individuals not yet eligible for antiretroviral therapy: rural KwaZulu-Natal, South Africa. *JAIDS.* 2011;56(3):e79.

Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, ... & Cross M. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *The Lancet*. 2013;380(9859):2095-2128.

Mutevedzi PC, Lessells RJ, Heller T, Bärnighausen T, Cooke GS, Newell ML. Scale-up of a decentralized HIV treatment programme in rural KwaZulu-Natal, South Africa: does rapid expansion affect patient outcomes? *Bull World Health Org.* 2010;88(8), 593-600.

McCrary J. Manipulation of the running variable in the regression discontinuity design: A density test. *J Econom*. 2008;142:698-714.

Meyer-Rath G, Brennan AT, Fox MP, et al. Rates and cost of hospitalization before and after initiation of antiretroviral therapy in urban and rural settings in South Africa. *JAIDS.* 2013;62:322-28.

Molenberghs G, Kenward MG. *Missing Data in Clinical Studies.* Chichester, UK: John Wiley and Sons Ltd, 2007.

Moscoe E, Bor J, Bärnighausen T. Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *J Clin Epidemiol*. 2015;68:122-133.

National Institute of Allergy and Infectious Diseases (NIAID). Effectiveness of immediate and deferred antiretroviral treatment in HIV-infected adults. March 20, 2009. http://clinicaltrials.gov/ct2/show/NCT00821171?term=START&cond=HIV&intr=antiretroviral&rank=2 (accessed Nov 21, 2011).

Nichols AL & Zeckhauser RJ. Targeting transfers through restrictions on recipients. *The American Economic Review*. 1982;372-377.

Omran AR. The epidemiologic transition: a theory of the epidemiology of population change. *The Milbank Memorial Fund Quarterly*. 1971;509-538.

Palella FJ, Delaney KM, Moorman AC, et al. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. N Engl J Med 1998; 338: 853–908.

Palella FJ, Jr., Deloria-Knoll M, Chmiel JS, et al. Survival benefit of initiating antiretroviral therapy in HIV-infected persons in different CD4+ cell strata. Ann Intern Med. 2003;138(8):620-626.

Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine.* 2002; 21:2175-2197.

Sanne I, Orrell C, Fox MP, Conradie F, Ive P, Zeinecker J, ... & CIPRA-SA Study Team. Nurse versus doctor management of HIV-infected patients receiving antiretroviral therapy (CIPRA-SA): a randomised non-inferiority trial. *The Lancet*. 2010;376(9734):33-40.

Severe P, Juste MA, Ambroise A, et al. Early versus standard antiretroviral therapy for HIV-infected

adults in Haiti. N Engl J Med. 2010;363(3):257-265.

Shigeoka H, Fushimi K. Supplier-induced demand for newborn treatment: Evidence from Japan. *Journal of Health Economics*. 2014;35:162-178.

Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ* 2003; 327(7429): 1459-61.

Sterne JA, May M, Costagliola D, et al. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. Lancet. 2009;373(9672):1352-1363.

National Department of Health (NDOH). National Antiretroviral Treatment Guidelines, 1st Edition. South Africa, 2004.

Tanser F, Bärnighausen T, Graspa E, Zaidi J, Newell M-L. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. *Science*. 2013;339(6122):966-971.

Tanser F, Hosegood V, Bärnighausen T, et al. Cohort profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. *Int J Epidemiol*. 2008;37:956.

Thistlethwaite D, Campbell D. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *Journal of Educational Psychology.* 1960;51:309-317.

Verbeke G, Molenberghs G. *Linear Mixed Models for Longitudinal Data.* Secaucus, NJ: Springer, 2000.

Weiss NS, Koepsell TD, & Psaty BM. (2008). Generalizability of the results of randomized trials. *Archives of internal medicine*, *168*(2), 133-135.

Wooldridge JM. *Econometric Analysis of Cross-Sectional and Panel Data*. Cambridge, MA: The MIT Press, 2002.

World DataBank, World Development Indicators and Global Development Finance. The World Bank Group, Washington, DC, 2012.

World Health Organization (WHO). (2003). Scaling up antiretroviral therapy in resource-limited settings: treatment guidelines for a public health approach. Geneva: WHO.

World Health Organization (WHO). (2010). Antiretroviral Therapy for HIV Infection in Adultsa and Adolescents: Recommendations for a Public Health Approach: 2010 revision. Geneva: WHO.

World Health Organization (WHO). (2013). Antiretroviral Therapy for HIV Infection in Adultsa and Adolescents: Recommendations for a Public Health Approach: 2013 revision. Geneva: WHO.

World Health Organization (WHO) & International Diabetes Federation. (2006). Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia. Geneva: WHO.

World Health Organization & International Society of Hypertension Writing Group (WHO/ISH). 2003 World Health Organization (WHO)/International Society of Hypertension (ISH) statement on management of hypertension. *Journal of Hypertension*. 2003; 21(11):1983-1992.

Zhao M, Konishi Y, & Glewwe P. Does information on health status lead to a healthier lifestyle? Evidence from China on the effect of hypertension diagnosis on food consumption. *Journal of Health Economics*. 2013;32(2):367-385.

Table 1, Treatment eligibility and rapid ART initiation

	Earliest CD4+ count	**Probability of initiating ART within six months**				**Sample**
	Specification; Range	$E[Y\|Z \uparrow c]$	$E[Y\|Z \downarrow c]$	*Difference*	*95% CI*	*N*
(1)	Linear; 0-350	0.67	0.36	0.31	(0.25, 0.37)	4113
(2)	Linear; 50-350	0.69	0.36	0.32	(0.26, 0.38)	3548
(3)	Linear; 100-300	0.66	0.42	0.24	(0.16, 0.31)	2471
(4)	Linear; 150-250	0.66	0.45	0.21	(0.10, 0.32)	1256
(5)	Linear; 175-225	0.65	0.44	0.21	(0.06, 0.37)	610
(6)	Quadratic; 0-350	0.67	0.47	0.20	(0.12, 0.29)	4113
(7)	Quadratic; 50-350	0.66	0.47	0.19	(0.10, 0.28)	3548
(8)	Quadratic; 100-300	0.68	0.47	0.21	(0.10, 0.33)	2471
(9)	Quadratic; 150-250	0.66	0.46	0.20	(0.04, 0.36)	1256
(10)	Quadratic; 175-225	0.77	0.60	0.17	(-0.05, 0.38)	610

Notes: Linear probability models. Each row is its own regression.

Table 2. Treatment eligibility and all-cause mortality: flexible parametric models.

Range: 50 – 350 cells

Time since first CD4+ count (*t*)	**Cumulative probability of death (1 - survival)** $F[t\|CD4 \uparrow 200]$	$F[t\|CD4 \downarrow 200]$	*Difference in F(t)*	*95% CI*
6 months	3.4%	3.7%	-0.3%	-2.5%, 2.0%
1 year	4.6%	6.5%	-1.8%	-4.7%, 1.1%
2 years	6.6%	10.9%	-4.3%	-8.0%, -0.6%
3 years	8.8%	13.6%	-4.8%	-9.0%, -0.6%
4 years	10.7%	15.3%	-4.5%	-9.2%, 0.1%
5 years	12.6%	16.6%	-4.0%	-9.5%, 1.5%
Years of life lost (over 5 year horizon)			-0.18	-0.26, -0.12

Time since first CD4+ count (*t*)	**Instantaneous hazard of death** $h[t\|CD4 \uparrow 200]$	$h[t\|CD4 \downarrow 200]$	*Hazard ratio*	*95% CI*
6 months	0.07	0.06	1.07	0.50. 2.30
1 year	0.01	0.06	0.24	0.10, 0.60
2 years	0.02	0.04	0.62	0.30, 1.26
3 years	0.02	0.02	0.96	0.47, 1.98
4 years	0.02	0.02	1.24	0.40, 3.90
5 years	0.02	0.01	1.41	0.36, 5.44

Note: Models estimated for patients presenting with CD4 counts between 50 and 350 cells, with linear functions estimated on either side of the threshold; n=3710. See Tables S1 and S2 for sensitivity analyses changing the bandwidth and including higher order polynomials in first CD4 count.

Table 3. Treatment eligibility and HIV/TB mortality: flexible parametric models.

HIV/TB-related mortality

Time since first CD4+ count (*t*)	**Cumulative probability of death (1 - survival)** $F[t\|CD4 \uparrow 200]$	$F[t\|CD4 \downarrow 200]$	*Difference in F(t)*	*95% CI*
6 months	2.9%	3.3%	-0.4%	-2.6%, 1.7%
1 year	4.0%	6.1%	-2.1%	-4.9%, 0.8%
2 years	5.5%	10.3%	-4.8%	-8.4%, -1.2%
3 years	7.1%	12.6%	-5.5%	-9.6%, -1.4%
4 years	8.7%	14.0%	-5.3%	-9.8%, -0.8%
5 years	10.1%	15.0%	-4.8%	-10.1%, 0.4%
Years of life lost to HIV (over 5 year horizon)			-0.20	-0.30, -0.14

Non-HIV/TB-related mortality

Time since first CD4+ count (*t*)	**Cumulative probability of death (1 - survival)** $F[t\|CD4 \uparrow 200]$	$F[t\|CD4 \downarrow 200]$	*Difference in F(t)*	*95% CI*
6 months	0.5%	0.5%	-0.0%	-0.8%, 0.7%
1 year	0.7%	0.6%	0.1%	-0.8%, 0.9%
2 years	1.1%	1.4%	-0.2%	-1.5%, 1.0%
3 years	2.0%	2.8%	-0.8%	-2.8%, 1.1%
4 years	2.8%	4.4%	-1.6%	-4.4%, 1.1%
5 years	3.7%	6.3%	-2.5%	-6.7%, 1.6%
Years of life lost to other causes (over 5 year horizon)			-0.04	-0.10, -0.00

Notes: Predictions from flexible parametric survival models, for patients presenting with CD4 counts of 50-350 cells; n=3710. Person-time was censored at the time of the competing event. Models for HIV-related mortality were estimated using four knots for the baseline log-cumulative-hazard, and four knots for the time-varying treatment effect. Due to small numbers of non-HIV-related deaths, those models were estimated using a spline with three knots for the baseline log-cumulative-hazard and two knots for the time-varying effects of covariates.

Table 4. Effect of ART eligibility on follow-up CD4 counts.

Linear Regression (Least Squares)

Time since first CD4+ count (*t*)	**Mean CD4 count at follow-up** $E[Y\|t, CD4 \uparrow 200]$	$E[Y\|t, CD4 \downarrow 200]$	*Difference*	*95% CI*
1 year	366	314	52	17, 87
2 years	416	355	61	19, 103
3 years	446	376	70	25, 115
4 years	472	390	82	18, 146
5 years	497	403	94	-8, 196

Linear Mixed Effects Model (Maximum Likelihood)

Time since first CD4+ count (*t*)	**Mean CD4 count at follow-up** $E[Y\|t, CD4 \uparrow 200]$	$E[Y\|t, CD4 \downarrow 200]$	*Difference*	*95% CI*
1 year	351	303	48	21, 76
2 years	416	345	71	32, 110
3 years	452	377	75	33, 117
4 years	474	406	69	13, 124
5 years	494	434	60	-26, 146

Notes: Predictions in the top panel are from linear regression discontinuity model with time-varying effects modeled as a restricted cubic spline, interacted with covariates. Predictions in the bottom panel are for a mixed effects model, in which individual specific intercepts and growth curves are modeled as random effects. This latter model is robust to missingness correlated with patients' own CD4 count history. In both panels, models were estimated for patients presenting with CD4 counts in the range 100-300; n=2557.

Table 5. Effect of ART eligibility on retention in care.

Linear Probability Models

Time since first CD4+ count (*t*)	**Probability Retained in HIV Care** $E[Y\|t, CD4 \uparrow 200]$	$E[Y\|t, CD4 \downarrow 200]$	*Difference*	*95% CI*
50 – 350 cells, (N = 3705)				
0-6 months	77.8%	57.1%	20.7%	14.8%, 26.6%
6-12 months	61.3%	47.8%	13.5%	7.2%, 19.8%
12-18 months	49.5%	40.2%	9.1%	2.8%, 15.4%
18-24 months	46.5%	36.7%	9.7%	3.5%, 16.0%
6-18 months	74.5%	61.6%	12.9%	6.9%, 18.9%
100 – 300 cells, (N = 2543)				
0-6 months	75.0%	58.5%	16.5%	9.3%, 23.6%
6-12 months	60.0%	48.1%	11.8%	4.1%, 19.5%
12-18 months	46.7%	38.8%	7.8%	0.2%, 15.5%
18-24 months	43.8%	34.4%	9.4%	1.7%, 17.0%
6-18 months	73.0%	61.6%	11.3%	4.0%, 18.7%
150 – 250 cells, (N = 1281)				
0-6 months	76.5%	61.3%	15.2%	4.9%, 25.5%
6-12 months	63.6%	47.8%	15.8%	4.7%, 26.9%
12-18 months	41.4%	37.9%	3.4%	-7.5%, 14.4%
18-24 months	42.4%	38.0%	4.4%	-6.6%, 15.4%
6-18 months	73.0%	62.0%	11.0%	0.5%, 21.5%

Notes: Each row is its own linear probability regression model. The outcome "retained at time t" was defined as any lab test (CD4 or viral load) or initiation of ART within a six-month interval. All patients regardless of their eligibility for treatment are instructed to return to the clinic every six months for lab tests. Confidence intervals are based on heteroskedasticity-robust standard errors.

Table 6. "Fuzzy RD": the effect of "rapid ART initiation" on probability of death, years of life lost, retention in care, and mean CD4 count among patient compliers

Effect estimate at threshold: $E[Y|CD4 \uparrow 200] - E[Y|CD4 \downarrow 200]$

Outcome	First Stage	ITT_{RDD}	$CACE_{RDD}$
Rapid ART initiation	32.2% (26.6, 37.8)		
Probability of death in:			
1 year		-2.0% (-4.0, -0.1)	-6.3% (-12.8, 0.2)
2 years		-4.1% (-7.2. -1.0)	-12.7% (-23.0, -2.3)
3 years		-5.1% (-9.0, -1.2)	-15.9% (-28.2, -3.5)
4 years		-4.7% (-9.1, -0.2)	-14.5% (-28.3, -0.8)
5 years		-3.7% (-9.0, 1.7)	-11.4% (-27.7, 5.0)
Years of life lost, over 5 year horizon		-0.18 (-0.24, -0.13)	-0.55 (-1.00, -0.10)
Follow-up CD4 count:			
1 year		72 (43, 101)	217
2 years		77 (41, 114)	233
3 years		75 (35, 116)	227
4 years		73 (18, 128)	221
5 years		71 (-15, 157)	215
Retention at 6-18 months		11.9% (5.9, 17.9)	37.5%

Notes: All models exclude patients who died in the first six months or who had less than six months of follow-up; thus results should be interpreted as conditional on survival to six months. As shown in Table 2, there was no significant difference in survival at six months between patients presenting just above vs. just below the eligibility threshold. Rapid ART initiation is an indicator for whether a patient initiated ART within six months of her first CD4 count. Differences in the probability of death in 1,2,…,5 years and differences in life years lost were estimated based on a flexible parametric survival model similar to Table 2. Differences in CD4 counts were estimated based on linear regression models similar to Table 4, panel A. All models are estimated with linear terms on either side of the threshold, for patients presenting with CD4 counts of 50-350; n=3449. 95% CIs shown. For differences in mortality probabilities, CACE confidence intervals are Normal-based CIs obtained by bootstrapping with 488 resamples.

Table 7, Treated and control complier means: retention in care and mortality

Panel A.	(1)	(2)	(3)	(4)
	Conditional on treated		*Conditional on <u>not</u> treated*	
Estimand	$E[Y\|CD4 \uparrow 200, ART6m = 1]$	$E[Y\|CD4 \downarrow 200, ART6m = 1]$	$E[Y\|CD4 \uparrow 200, ART6m = 0]$	$E[Y\|CD4 \downarrow 200, ART6m = 0]$
Latent type	*Always-takers + compliers*	*Always-takers*	*Never-takers*	*Never-takers + compliers*
Retained 6-18 months	89.7%	79.9%	45.2%	54.1%
Death by 3 years	-	-	10.6%	13.1%

Panel B. **Time since first CD4+ count (*t*)**	(5) $E[Y\|Treated\ Compliers]$	(6) $E[Y\|Control\ Compliers]$
Retained 6-18 months	100%*	63.1%
Death by 3 years	0%*	15.6%

Notes: Panel A displays predicted retention and mortality risk, limiting the sample to patients who initiated ART within six months, i.e. the "treated", models (1) and (2) and to patients who did not initiate ART within six months, i.e. the "not treated", models (3) and (4). Estimates are from linear probability models (retention) and flexible parametric survival models (mortality) estimated for patients surviving at 6 months. Panel B displays predicted probabilities of retention and death for treatment and control compliers. The conditional-on-treated flexible parametric survival model did not converge and treated complier mean was obtained by subtracting off CACE from the control complier mean. *For treated compliers, 6-18mo retention was estimated to be 100.6%; 3-year mortality was estimated to be -0.0%; due to imprecision, probability estimates are not bounded on the 0,1 interval.

Table 8. Effect heterogeneity by baseline characteristics: reduced form hazard models

Exponential Hazard Models

	Mortality hazard			
50 – 350 cells	$h\|CD4 \uparrow 200$	$h\|CD4 \downarrow 200$	*Hazard Ratio*	*95% CI*
A. Full sample (n=3710)	2.9 deaths per 100 person-years	4.4 deaths per 100 person-years	0.65	(0.45, 0.94)
B. Distance to clinic				
Not in DSA (n=677)	3.6	4.7	0.77	(0.34, 1.75)
0-2 km (n=1314)	2.5	2.9	0.86	(0.44, 1.69)
>2 km (n=1719)	2.8	5.8	0.49	(0.29, 0.83)
>2 km * CD4elig			0.57	(0.24, 1.33)
C. Other HIV patient in the household				
No (n=2913)	2.8	5.0	0.56	(0.37, 0.84)
Yes (n=797)	3.1	2.2	1.38	(0.52, 3.66)
No * CD4elig			0.41	(0.14, 1.17)
D. Age of patient				
<30 (n=1512)	1.9	3.7	0.51	(0.26, 0.99)
30-45 (n=1542)	2.9	4.7	0.62	(0.35, 1.11)
45+ (n=656)	5.5	5.5	0.99	(0.50, 1.99)
45+ * CD4elig			0.58	(0.25, 1.31)
E. Gender of patient				
Male (n=1042)	4.9	4.5	1.09	(0.60, 1.96)
Female (n=2668)	2.2	4.8	0.47	(0.29, 0.75)
Female * CD4elig			0.43	(0.20, 0.92)
F. Education level of patient				
<12 years (n=2539)	3.3	4.8	0.70	(0.46, 1.08)
12+ years (n=1171)	2.0	3.4	0.57	(0.28, 1.18)
12+ years * CD4elig			0.81	(0.35, 1.88)

Notes: Each row was estimated in a separate exponential hazard model, estimated over the range 50 to 350 cells/mm^3, with separate linear terms on either side of the threshold. For each panel, the first two rows display the results of separate regressions for each subgroup. The final row in panels B-F presents the hazard ratio on the interaction term between the group and the threshold rule in a pooled model allowing for different slopes and intercepts for each group.

Figure 1. Distribution of CD4+ counts at clinical enrollment

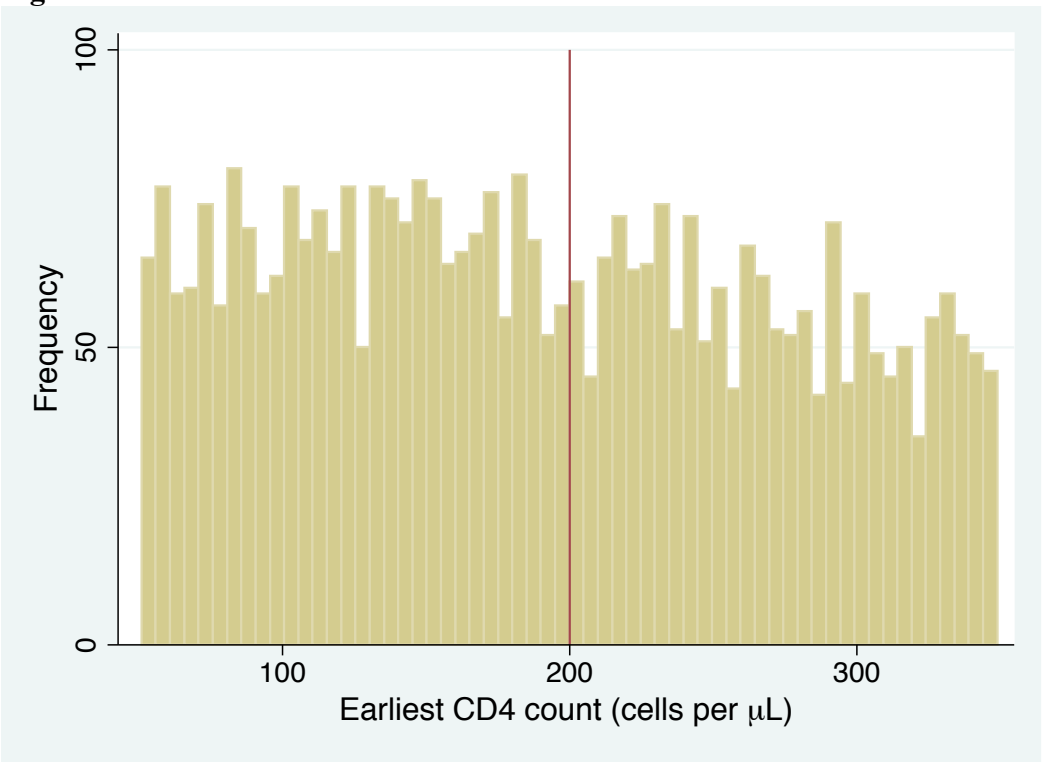
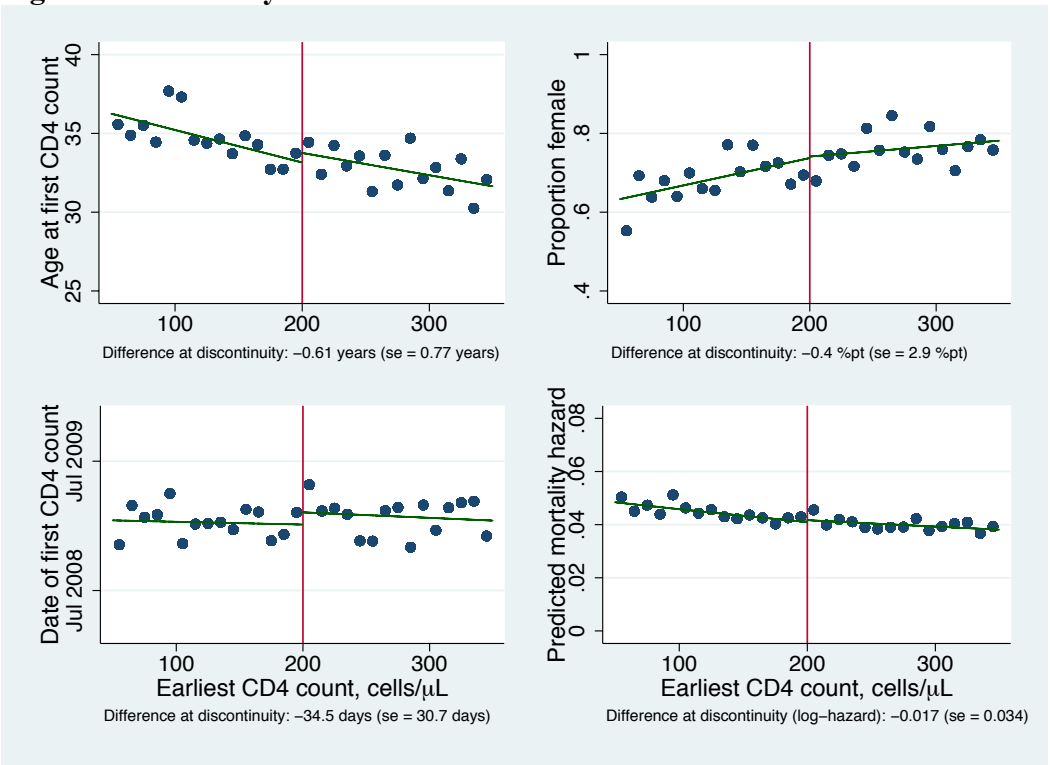
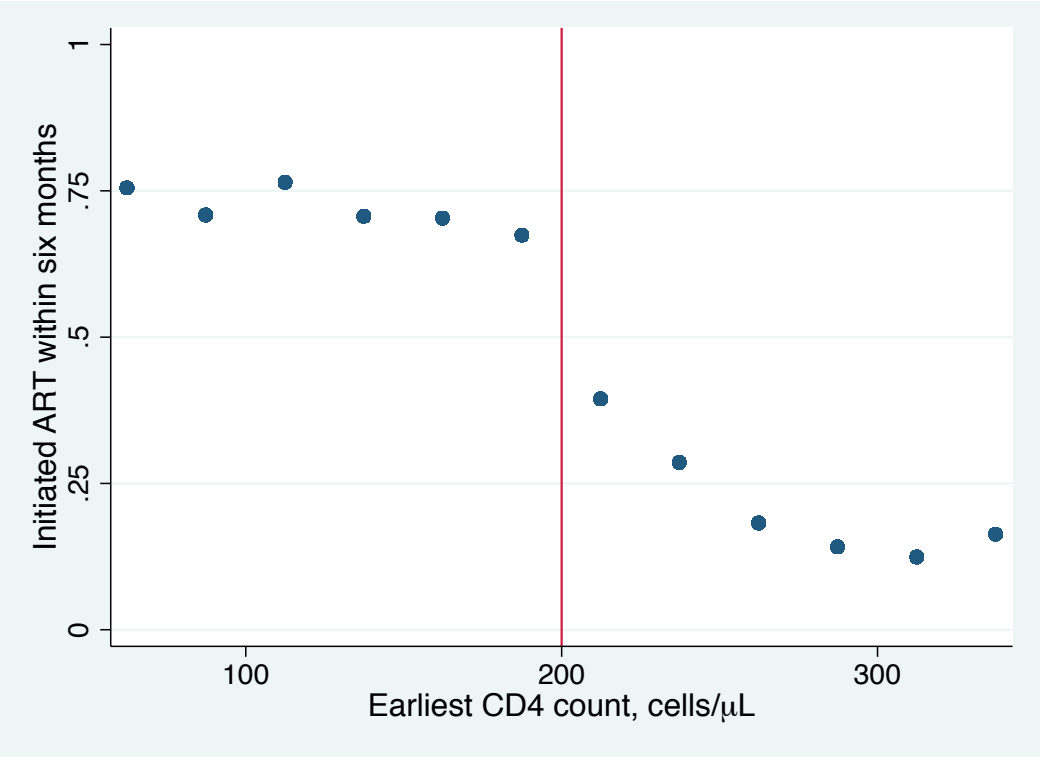


Figure 2. Continuity in baseline covariates



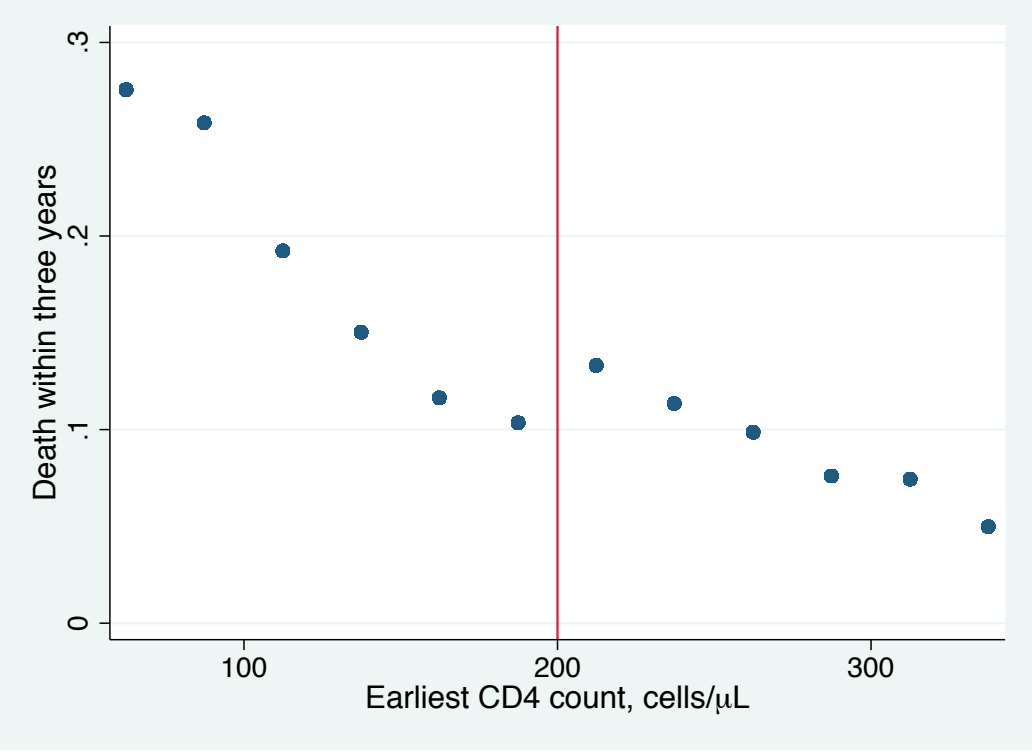
Notes: Figure plots age, sex, and date of first CD4 count against value of first CD4 count. Bottom-right panel displays mortality hazards predicted using baseline covariates. Log-hazards were predicted in an exponential regression model, controlling for sex, age, age^2, and their interactions. Geometric mean hazards are shown for 10-cell CD4+ count bins. Fitted lines were estimated by regressing predicted log hazards on CD4 count and exponentiating the predictions.

Figure 3. Baseline CD4+ count and probability of ART initiation



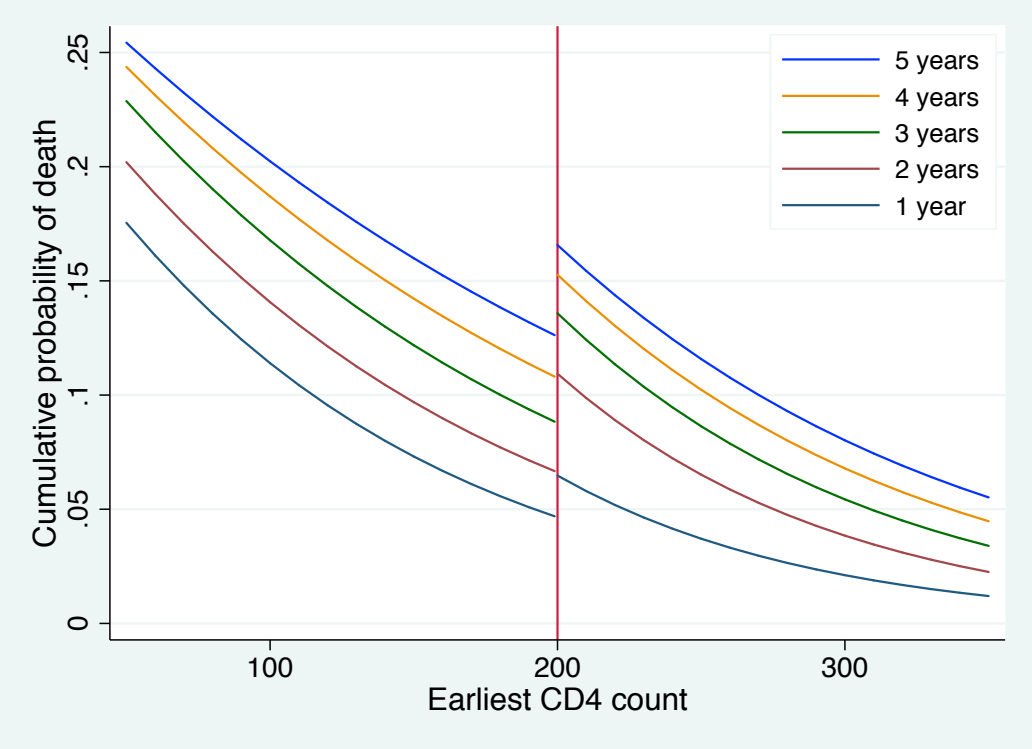
Notes: Kaplan-Meier estimates of probability that a patient initiated ART within X months of first CD4+ count in care. Follow-up time was censored at date of death or last survey visit.

Figure 4. Baseline CD4+ count and probability of death in three years



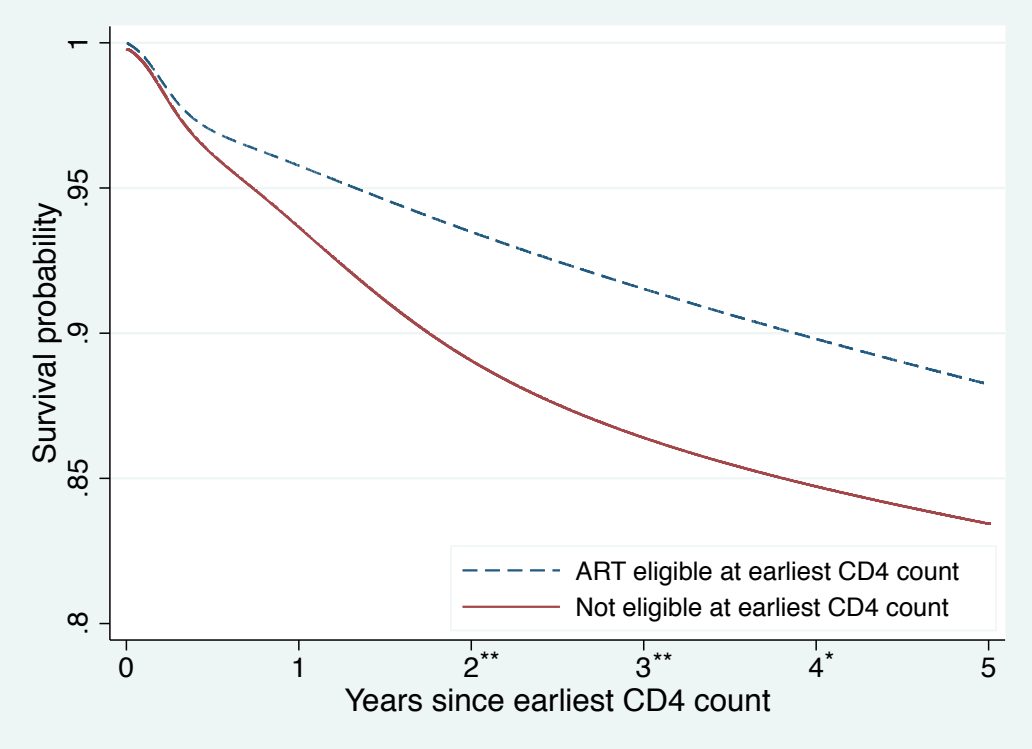
Notes: Kaplan-Meier estimates of probability that a patient initiated ART within X months of first CD4+ count in care. Follow-up time was censored at date of death or last survey visit.

Figure 5. Baseline CD4+ count and cumulative probability of death, as predicted in flexible parametric survival model



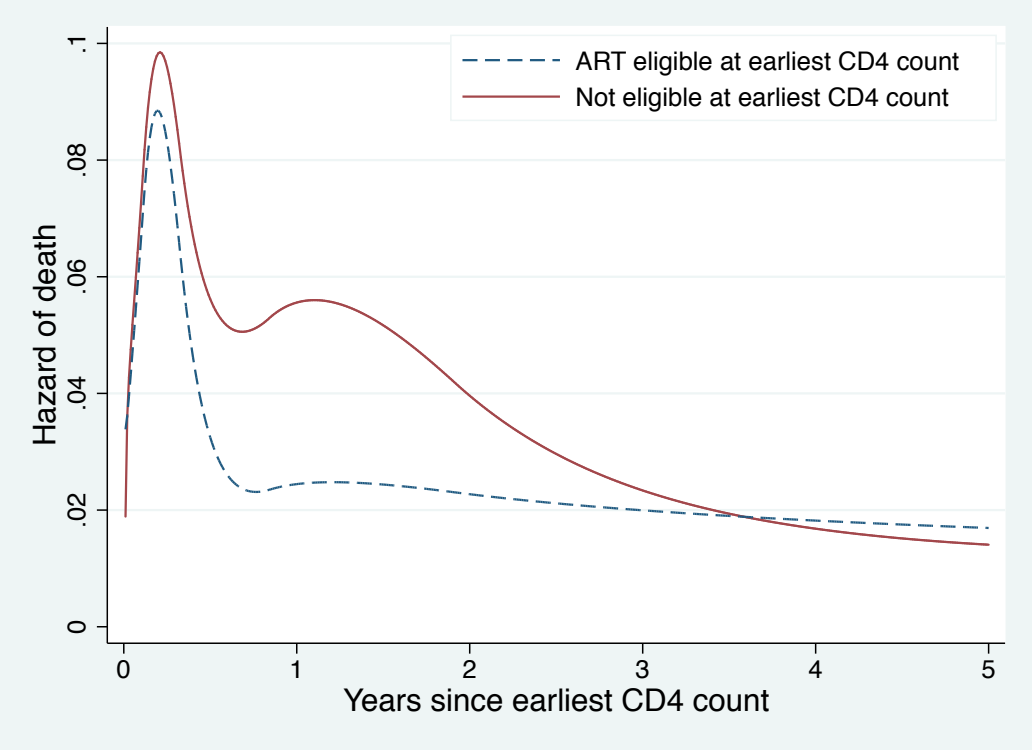
Notes: Predicted probabilities of death within 1, 2,…,5 years based on flexible parametric survival model, estimated for range 50-350 CD4 cells.

Figure 6. Predicted survival curves for patients presenting at the 200-cell threshold



Survival curves predicted for patients presenting on either side of 200 CD4 count threshold. Predicted based on flexible-parametric hazard model. Significance of difference between survival curves at annual intervals: ** p<.05; * p<.1

Figure 7. Baseline CD4+ count and mortality hazard for patients presenting at the 200-cell threshold



Instantaneous mortality hazards predicted for patients presenting on either side of 200 CD4 count threshold, based on flexible-parametric hazard model.

Figure 8. Mean CD4+ count at 12 months follow-up among patients surviving and still in care

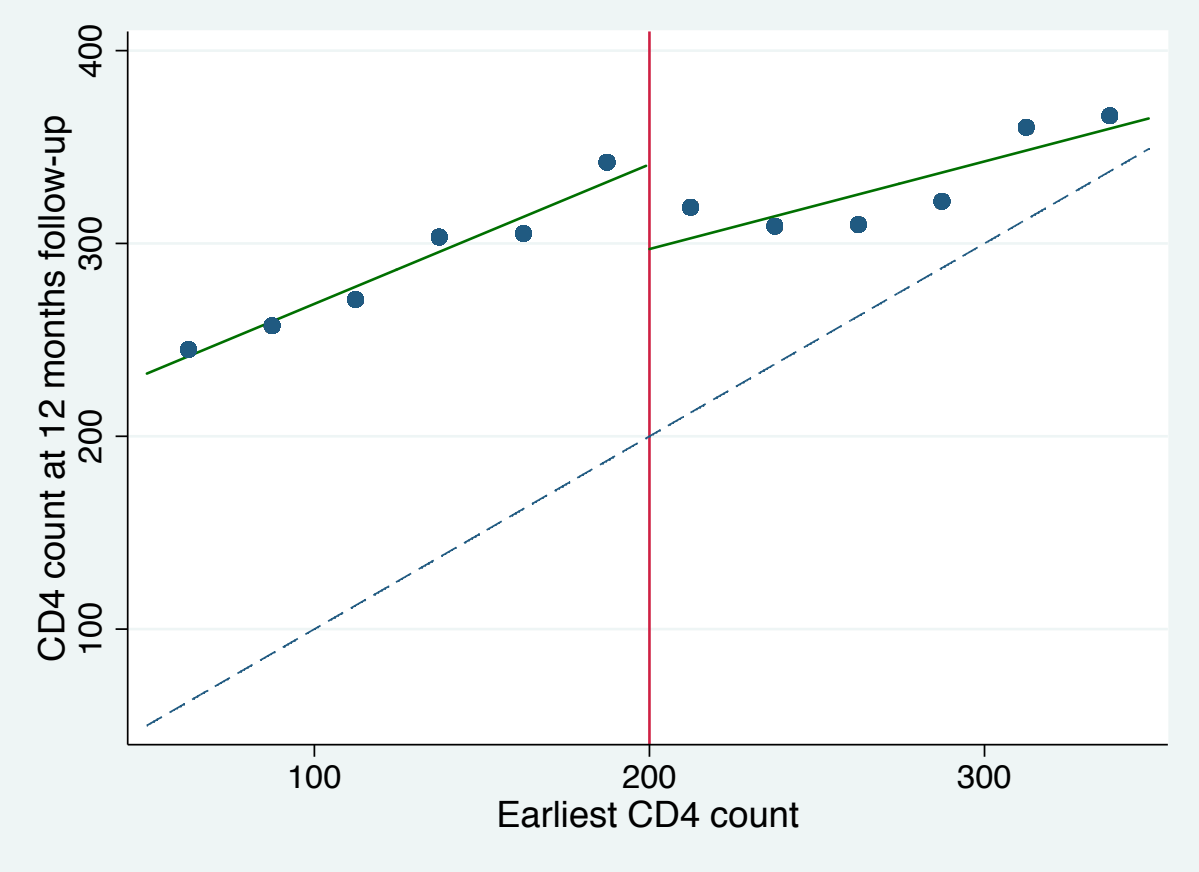


Figure displays mean CD4 counts for 1753 (of 4391) patients with follow-up CD4 counts between 9 and 15 months follow-up. For patients with multiple CD4 counts in this interval, the test date closest to 12 months was retained. Dotted 45° line shows the scenario if there had been no change in CD4 counts at 12 months.

Figure 9. Predicted CD4 counts for patients presenting at the 200-cell threshold

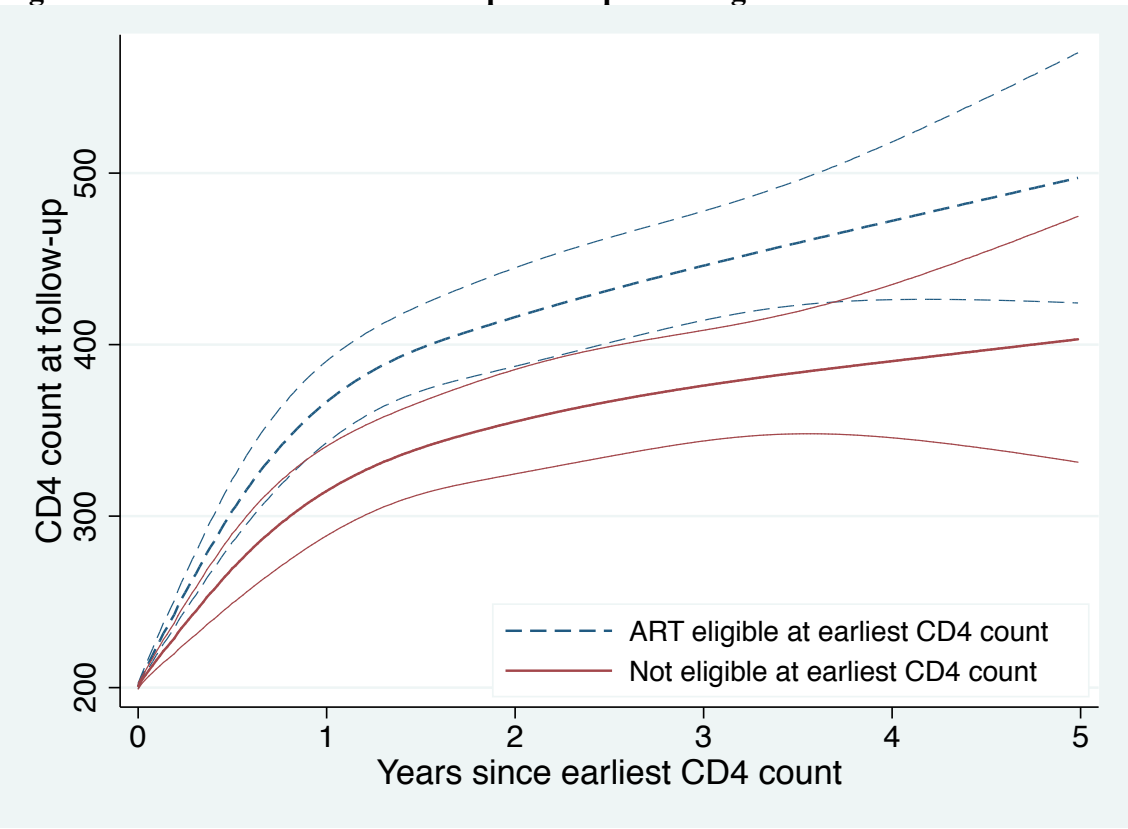


Figure displays predicted mean CD4 counts over time for patients presenting with an initial CD4 count just below (eligible) vs. just above (not eligible) the 200-cell threshold. Linear regression-discontinuity models were estimated with the effect of time modeled as a cubic spline, and interacted with the regression discontinuity coefficient and linear terms on either side of the discontinuity. Patients presenting with CD4 counts between 100 and 300 cells were included. The model was estimated based on data from survivors retained in care; follow-up was censored at the date of a patient's last CD4 count. 95% confidence bands are shown.

Figure 10. Distributions of CD4+ counts for patients presenting at the 200-cell threshold

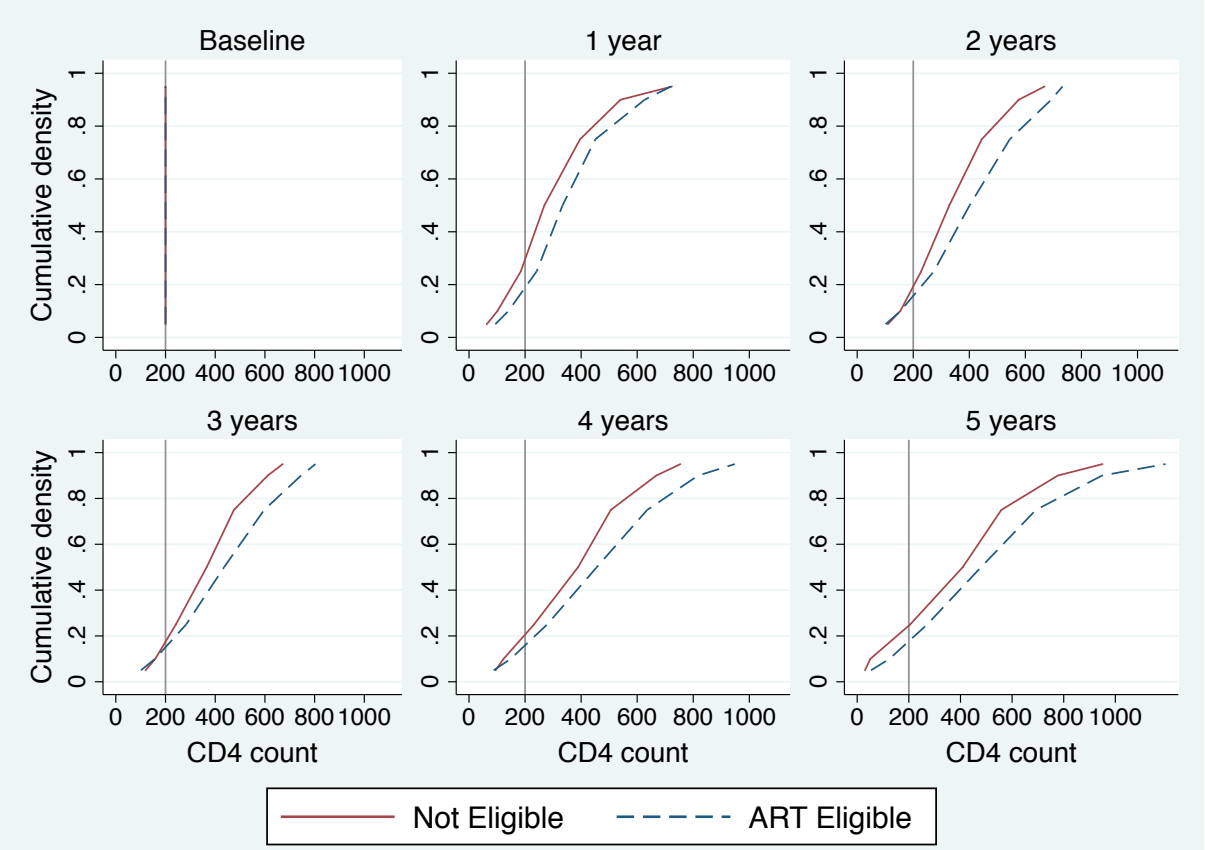


Figure displays cumulative density functions of CD4 counts at baseline and 1, 2, …, and 5 years follow-up, for patients presenting with an initial CD4 count just below (eligible) vs. just above (not eligible) the 200-cell threshold. CDFs are constructed as 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentile predictions from quantile regression-discontinuity models, estimated with the effect of time modeled as a cubic spline, and interacted with the regression discontinuity coefficient and linear terms on either side of the discontinuity. Patients presenting with CD4 counts between 100 and 300 cells were included. The model was estimated based on data from survivors retained in care; follow-up was censored at the date of a patient's last CD4 count. 95% confidence bands are shown.

Figure 11. Retention in clinical HIV care at 0-6, 6-12, 12-18, and 18-24 months

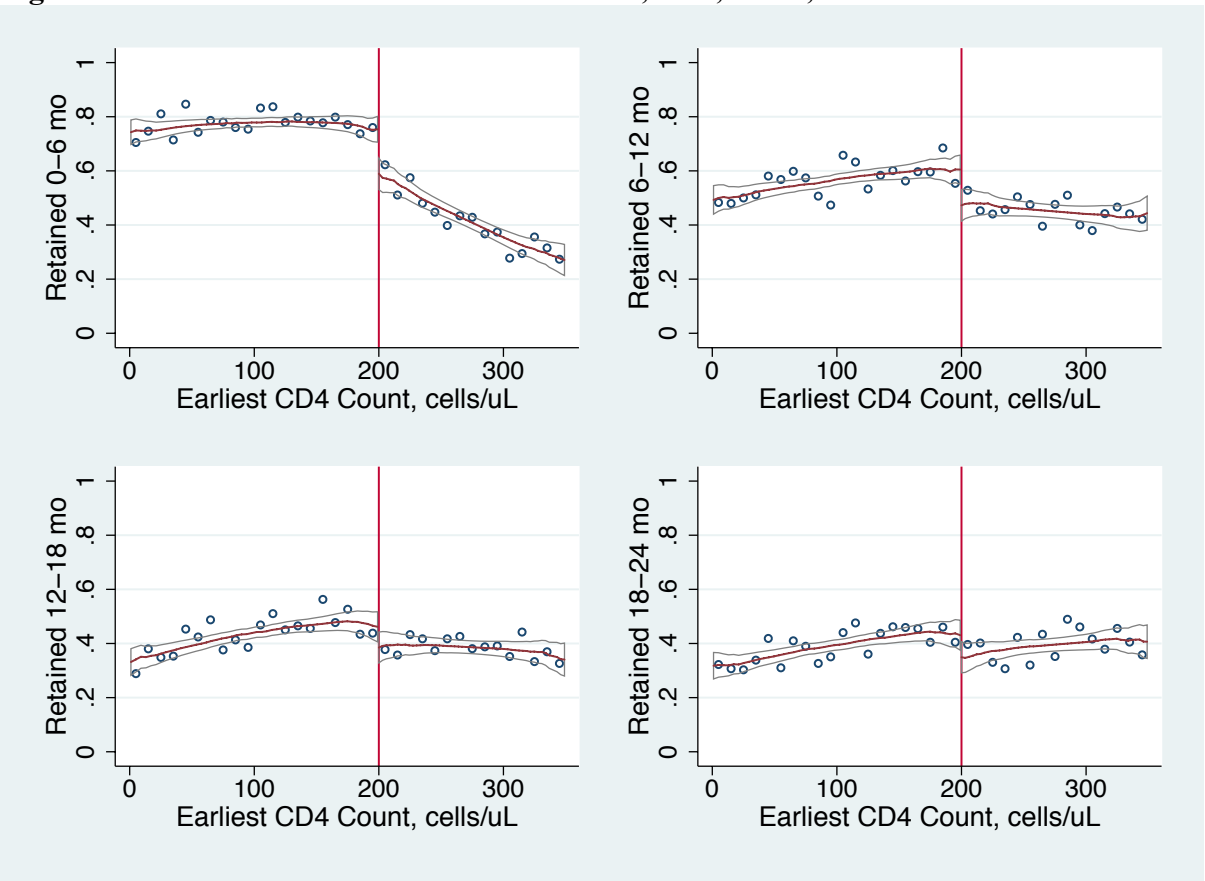


Figure displays probabilities of retention in care at 0-6, 6-12, 12-18, and 18-24 months by the value of a patient's first CD4 count. Regardless of their baseline eligibility for ART, all patients are to have a CD4 count and/or viral load laboratory test every six months as part of routine clinical care, according to guidelines. Retention in care was defined as having a CD4 count, viral load, or initiating ART, within six month intervals following a patient's first CD4 count. Fitted lines are local linear regressions with a rectangular kernel and bandwidth of 100 cells. 95% confidence bands are shown.

Table S1. Treatment eligibility and all-cause mortality: flexible parametric models, varying bandwidth.

Range: 0 – 350 cells	*(n=4391)*			
Time since first CD4+ count (*t*)	**Cumulative probability of death (1 - survival)** $F[t\|CD4 \uparrow 200]$	$F[t\|CD4 \downarrow 200]$	*Difference in F(t)*	*95% CI*
6 months	3.0%	3.9%	-0.8%	-3.0%, 1.4%
1 year	4.2%	6.4%	-2.1%	-4.8%, 0.6%
2 years	6.5%	11.0%	-4.4%	-8.0%, -0.9%
3 years	8.5%	13.6%	-5.1%	-9.1%, -1.2%
4 years	10.2%	15.3%	-5.1%	-9.5%, -0.7%
5 years	11.8%	16.6%	-4.8%	-9.9%, 0.3%
Years of life lost (over 5 year horizon)			-0.19	-0.30, -0.12
Range: 50 – 350 cells	*(n=3710)*			
6 months	3.4%	3.7%	-0.3%	-2.5%, 2.0%
1 year	4.6%	6.5%	-1.8%	-4.7%, 1.1%
2 years	6.6%	10.9%	-4.3%	-8.0%, -0.6%
3 years	8.8%	13.6%	-4.8%	-9.0%, -0.6%
4 years	10.7%	15.3%	-4.5%	-9.2%, 0.1%
5 years	12.6%	16.6%	-4.0%	-9.5%, 1.5%
Years of life lost (over 5 year horizon)			-0.18	-0.26, -0.12
Range: 100 – 300 cells	*(n=2557)*			
6 months	2.6%	3.3%	-0.7%	-3.2%, 1.8%
1 year	4.3%	6.4%	-2.1%	-5.5%, 1.3%
2 years	6.5%	11.6%	-5.0%	-9.5%, -0.5%
3 years	8.7%	13.7%	-5.0%	-10.2%, 0.1%
4 years	10.9%	15.1%	-4.1%	-9.8%, 1.5%
5 years	13.1%	16.1%	-3.0%	-9.6%, 3.7%
Years of life lost (over 5 year horizon)			-0.18	-0.26, -0.13
Range: 150 – 250 cells	*(n=1293)*			
6 months	2.1%	2.1%	-0.0%	-2.6%, 2.6%
1 year	3.7%	5.2%	-1.5%	-5.7%, 2.7%
2 years	6.7%	10.9%	-4.2%	-10.4%, 1.9%
3 years	9.8%	12.9%	-3.1%	-10.4%, 4.2%
4 years	11.4%	14.6%	-3.2%	-11.2%, 4.9%
5 years	12.5%	16.1%	-3.6%	-12.9%, 5.7%
Years of life lost (over 5 year horizon)			-0.14	-0.20, -0.10
Range: 175 – 225 cells	*(n=623)*			
6 months	2.9%	1.2%	1.6%	-2.5%, 5.7%
1 year	5.7%	4.1%	1.5%	-5.4%, 8.4%
2 years	7.4%	10.4%	-3.0%	-12.0%, 6.1%
3 years	8.3%	14.5%	-6.2%	-16.5%, 4.1%
4 years	9.0%	17.5%	-8.5%	-20.2%, 3.2%
5 years	9.5%	19.9%	-10.4%	-24.5%, 3.8%
Years of life lost (over 5 year horizon)			-0.21	-0.27, -0.14

Note: Each panel was estimated in a separate flexible parametric survival model with four knots in the spline of log-time (the 175-225 model used 2 knots due to non-convergence with 3 or 4 knots). All models control for separate linear terms in earliest CD4 count on either side of the threshold. The range restriction is equivalent to local linear regression with a rectangular kernel.

Table S2. Treatment eligibility and all-cause mortality: flexible parametric models, <u>controlling for higher order polynomials</u> in earliest CD4 count.

Linear				
Time since first CD4+ count (*t*)	**Cumulative probability of death (1 - survival)** $F[t\|CD4 \uparrow 200]$	$F[t\|CD4 \downarrow 200]$	*Difference in F(t)*	*95% CI*
6 months	3.1%	3.6%	-0.5%	-2.7%, 1.7%
1 year	4.6%	6.7%	-2.1%	-4.9%, 0.7%
2 years	6.7%	10.8%	-4.1%	-7.8%, -0.5%
3 years	8.7%	13.4%	-4.6%	-8.7%, -0.6%
4 years	10.7%	15.2%	-4.5%	-9.2%, 0.1%
5 years	12.5%	16.8%	-4.3%	-9.7%, 1.2%
Years of life lost (over 5 year horizon)			-0.18	-0.26, -0.12
Quadratic				
6 months	3.1%	3.6%	-0.5%	-3.1%, 2.1%
1 year	4.6%	6.7%	-2.1%	-5.7%, 1.5%
2 years	6.7%	10.8%	-4.1%	-9.2%, 0.9%
3 years	8.7%	13.4%	-4.6%	-10.6%, 1.3%
4 years	10.7%	15.3%	-4.6%	-11.5%, 2.3%
5 years	12.5%	16.8%	-4.3%	-12.2%, 3.6%
Years of life lost (over 5 year horizon)			-0.18	-0.27, -0.12
Cubic				
6 months	3.1%	3.5%	-0.4%	-3.4%, 2.6%
1 year	4.6%	6.4%	-1.8%	-6.2%, 2.6%
2 years	6.6%	10.3%	-3.7%	-10.1%, 2.8%
3 years	8.7%	12.7%	-4.1%	-11.8%, 3.7%
4 years	10.6%	14.5%	-3.9%	-12.9%, 5.1%
5 years	12.4%	16.0%	-3.6%	-13.8%, 6.6%
Years of life lost (over 5 year horizon)			-0.15	-0.25, -0.09
Quartic				
6 months	2.8%	3.7%	-0.9%	-4.2%, 2.3%
1 year	4.2%	6.7%	-2.5%	-7.6%, 2.6%
2 years	6.3%	10.6%	-4.3%	-11.8%, 3.2%
3 years	8.2%	13.1%	-4.9%	-14.1%, 4.2%
4 years	9.9%	15.1%	-5.2%	-15.8%, 5.3%
5 years	11.4%	16.7%	-5.4%	-17.2%, 6.5%
Years of life lost (over 5 year horizon)			-0.20	-0.38, -0.10

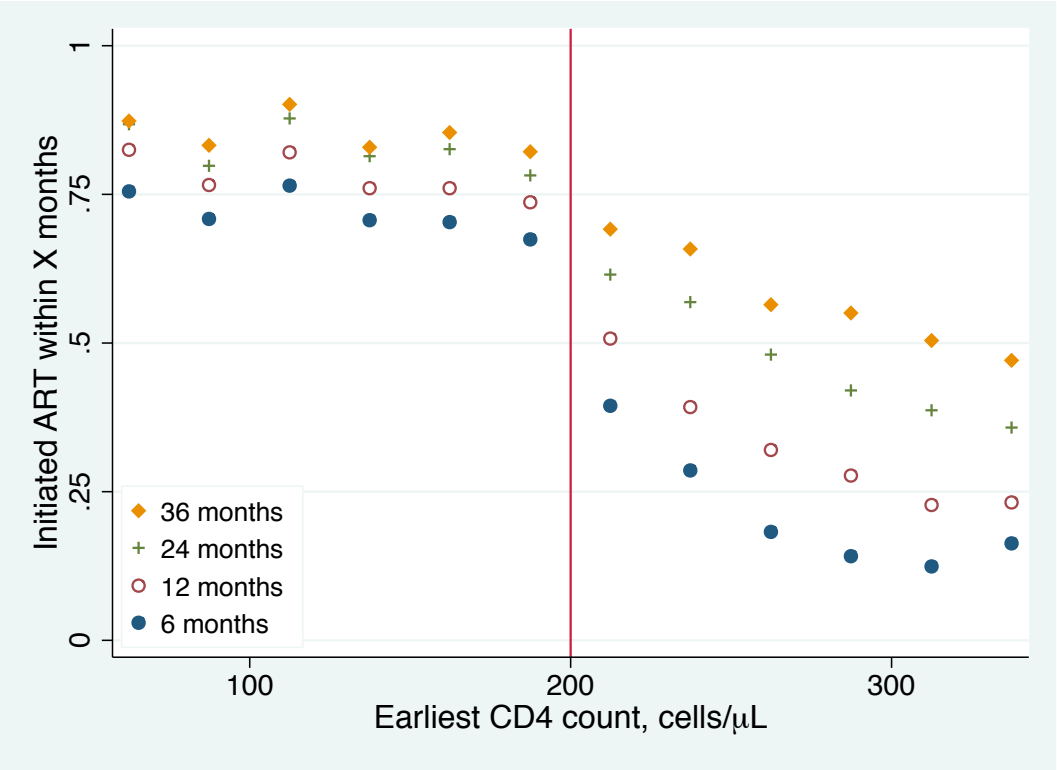
Note: Each panel is estimated in a separate flexible parametric survival model with four knots in the spline of log-time and four knots in the time-varying effect of covariates. The linear effect of earliest CD4 count was allowed to vary over time; higher order polynomial terms were modeled as time-invariant, proportional effects. Separate polynomial functions of earliest CD4 count were included on either side of the threshold. Models were estimated for patients presenting with CD4 counts of 50-350 cells; n=3710.

Table S3, Treatment eligibility and survival: hazard regression results.

	Earliest CD4+ count	**(a) Exponential**		**(b) Weibull**		**Sample**	
	Range, frailty distribution	HR_{RD}	*95% CI*	HR_{RD}	*95% CI*	*N*	*Deaths*
All-cause mortality							
(1)	0-350	0.59	(0.42, 0.83)	0.62	(0.44, 0.87)	4391	820
(2)	50-350	0.65	(0.45, 0.94)	0.67	(0.46, 0.96)	3710	539
(3)	100-300	0.66	(0.42, 1.04)	0.67	(0.43, 1.06)	2557	331
(4)	150-250	0.68	(0.35, 1.32)	0.71	(0.37, 1.36)	1293	153
(5)	175-225	0.54	(0.21, 1.41)	0.54	(0.21, 1.42)	623	73
(6)	0-350, Gamma	0.44	(0.25, 0.75)	0.43	(0.24, 0.77)	4391	820
(7)	0-350, Inv. Gaussian	0.44	(0.25, 0.78)	0.49	(0.30, 0.80)	4391	820
(8)	50-350, Gamma	0.45	(0.24, 0.84)	0.41	(0.19, 0.85)	3710	539
(9)	50-350, Inv. Gaussian	Did not converge		0.52	(0.29, 0.92)	3710	539
Non-HIV-related mortality							
(10)	0-350	0.94	(0.39, 2.26)	0.96	(0.40, 2.33)	4391	115
(11)	0-350, Inv. Gaussian	0.92	(0.22, 3.95)	0.93	(0.25, 3.45)	4391	115
HIV-related mortality							
(12)	0-350	0.58	(0.39, 0.86)	0.61	(0.41, 0.90)	4391	640
(13)	0-350, Inv. Gaussian	0.43	(0.23, 0.82)	0.48	(0.27, 0.84)	4391	640

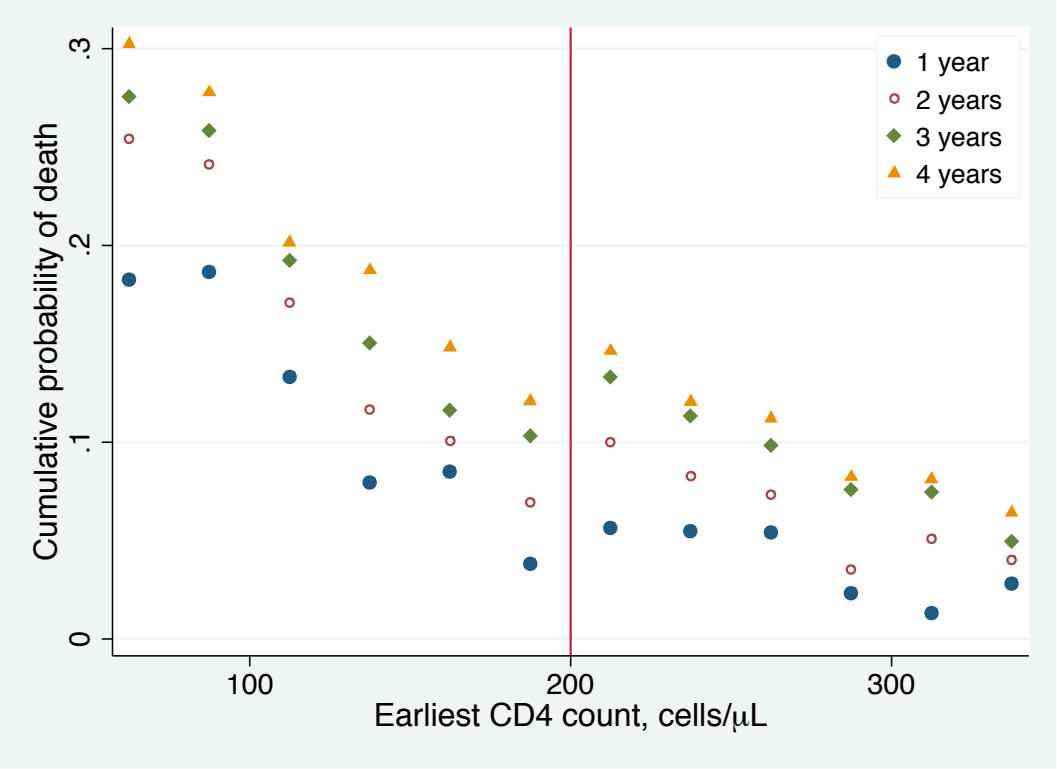
Each hazard ratio is estimated in its own regression. Models control for separate linear terms in earliest CD4+ count on either side of the threshold. Models in rows 6-9 present hazard ratios conditional on individual-level frailties (random effects). Models in rows 10-13 display hazard models for HIV-related and non-HIV related mortality. Data on cause of death were available through 2011.

Figure S1. Baseline CD4+ count and probability of ART initiation within X months, among survivors



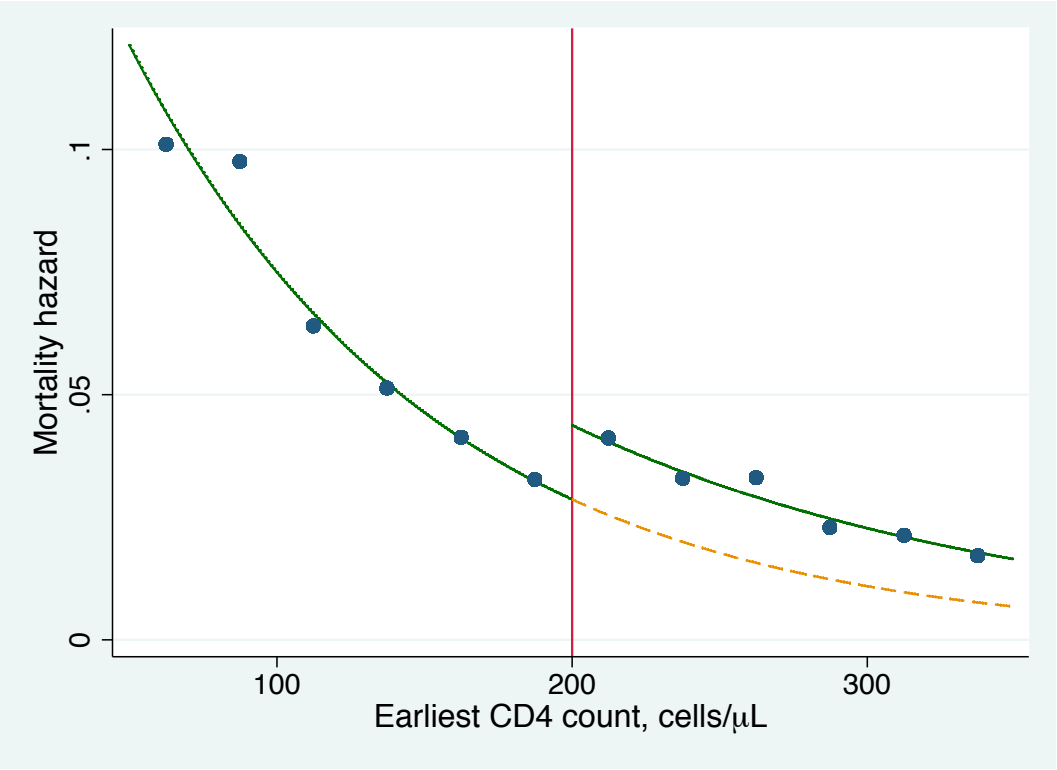
Note: Kaplan-Meier estimates of the probability of initiation among survivors. Follow-up time was censored at date of death or last survey visit.

Figure S2. Baseline CD4+ count and probability of death within X months



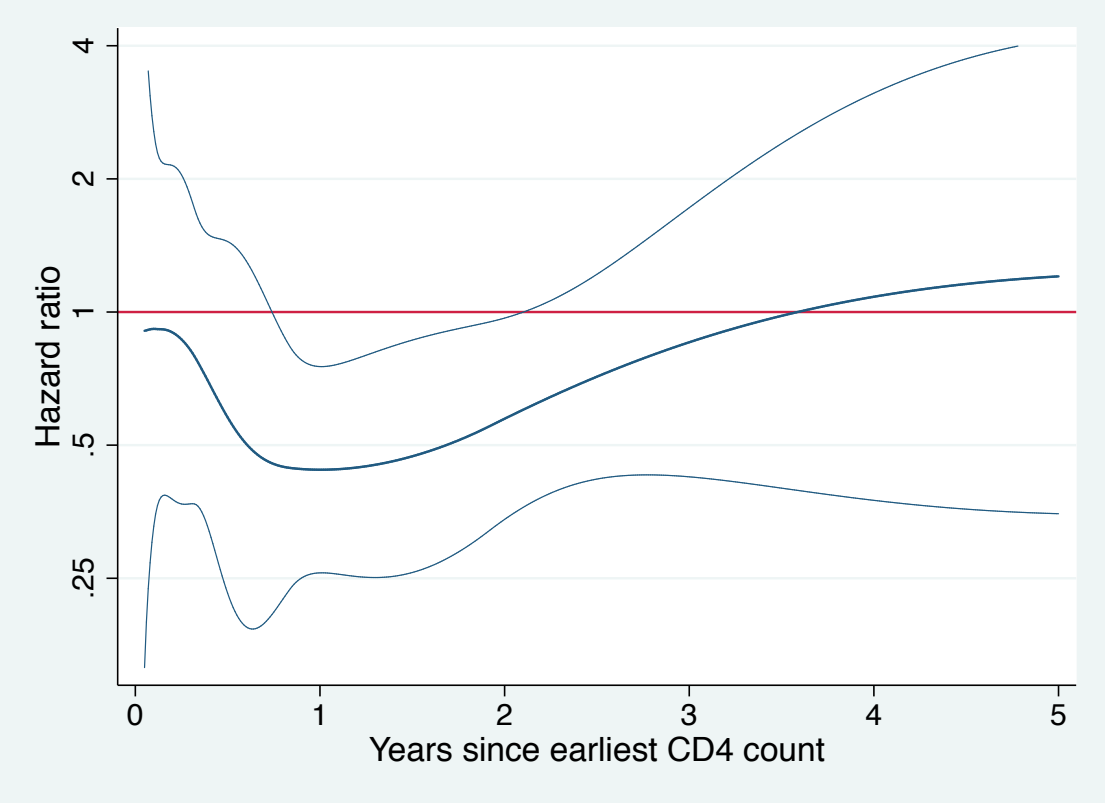
Note: Kaplan-Meier estimates of the probability of death. Follow-up time was censored at date of death or last survey visit.

Figure S3. Baseline CD4+ count and hazard of death



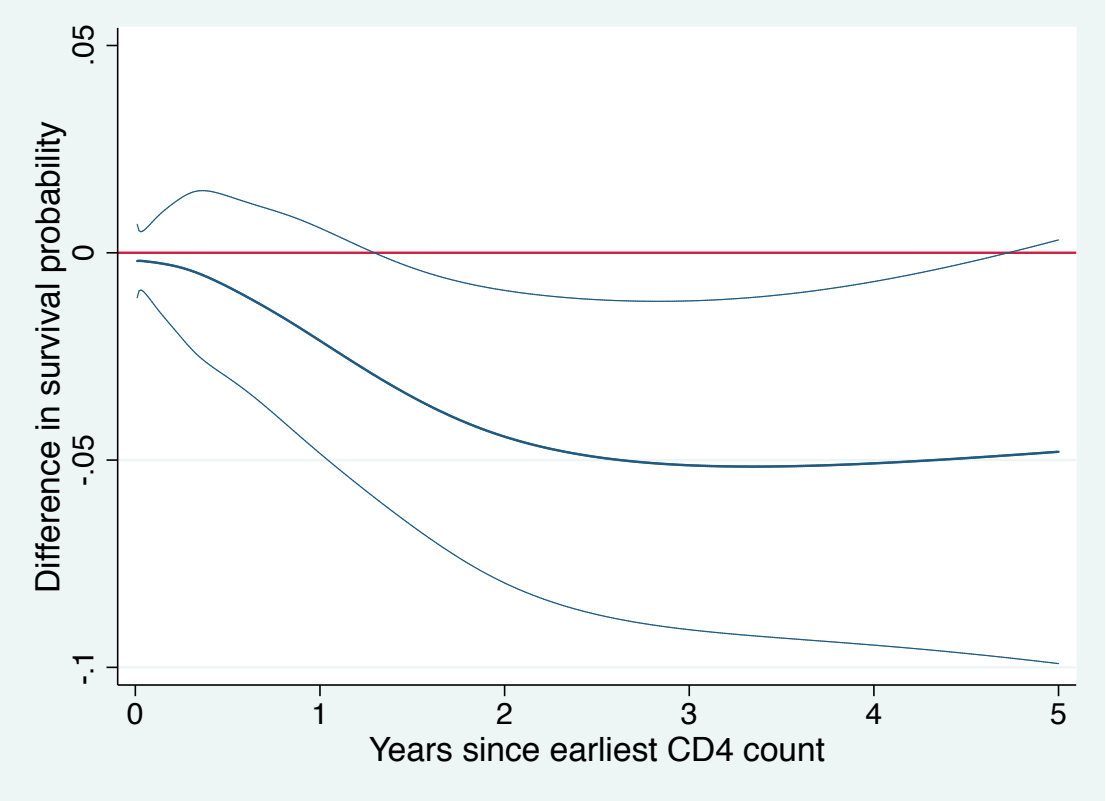
Predictions from exponential hazard model.

Figure S4. Baseline CD4+ count and mortality hazard ratio, with 95% CI



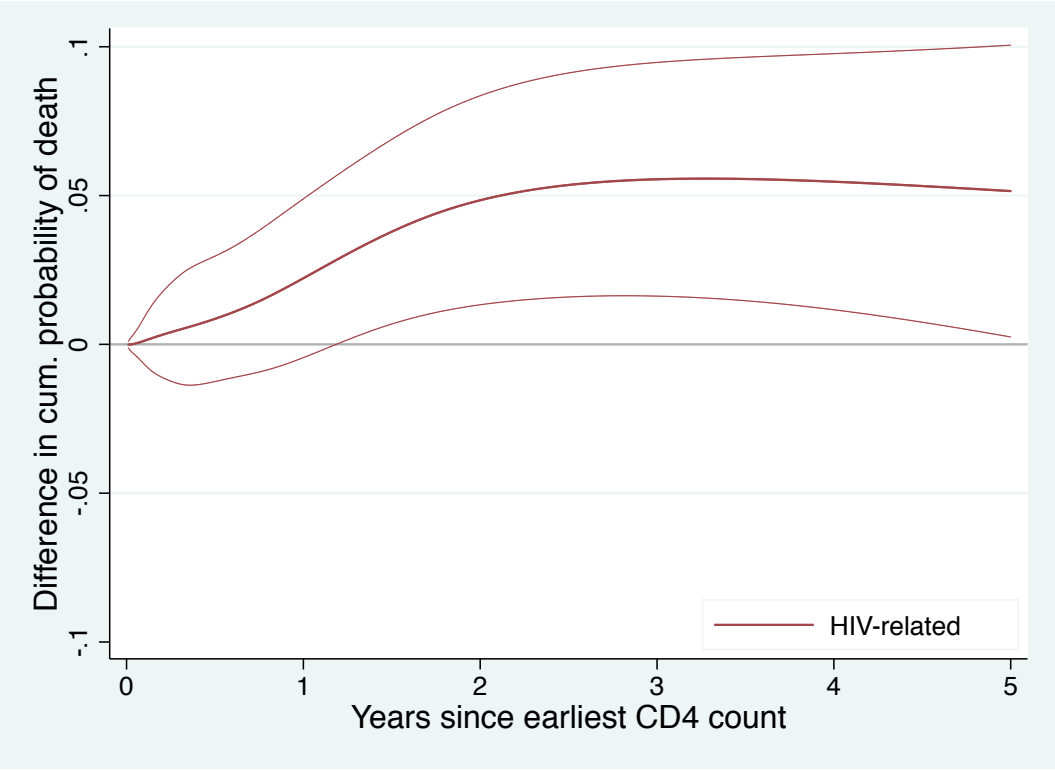
Predictions from flexible parametric survival model.

Figure S5. Baseline CD4+ count and survivorship gap, with 95% CI



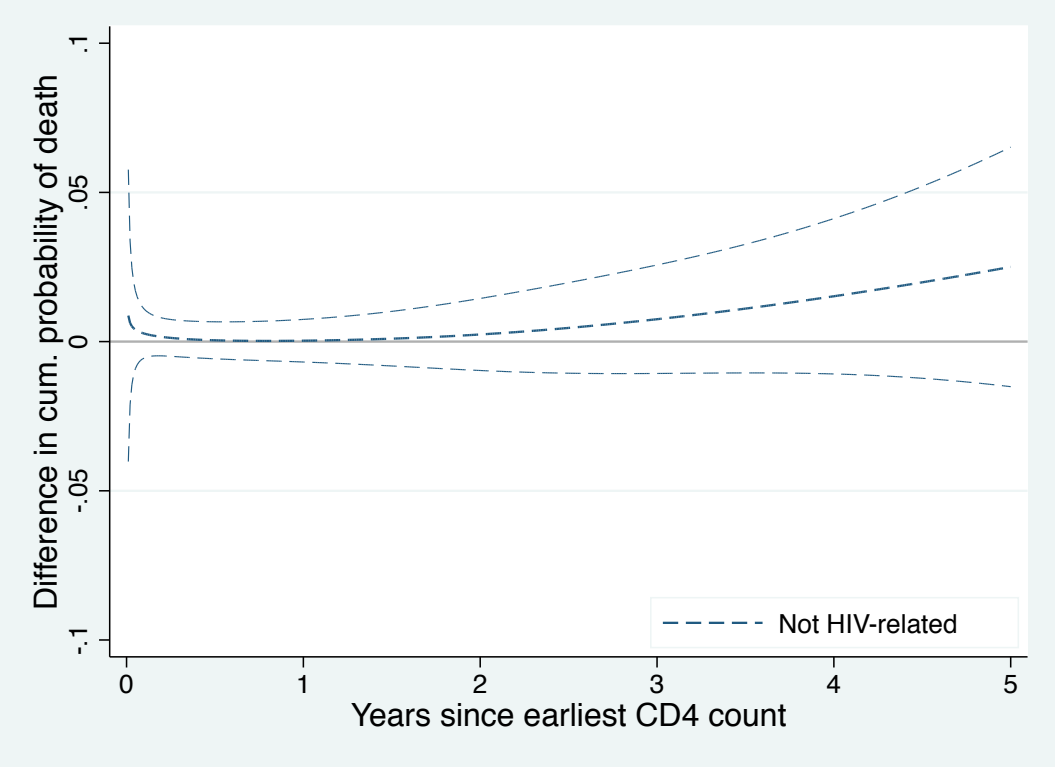
Predictions from flexible parametric survival model.

Figure S6. Excess mortality among non-eligible patients, HIV-related causes



Calculated as one minus the difference in survival, as predicted in flexible parametric survival models of time to HIV-related death.

Figure S7. Excess mortality among non-eligible patients, HIV-related causes



Calculated as one minus the difference in survival, as predicted in flexible parametric survival models of time to HIV-related death. Due to small number of deaths, model would not converge with 5 knots; instead used 4 knots for baseline log-cum-hazard and 3 knots for time-varying effect.

Figure S8. Predicted CD4 counts for patients presenting at the 200-cell threshold: mixed effects model, to adjust for missing data

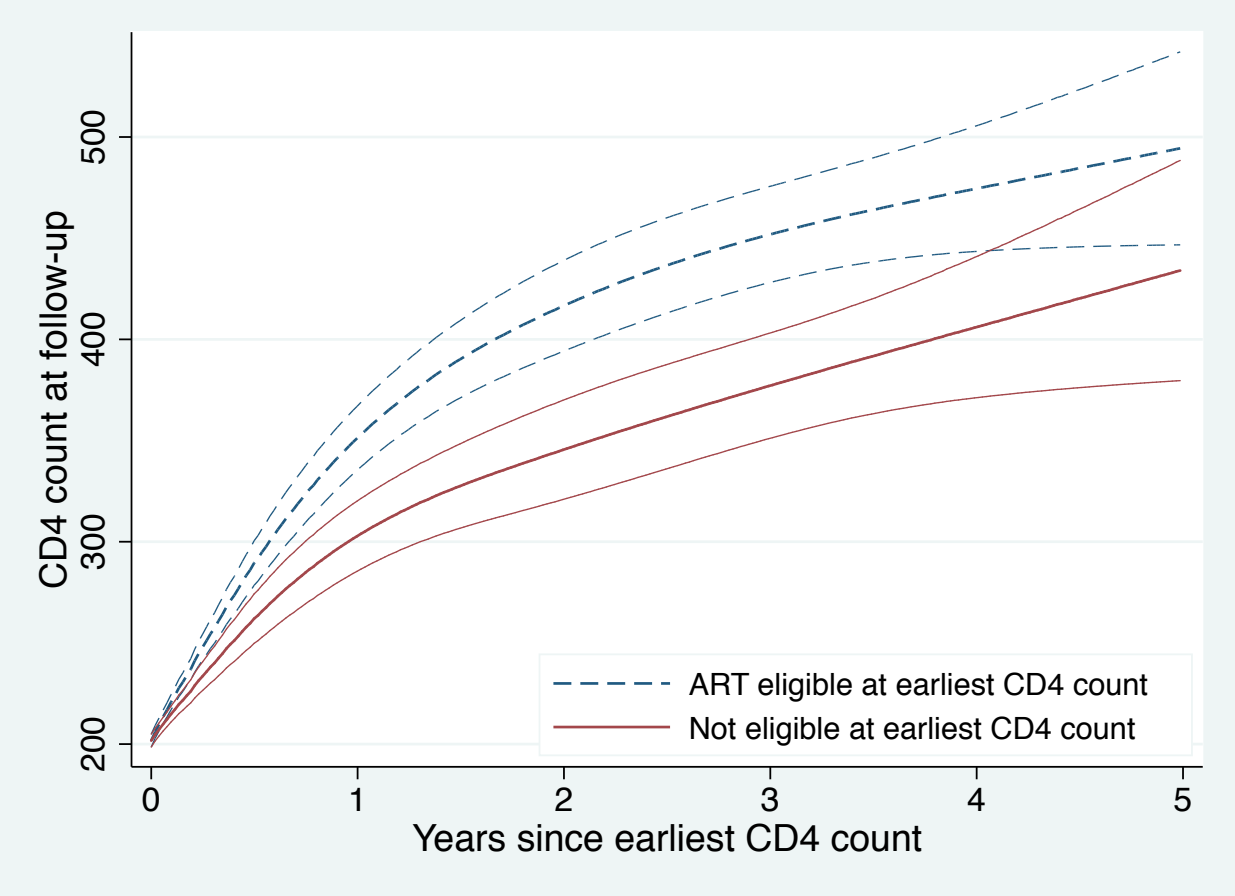


Figure displays predicted mean CD4 counts over time for patients presenting with an initial CD4 count just below (eligible) vs. just above (not eligible) the 200-cell threshold. Linear mixed-effects regression-discontinuity models were estimated with the effect of time modeled as a cubic spline, and interacted with the regression discontinuity coefficient and linear terms on either side of the discontinuity. Patients presenting with CD4 counts between 100 and 300 cells were included. The model was estimated based on data from survivors retained in care; follow-up was censored at the date of a patient's last CD4 count. 95% confidence bands are shown. This model differs from Figure 9 in that by modeling random intercepts and random time-varying effects, the model adjusts for any missingness that is correlated with a patient's CD4 count history.