WHEN TO START HIV TREATMENT: EVIDENCE FROM A REGRESSION DISCONTINUITY STUDY IN SOUTH AFRICA

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

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

PAA 2015 SUBMISSION

September 24, 2014

ABSTRACT: Early initiation of antiretroviral therapy (ART) may improve survival for people infected with HIV. To date, no experimental or quasi-experimental evidence exists on the survival impact of early vs. delayed ART in sub-Saharan Africa. We estimate the causal effects and cost-effectiveness of early ART in a real world setting. Using a quasi-experimental regression discontinuity design, we analyze data on 4391 HIV patients from a large demographic surveillance site in rural South Africa. Like many clinical therapies, ART is assigned based on a threshold rule, with eligibility determined by a patient's CD4+ cell count being below a cut-off. Regression discontinuity treatment effects were estimated 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 a CD4+ count just below 200 cells/μL were 4.3% points (95% CI 0.6, 8.0) more likely to be alive at two years than patients presenting with a CD4+ count just above the cut-off, an advantage that persisted at five years. These effects imply a 14.9% point two-year survival advantage for patients who actually initiated ART because they had an eligible CD4+ count. Large, persistent gains in clinical retention and immune function were also observed among patients who were ART eligible. The additional medical care provided to ART-eligible patients implies a cost of $1967 per life year saved to treating patients with CD4+ counts close to 200-cells. *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, David Canning, Gabriel Chodorow-Reich, Matthew Fox, Miguel Hernan, Guido Imbens, Portia Mutevedzi, Kevi Naidu, Marie-Louise Newell, Deenan Pillay, Heidi Williams, and seminar participants at the International Workshop on HIV/AIDS Observational Databases, Atlantic Causal Inference Conference, University of Witwatersrand, and Boston University. All errors and omissions are our own.

I. INTRODUCTION

Antiretroviral therapy (ART) slows disease progression and improves survival for people infected with HIV (Egger et al. 1997; Palella et al. 1998). However, there has been much debate about when in the progression of HIV disease patients should start ART, which once initiated must be taken for the rest of a patient's life (Lane 2003). Timing of ART initiation may have implications for morbidity, mortality, treatment side effects, quality of life, health-related expenditures, and economic productivity. Earlier ART may also have population health implications with reductions in HIV transmission to HIV-negative partners (Cohen et al. 2011; Tanser et al. 2013) and increased potential for development of drug-resistant strains of HIV. Information on the costs and benefits of earlier treatment initiation has implications for clinical practice and resource allocation decisions.

This paper provides the first causal evidence on the survival benefits of early vs. delayed ART in sub-Saharan Africa. To identify causal effects, we use a regression-discontinuity design, exploiting the threshold rule used to determine treatment eligibility during the period of study. Data come from a large demographic surveillance site in rural South Africa, which has been linked at the individual level with clinical records from the public sector ART program that serves the region. We find very large effects: patients eligible for ART were 4.3% points more likely to be alive at 2 years than patients who were ineligible at baseline – a 39% relative reduction in mortality risk. This difference in survival persisted five years later. These effects imply survival gains of nearly 15 percentage points for patients who were initiated on ART because they were eligible vis-à-vis patients who were barred from initiating because they were ineligible. Over a five-year horizon, early eligibility saved 0.18 years of life at a cost of $1967 per year of life.

There has been much debate in the economic literature on the causal effect of medical care on health (Grossman 1972; Card & Dobkin 2009; Doyle 2005; Finkelstein et al. 2012; Baicker et al. 2013).

Although therapeutic medicine played a small role in historical advances in life expectancy vis-à-vis nutrition (Fogel 2004) and sanitary interventions (Cutler & Miller 2005), its contribution increased with technological advances in antimicrobials, emergency medicine, and clinical management of chronic illness (Cutler, Deaton, & Lleras-Muney 2006). Diffusion of medical technology has figured prominently in longevity gains in low and middle-income countries over the last thirty years. Perhaps the starkest example is the mass provision of ART for HIV in developing countries. In the 1990s, HIV-related mortality lowered life expectancy by almost twenty years in some of the hardest hit countries in southern Africa (WHO 2012). An effective combination of antiviral drugs went to market in 1996, but was prohibitively expensive for most of the world's HIV-infected population. Since the early 2000s, however, falling drug prices, donor support, and bulk procurement have enabled widespread public sector provision of ART in many developing countries. ART scale-up has led to gains in labor supply and productivity (Thirumurthy et al. 2008; Habyarimana et al. 2010; McLaren 2010; Bor et al. 2012); improvements in measures of household wellbeing (d'Adda et al. 2009; Graff Ziven et al. 2009; Bor et al. 2012; Lucas & Wilson 2013); and some evidence of spillover effects on HIV infection (Friedman 2013; Tanser et al. 2013) and human capital investment (Baranov & Kohler 2013). Most importantly, scale-up of ART has led to large increases in longevity with gains in population adult life expectancy of more than a decade in some HIV-endemic regions (Bor et al. 2013). Existing evidence points to large benefits of ART vis-à-vis a world without ART. However, causal evidence to inform the timing of ART initiation for HIV patients in sub-Saharan Africa is lacking. Information about the marginal benefits of earlier ART is critical for further improvements in the effectiveness and efficiency of HIV treatment programs, which are financed primarily through public and donor funds.

Clinical and health policy decisions require evidence on the 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 (for an exception, see Almond et al. 2010).

Observational studies are vulnerable to the endogenous selection of patients into treatment options. Although randomization solves this problem, medical ethics place constraints on the types of information that clinical RCTs can provide. First, the standard of informed consent implies opt-in participation for most trials, which leads to non-representative experimental samples. Second, both treated and control patients are monitored carefully for adverse events, raising the potential for Hawthorne effects (Landsberger 1958) and implying that controls rarely receive true standard of care (Severe et al. 2011). Most RCTs do not provide evidence on the real world effectiveness of clinical interventions implemented at scale. Third, medical ethics dictate that a trial can only be started if clinical equipoise can be established, i.e. if there is "genuine uncertainty within the expert medical community… regarding the comparative therapeutic merits of each arm of the trial" (Freedman 1987); similarly, clinical trials must be stopped if intermediate results indicate that the treatment is protective (or harmful) vis-à-vis the control condition at a predetermined level of statistical significance. Starting and stopping rules for clinical trials are governed by statistical tests of the direction of an effect, with no attention to its magnitude. Ethical requirements thus generate a paradox: the larger a treatment effect (and hence the more likely it is that a policy maker might care about it), the less precisely the effect can be estimated in an RCT and the less one can know about its generalizability across populations and institutional contexts. Imprecise and inaccurate estimates of effect magnitude can lead to errors in clinical decision-making and resource allocation, costing dollars and lives at the population level. Ethical constraints thus drive a wedge between what can be identified in an RCT and what is required to optimize medical and health-policy decision-making. This "epistemological gap" suggests an important role for quasi-experiments in determining the real world effectiveness of medical inputs.

Our study builds on Almond et al. (2010), which used a regression discontinuity design to evaluate the costs and benefits of neonatal intensive care provided to infants born below the threshold for "very low birth weight". Threshold decision rules are very common in clinical care, e.g. in the diagnosis of

diabetes, hypertension, and high cholesterol. The imprecision of laboratory measures leads to a "strong" RD design with a "local randomization" interpretation at the threshold. Tests occur immediately prior to diagnosis, in contrast to other examples such as age (Card & Dobkin 2009) and distance from an administrative boundary (Chen et al. 2013), which may have effects on outcomes unrelated to treatment assignment. Measurement error in lab results implies that the precise value of a diagnostic measurement has no direct implications for health except through changes in clinical care that follow from that measurement. Additionally, 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 technician (often off-site) and are thus less vulnerable to manipulation and heaping in the assignment variable (Almond et al. 2010; Barreca et al. 2011; Shigeoka & Fushimi 2014). In spite of many applications, few RD studies have been published in the medical literature (see Moscoe, Bor, Bärnighausen 2014 for a recent review, and Bor et al. 2014 for a recent example).

Threshold rules are particularly important in low-income and developing country settings where medical care is delivered largely by health workers with limited training, large patient loads, and who rely on standardized guidelines for care (WHO 2013). Evidence on "checklists" suggests that guidelines may yield superior outcomes to clinical judgment in well-resourced settings as well (Haynes et al. 2009). In Almond et al. (2010), very low birth weight diagnoses affected treatment decisions primarily in lower quality hospitals, which lacked capacity for continued observation of newborns in neonatal intensive care units. Our study evaluates a nurse-led HIV treatment program implemented through public-sector clinics in a poor, largely rural area of South Africa. In clinical settings such as this one, where clinicians rely on national guidelines, optimizing diagnostic and treatment thresholds will have large population health impacts. Provider behavior regarding when to adhere to guidelines is also of interest in its own right (Shigeoka & Fushimi 2014). We provide evidence (IN PROGRESS) on the types of patients that are prioritized for treatment by assessing when threshold rules are ignored.

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.

Section II provides background on the history of clinical guidelines for ART treatment and existing research on the topic of when to start ART. 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 describes the characteristics of patients treated according to the threshold rule and presents evidence on effect heterogeneity. Section VIII presents results on cost-effectiveness. Section IX concludes.

II. BACKGROUND

II.A. When To Start ART: Biological and Theoretical Considerations

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.) The question about "when to start" ART has largely been a question about the appropriate CD4 count threshold.

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

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 only 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), citing a "public health" approach (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 (ART) 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 earlier treatment is limited (WHO 2013). WHO itself cited "strong" evidence that early ART reduces HIV transmission, but 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).

Clinical 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 those 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.C. Rational For Study Design

In this study, we use a quasi-experimental regression discontinuity (RD) design to identify the causal effect of early vs. delayed ART initiation among HIV patients in rural South Africa. 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 (Bor, et al. 2013, Lee & Lemieux 2010, Imbens & Lemieux 2008, Campbell & Thistlewaite 1960).

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 between January 2007 and August 2011. Patients were eligible for ART if their CD4 count was below 200 cells/μL or if they had Stage IV AIDS illness, according to national guidelines during the study period. 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. We find a large, discontinuous change 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. 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 using novel methods to model survival times in the context of fuzzy RD.

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 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 enabled longitudinal follow-up of patient survival regardless of whether they were 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.

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. If the patient and provider decided to initiate ART, the patient was first required to attend a series of weekly treatment literacy and adherence counseling sessions, except in cases of medical emergency. 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 chosen 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, rendering the initial CD4 count obsolete (CONFIRM).

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, as measured in routine clinical monitoring of patients retained in care, and time to next clinic visit, a measure of retention in care. Costs were calculated by estimating the expected number of "years on ART", "years in pre-ART care", and "years not in care" 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).

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 observations are censored, e.g., survival times exceed the period of observation, models that accommodate censoring are required in order for these observations to be included. The absence of RD applications with non-linear and survival outcomes may be a barrier to uptake in the medical literature (Bor 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 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 intent-to-treat effect of being CD4-count eligible for observations presenting with CD4 counts close to 200 cells.

Equation 1

$$E\left[Y_i \mid CD4_i\right] = b_1(CD4_i - 200) + b_2 1\left[CD4_i < 200\right] + b_3(CD4_i - 200) * 1\left[CD4_i < 200\right]$$

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

Equation 2

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

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 is interacted with each of the terms from the regression discontinuity model on the right-hand side of Equation 1. The 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, estimated only in the area around the threshold. The causal effect of treatment eligibility on the difference scale is equal to:

Equation 3

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

This effect may vary over time, and if g(.) is a linear (identity) link function, is simply equal to $f\left(t \mid \mathbf{b}_2, k_2\right)$. The ratio of means $E\left[Y_i \mid t, CD4_i - 200\right] / E\left[Y_i \mid t, CD4_i \urcorner 200\right]$ is also identified due to the Slutsky 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, $\log\left(-\log[S(t \mid CD4_i)]\right) = \log\left(H(t \mid CD4_i\right)$. 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; and the time varying 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 (Almond et al. 2010). However, this approach discards data if some units are not followed up for the full interval. Further, if censoring times are random and 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 other applications in which censoring times are random, e.g. non-selective migration outside of the study area.) 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 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 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.) 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, "complier difference-in-survival (t)" LATE parameter.

A limitation for RDD is that Kaplan-Meier cannot accommodate continuous covariates, and estimation relies on local linear predictions at the threshold. FPSM offers a flexible parametric alternative, using restricted cubic regression splines to describe the baseline integrated hazard and the ratio of integrated hazards across covariates. Spline functions are approximations to the underlying true functional 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 functions subject to: i) linearity outside

[†] Consider the individual-specific hazard model $h_i \mid t, D_i, V_i, W_i, \gamma(t) = h_0(t) V_i \exp(\gamma(t) W_i D_i)$, 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 $\gamma(t)$ 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.

the outermost knots; and ii) continuous first and second derivatives at the knots (CITE). In practice, 3-5 knots placed at quantiles of the data are generally enough to describe most functions (Harrell 2001), and are almost certainly enough to describe functions that are monotonically increasing (e.g., log cumulative hazard as a function of time) or monotonically 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 RDD, 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 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. The Stata command `stpm2` was used for all flexible parametric survival analysis (Lambert & Royston 2009).

We report instantaneous 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.

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(.)* 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 determine the possible extent of any bias, we assessed the effect of treatment eligibility on clinical attrition in linear probability models. We then estimated a linear mixed effects model by maximum likelihood, which accounts 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 RDD models using quantile regression (Fransden et al. 2012), with the identical specification as the OLS model, and obtained predictions for the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, and 95^{th} 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. Clinicians and patients may be interested in another type of causal effect: the effect of early vs. deferred ART initiation on patients that initiated ART rapidly because their CD4 count was below 200. Under the assumption that rapid ART initiation is indeed the only pathway through which earlier ART eligibility would affect health (i.e., the *exclusion restriction* or *excludability assumption*), and the additional assumption of *monotonicity* (explained below), we can divide the intent-to-treat estimates by the "first stage" effect on ART initiation to obtain this effect. This is simply the Wald (ratio) instrumental variables (IV) estimator. This estimand is analogous to the *complier average causal effect* or *CACE* in a randomized experiment using an encouragement design; here, we denote the estimand as $CACE_{RDD}$ to signal that the estimate is identified at the threshold. The monotonicity assumption requires that no patient who would not have initiated ART if eligible would have initiated ART if ineligible and vice-versa; i.e., that no patient defied their treatment eligibility status. The usual IV assumption that eligibility is as good as randomly assigned is assured at the threshold by the nature of the study design. (This approach is sometimes known as a "fuzzy" regression discontinuity design, since treatment assignment is probabilistic, not deterministic.) $CACE_{RDD}$ is denoted by the following equation:

Equation 4

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

We obtained $CACE_{RDD}$ estimates of the effect of early vs. deferred ART on the probability 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 the average CD4 count among survivors retained in care.

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.

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)[‡].

Third, support for the validity of the study design can be found by assessing continuity in baseline observables. Figure 2 displays mortality hazards predicted as a function of sex, age, age-squared, sex-by-age interactions, and date of first CD4+ count. This figure is similar to a balancing 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).

[‡] We conducted a statistical test of continuity in the density function of earliest CD4 counts at 200 (McCrary 2008). 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. Figure 5 presents predicted probabilities of death at 1, 2, … and 5 years, based on a flexible parametric survival model, 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. At 6 months, there was no significant difference in the probability of death (Table 2, panel 1: risk difference = 0.3% points, 95% CI -2.0, 2.5). However, by 2 years, 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 in the first 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. 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, there was substantial excess mortality between about six months and three years. By three years, however, 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.

Note that these models seek to flexibly specify the shapes of the survival curves, without any assumption of proportional hazards. This means that a single meaningful hazard ratio cannot be obtained. However, other contrasts (of the survival curve) 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).

To assess the robustness of these results, we estimated flexible parametric survival models varying the bandwidth and functional form of earliest CD4 count. These results are presented in Tables S1 and S2.

Results were consistent with the results in Table 2. We also estimated more 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. 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). 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, 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).

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 RDD 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 5, 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 5, 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. Clinical Benefits of Rapid Treatment Initiation: Complier Causal Effects

Under plausible assumptions described above, we can assess the causal effect of rapid ART initiation (within six months) on survival and immune health 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 5 presents RDD 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 delayed 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.

VI. SENSITIVITY ANALYSES

Table S1 displays the results of FPSMs, reducing the window of data used in the analysis down to +/- 50 cells/μL. Table S2 displays results of FPSMs, including up to fourth-degree polynomials on either side of the threshold. In both cases, results are nearly identical, although some precision is lost with the narrower bandwidths and higher order functions.

VII. TREATMENT EFFECT HETEROGENITY

VII.A. Gender STILL TO DO

VII.B. Distance to Clinic STILL TO DO

VII.C. Clinic Size / Quality STILL TO DO

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 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 modeling 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 morbidity that we identify, these are lower-bound estimates on the clinical cost-effectiveness of raising the ART eligibility threshold from 200 cells/μL.

IX. CONCLUSION

This study assessed the health benefits of early vs. delayed ART treatment eligibility for patients presenting with CD4 counts close to 200-cell/μL threshold, using data from a public sector ART program in rural South Africa. ART eligibility at baseline had a large and statistically significant impact on both survival and immune health. These effects are large. 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. These gains in survival were attained at a modest cost of $1967 per life year saved. 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. JAIDS 2013). The divergence in both survival and immune health for eligible and ineligible patients occurred between about six months and two years. These effects appear to be permanent, with large gaps in both survival and immune health persisting five years after patients' first CD4 count.

These are the first quasi-experimental estimates of the survival benefits of early vs. delayed ART eligibility in sub-Saharan Africa. To date no experimental evidence exists and new trials are unlikely to be forthcoming given that current recommendations would make equipoise difficult to defend (WHO 2013). Guidelines for "when to start" ART have swung back and forth over the short history of highly active therapy (De Cock, El Sadr 2013). 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.

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 experiments. *Epidemiology.* In Press, 2014.

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.

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

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. *The Lancet infectious diseases*. 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.

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. (2008). 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.

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

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. *J Acquir Immune Defic Syndr* 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 in medicine, epidemiology, and public health: a review of current and best practice. 2014. *Journal of Clinical Epidemiology*, in press.

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).

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.

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.

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.

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

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

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.

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

Table 1, Treatment eligibility and rapid ART initiation

	Earliest CD4+ count	**Probability of initiating ART within six months**				**Sample**
	Specification; Range	$E[Y\|Z \downarrow c]$	$E[Y\|Z \uparrow c]$	*Difference*	*95% CI*	*N*
(1)	Linear; 0-350	0.36	0.67	0.31	(0.25, 0.37)	4113
(2)	Linear; 50-350	0.36	0.69	0.32	(0.26, 0.38)	3548
(3)	Linear; 100-300	0.42	0.66	0.24	(0.16, 0.31)	2471
(4)	Linear; 150-250	0.45	0.66	0.21	(0.10, 0.32)	1256
(5)	Linear; 175-225	0.44	0.65	0.21	(0.06, 0.37)	610
(6)	Quadratic; 0-350	0.47	0.67	0.20	(0.12, 0.29)	4113
(7)	Quadratic; 50-350	0.47	0.66	0.19	(0.10, 0.28)	3548
(8)	Quadratic; 100-300	0.47	0.68	0.21	(0.10, 0.33)	2471
(9)	Quadratic; 150-250	0.46	0.66	0.20	(0.04, 0.36)	1256
(10)	Quadratic; 175-225	0.60	0.77	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)** $F[t\|CD4 \downarrow 200]$	$F[t\|CD4 \uparrow 200]$	*Difference in F(t)*	*95% CI*
6 months	3.7%	3.4%	0.3%	2.5%, -2.0%
1 year	6.5%	4.6%	1.8%	4.7%, -1.1%
2 years	10.9%	6.6%	4.3%	8.0%, 0.6%
3 years	13.6%	8.8%	4.8%	9.0%, 0.6%
4 years	15.3%	10.7%	4.5%	9.2%, -0.1%
5 years	16.6%	12.6%	4.0%	9.5%, -1.5%
Years of life saved (over 5 year horizon)			0.18	0.12, 0.26

Time since first CD4+ count (*t*)	**Instantaneous hazard of death** $h[t\|CD4 \downarrow 200]$	$h[t\|CD4 \uparrow 200]$	*Hazard ratio*	*95% CI*
6 months	0.06	0.07	1.07	0.50. 2.30
1 year	0.06	0.01	0.24	0.10, 0.60
2 years	0.04	0.02	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	0.01	0.02	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.

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

Non-HIV/TB-related mortality

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

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. "Fuzzy RDD": the effect of "rapid ART initiation" on probability of death, years of life lost, 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		-3.4% (-6.0, -0.7)	10.6%
2 years		-4.8% (-8.2. -1.3)	14.9%
3 years		-5.0% (-9.2, -0.8)	15.5%
4 years		-4.2% (-9.0, 0.6)	13.0%
5 years		-3.1% (-9.1, 2.9)	9.6%
Years of life lost, over 5 year horizon		0.19 (0.14, 0.27)	0.59
Follow-up CD4 count:			
1 year		72 (43, 101)	224
2 years		77 (41, 114)	239
3 years		75 (35, 116)	233
4 years		73 (18, 128)	227
5 years		71 (-15, 157)	220

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, so selection is not a concern. 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 mixed effects models similar to Table 4. All models are estimated with linear terms on either side of the threshold, for patients presenting with CD4 counts of 50-350; n=3449.

Figure 1. Distribution of CD4+ counts at clinical enrollment

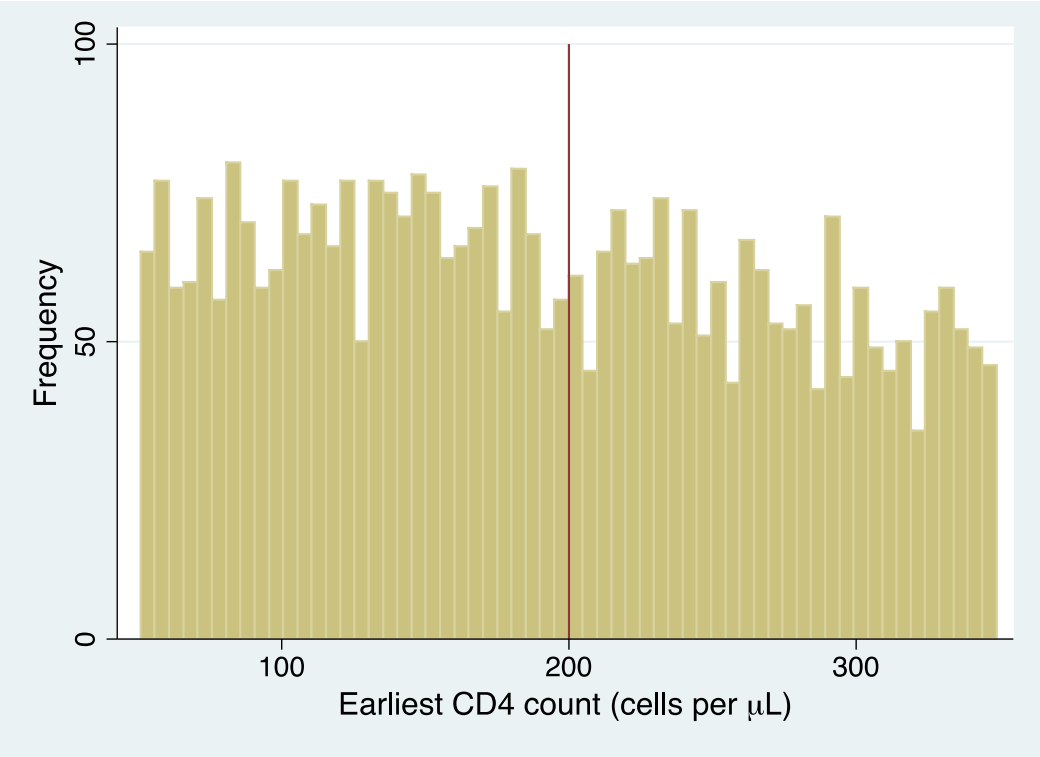
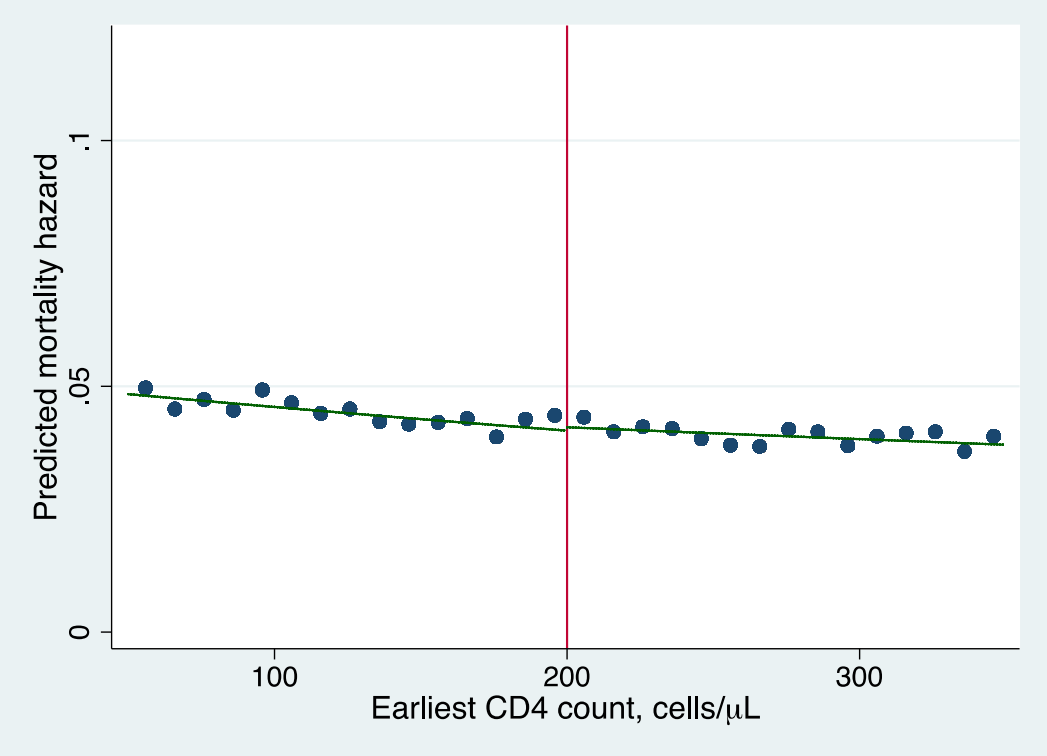
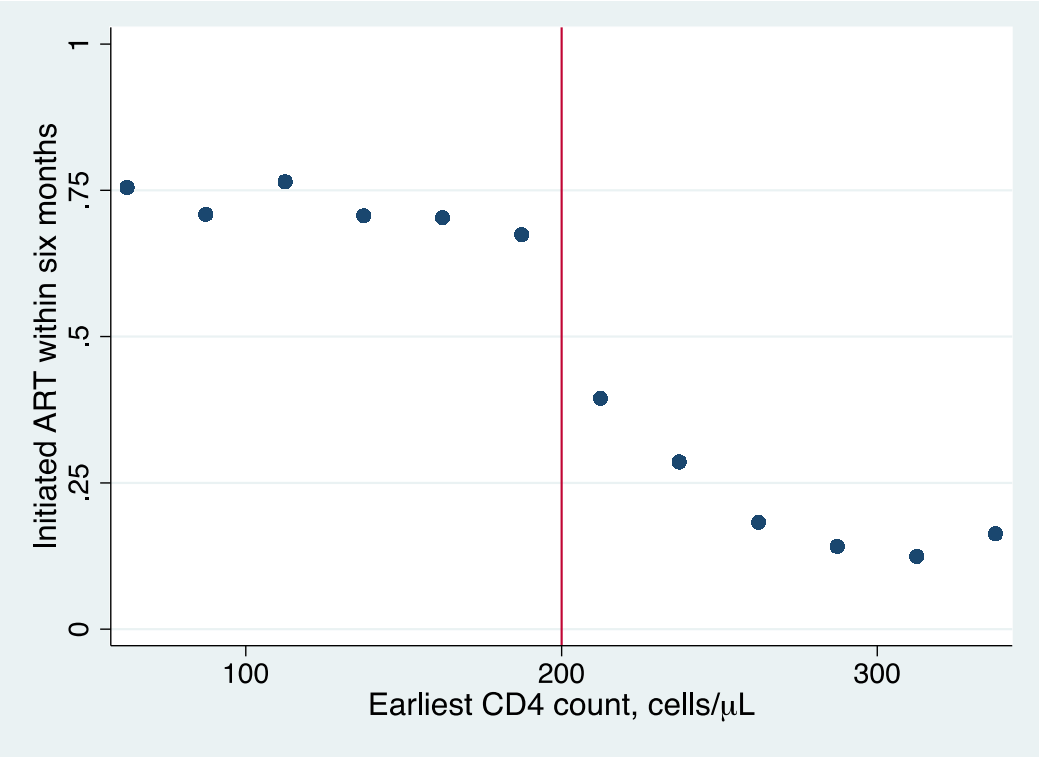


Figure 2. Mortality hazards predicted using baseline covariates



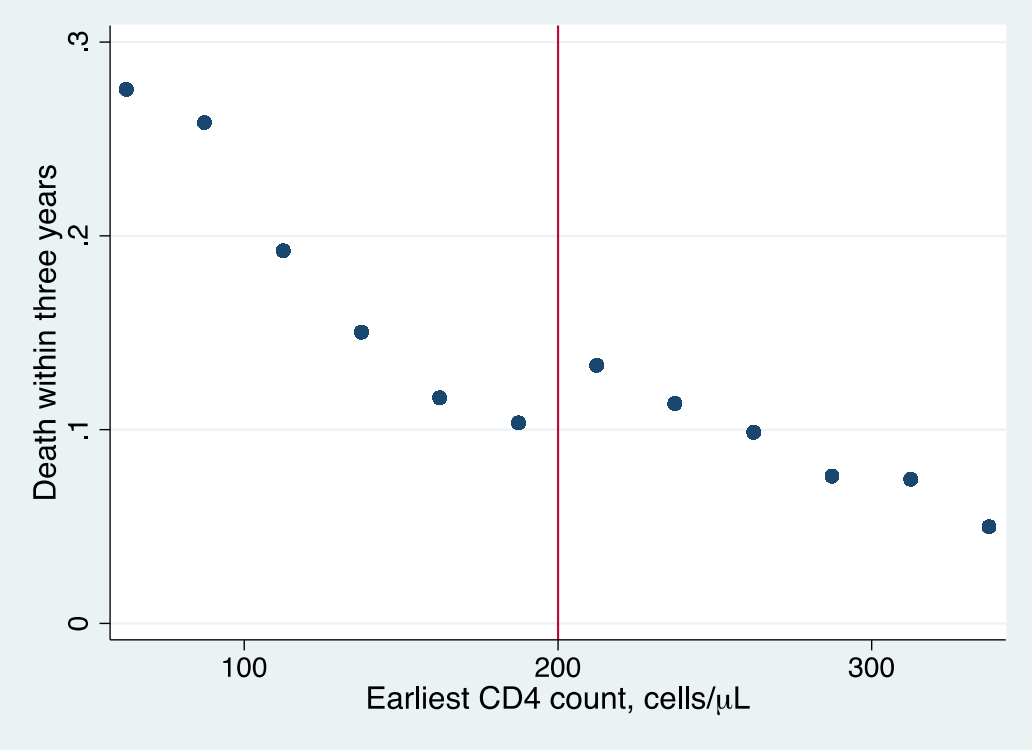
Notes: Figure plots mortality hazards predicted using sex and age at first CD4+ count. Log-hazards were predicted in an exponential regression model, controlling for sex, age, age^2, and their interactions. Geometric mean hazards are shown for 10-cell CD4+ count bins. Fitted lines were estimated by regressing the predicted log hazards on CD4 count, an indicator for CD4>200, and the interaction of these two variables, and then exponentiating the predictions.

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



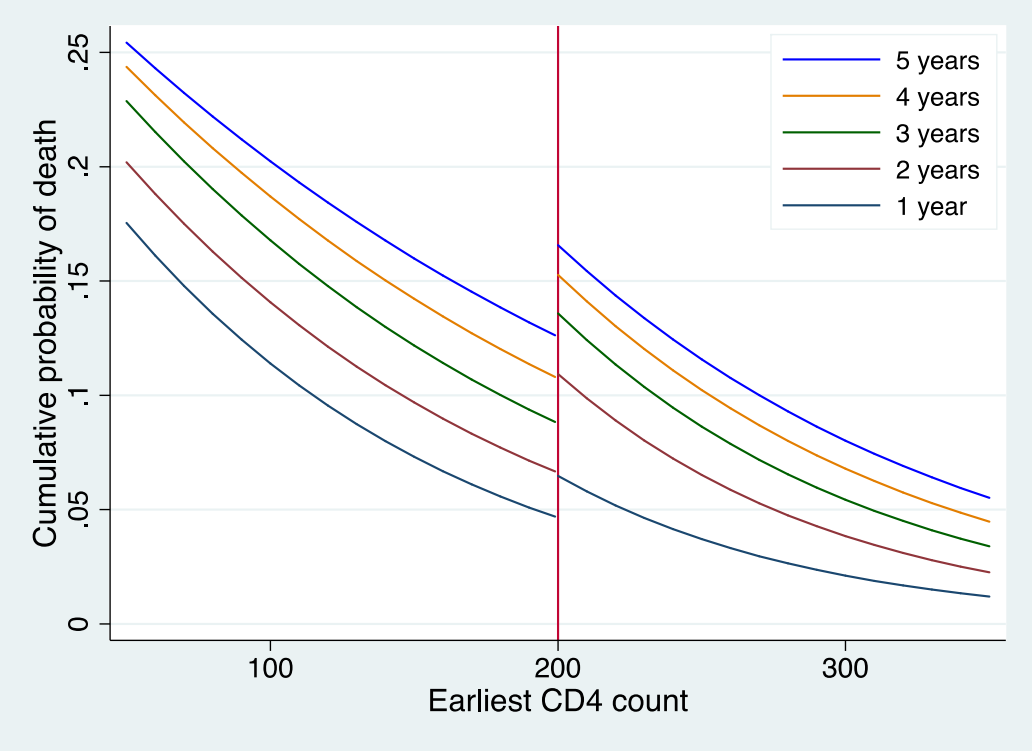
Notes: Kaplan-Meier estimates of probability that a patient initiated ART within X months of first CD4+ count in care. Follow-up time was censored at date of death or last survey visit.

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



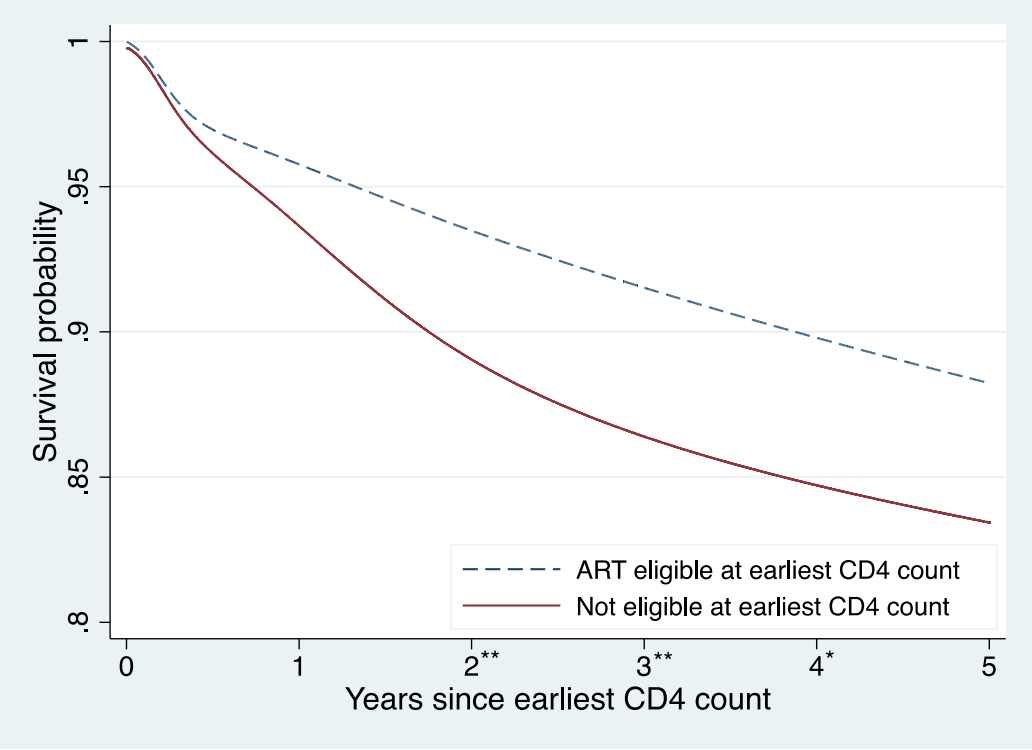
Notes: Kaplan-Meier estimates of probability that a patient initiated ART within X months of first CD4+ count in care. Follow-up time was censored at date of death or last survey visit.

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



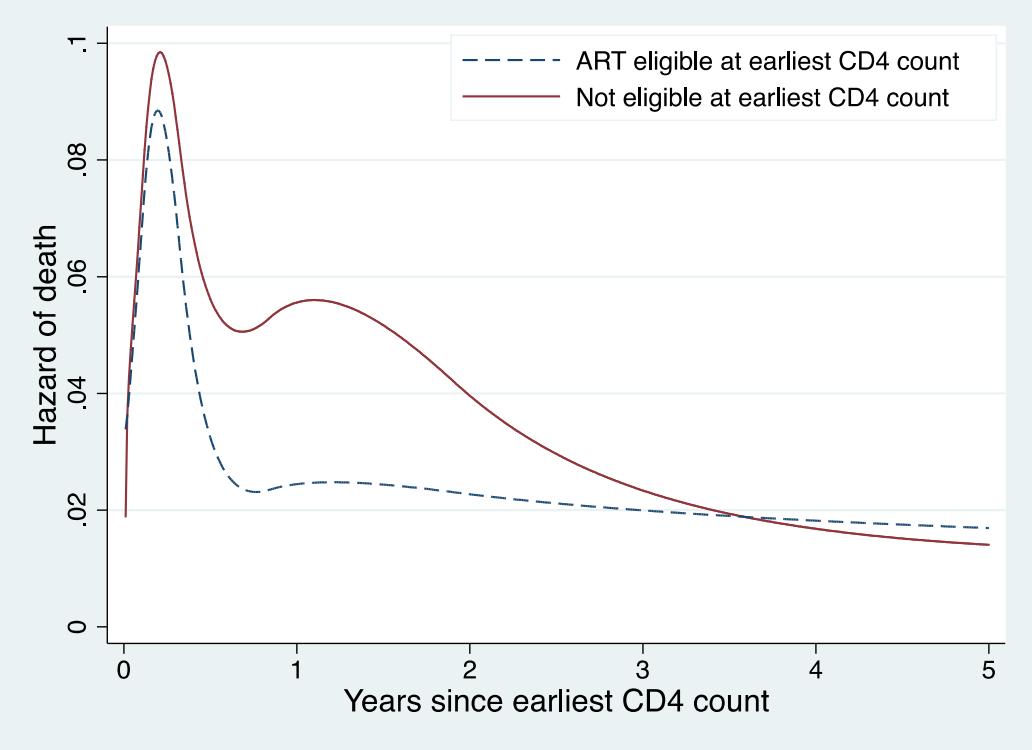
Notes: Predicted probabilities of death within 1, 2,…,5 years based on flexible parametric survival model, estimated for range 50-350 CD4 cells.

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

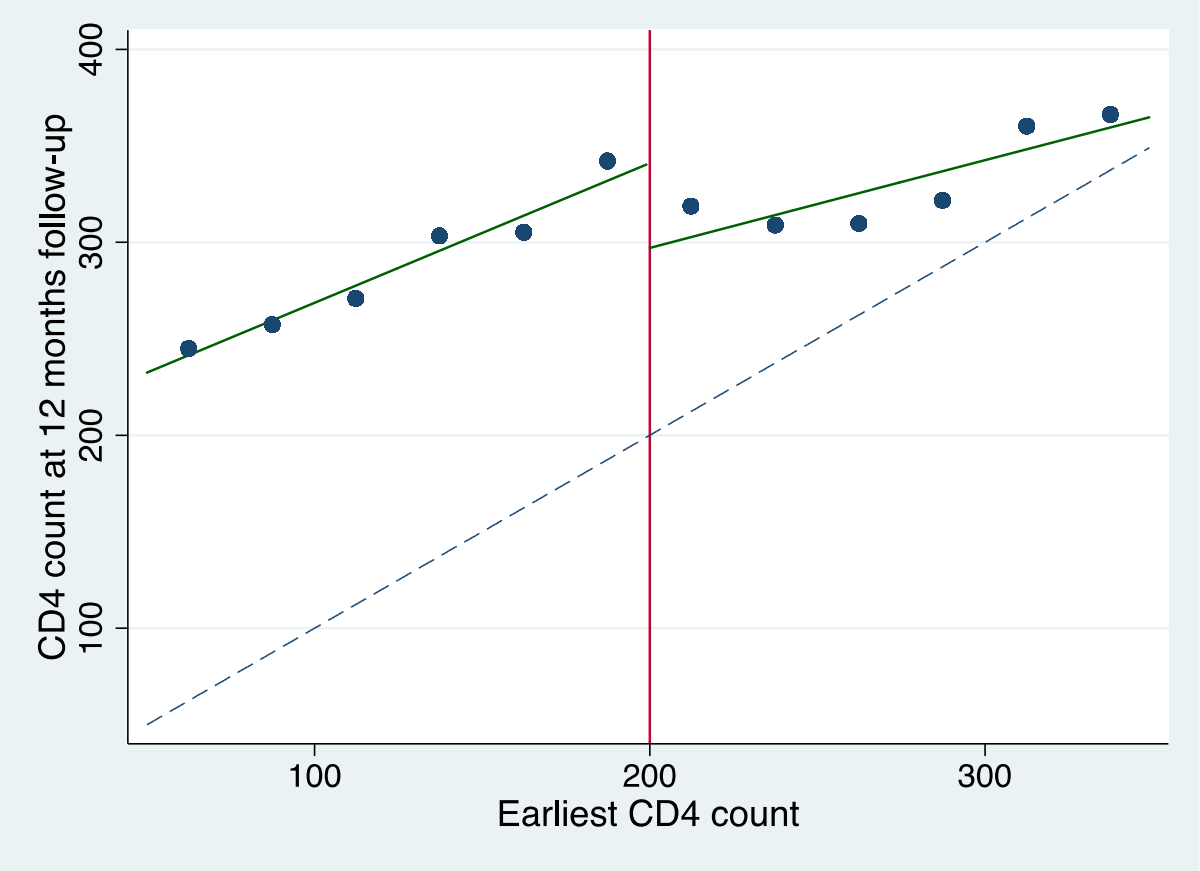


Figure displays mean CD4 counts for 1753 (of 4391) patients with follow-up CD4 counts between 9 and 15 months follow-up. For patients with multiple CD4 counts in this interval, the test date closest to 12 months was retained. Dotted line is the 45^{o} line.

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

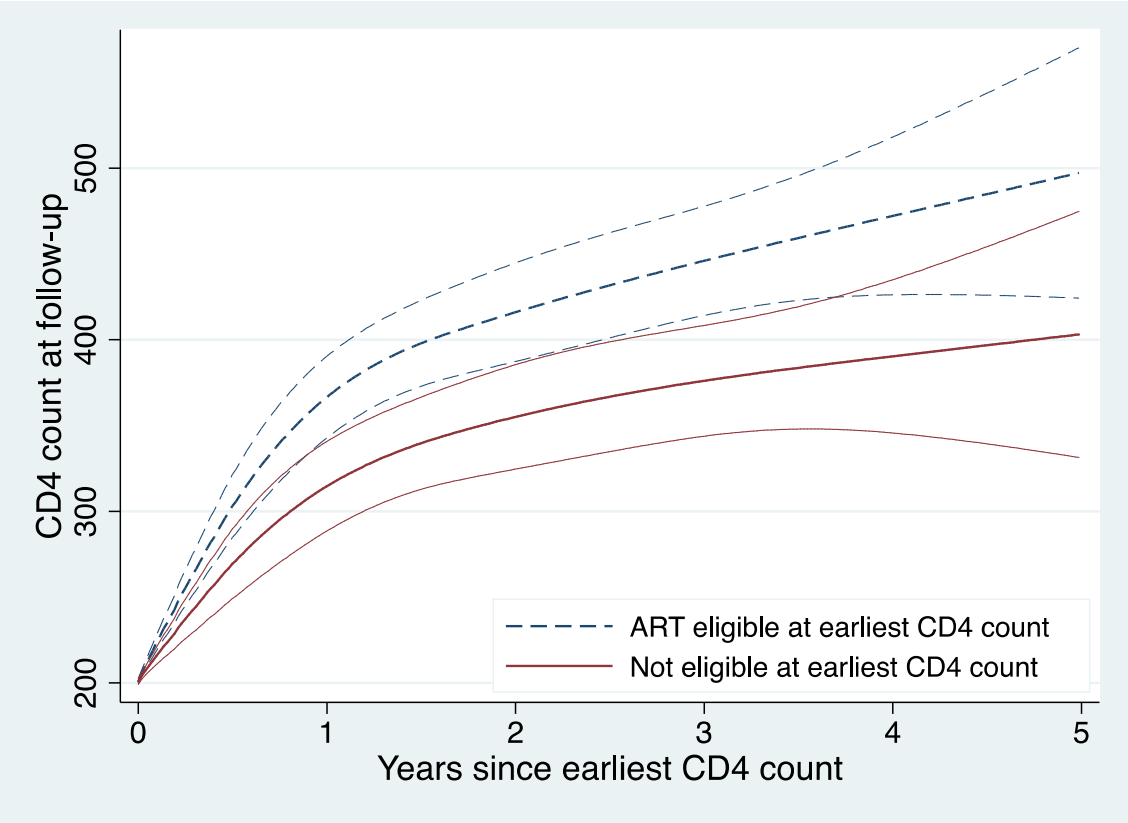


Figure displays predicted mean CD4 counts over time for patients presenting with an initial CD4 count just below (eligible) vs. just above (not eligible) the 200-cell threshold. Linear regression-discontinuity models were estimated with the effect of time modeled as a cubic spline, and interacted with 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

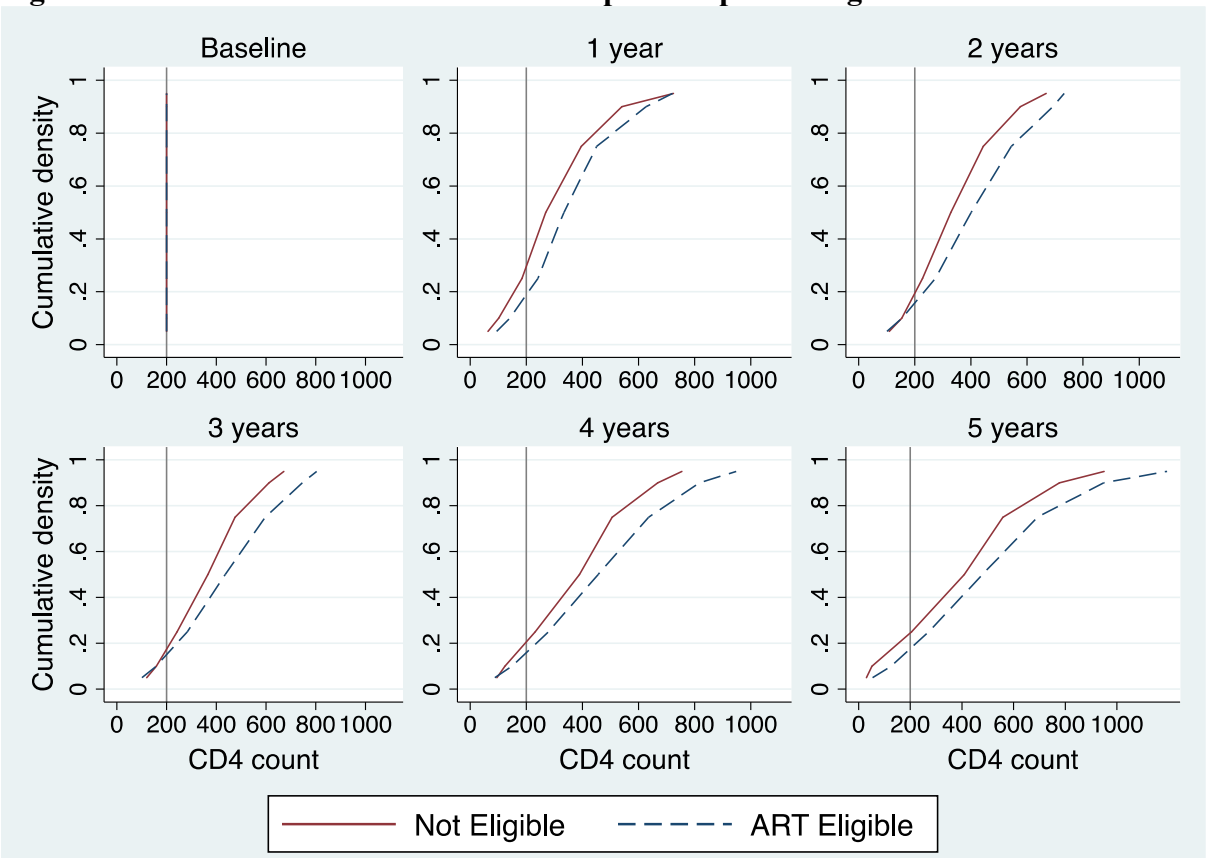


Figure displays cumulative density functions of CD4 counts at baseline and 1, 2, …, and 5 years follow-up, for patients presenting with an initial CD4 count just below (eligible) vs. just above (not eligible) the 200-cell threshold. CDFs are constructed as 5th, 10th, 25th, 50th, 75th, 90th, and 95th 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.

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

Note: Each panel is estimated in a separate flexible parametric survival model with four knots in the spline of log-time. All models control for separate linear functions of earliest CD4 count on either side of the threshold.

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

Linear

Time since first CD4+ count (*t*)	Cumulative probability of death (1 - survival) $F[t\|CD4 \downarrow 200]$	$F[t\|CD4 \uparrow 200]$	*Difference in F(t)*	*95% CI*
6 months	3.6%	3.1%	0.5%	2.7%, -1.7%
1 year	6.7%	4.6%	2.1%	4.9%, -0.7%
2 years	10.8%	6.7%	4.1%	7.8%, 0.5%
3 years	13.4%	8.7%	4.6%	8.7%, 0.6%
4 years	15.2%	10.7%	4.5%	9.2%, -0.1%
5 years	16.8%	12.5%	4.3%	9.7%, -1.2%
Years of life saved (over 5 year horizon)			0.18	0.12, 0.26

Quadratic

6 months	3.6%	3.1%	0.5%	3.1%, -2.1%
1 year	6.7%	4.6%	2.1%	5.7%, -1.5%
2 years	10.8%	6.7%	4.1%	9.2%, -0.9%
3 years	13.4%	8.7%	4.6%	10.6%, -1.3%
4 years	15.3%	10.7%	4.6%	11.5%, -2.3%
5 years	16.8%	12.5%	4.3%	12.2%, -3.6%
Years of life saved (over 5 year horizon)			0.18	0.12, 0.27

Cubic

6 months	3.5%	3.1%	0.4%	3.4%, -2.6%
1 year	6.4%	4.6%	1.8%	6.2%, -2.6%
2 years	10.3%	6.6%	3.7%	10.1%, -2.8%
3 years	12.7%	8.7%	4.1%	11.8%, -3.7%
4 years	14.5%	10.6%	3.9%	12.9%, -5.1%
5 years	16.0%	12.4%	3.6%	13.8%, -6.6%
Years of life saved (over 5 year horizon)			0.15	0.09, 0.25

Quartic

6 months	3.7%	2.8%	0.9%	4.2%, -2.3%
1 year	6.7%	4.2%	2.5%	7.6%, -2.6%
2 years	10.6%	6.3%	4.3%	11.8%, -3.2%
3 years	13.1%	8.2%	4.9%	14.1%, -4.2%
4 years	15.1%	9.9%	5.2%	15.8%, -5.3%
5 years	16.7%	11.4%	5.4%	17.2%, -6.5%
Years of life saved (over 5 year horizon)			0.20	0.10, 0.38

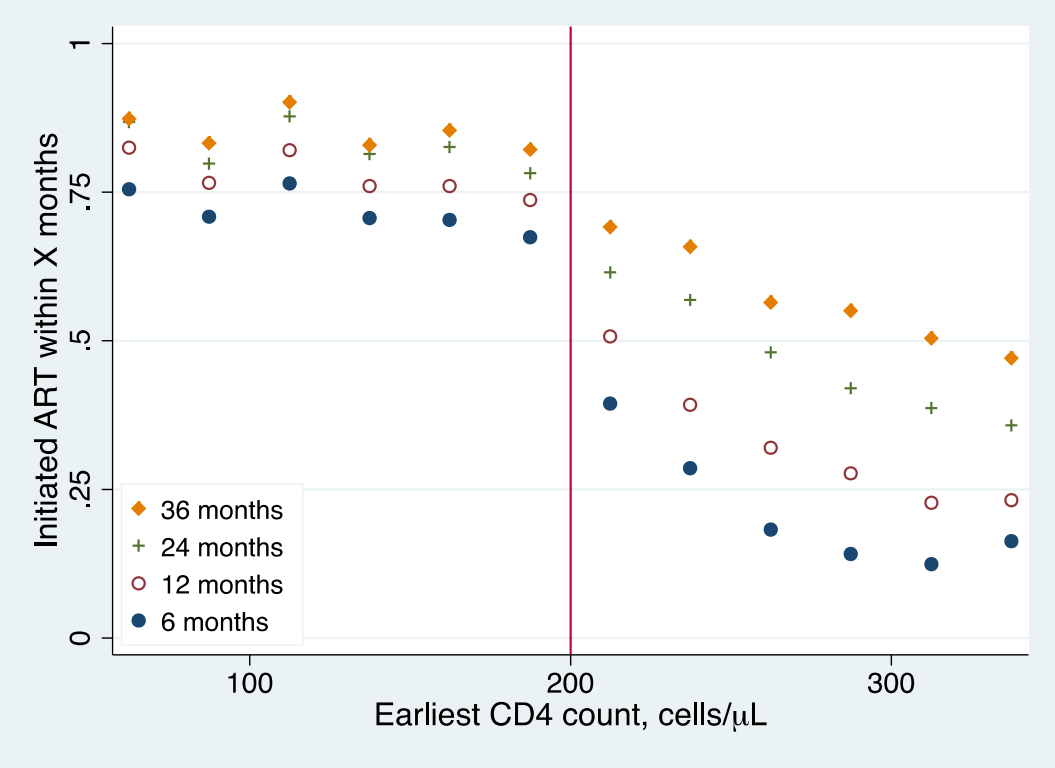
Note: Each panel is estimated in a separate flexible parametric survival model with four knots in the spline of log-time and four knots in the time-varying effect of covariates. The linear effect of earliest CD4 count was allowed to vary over time; higher order polynomial terms were modeled as time-invariant, 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

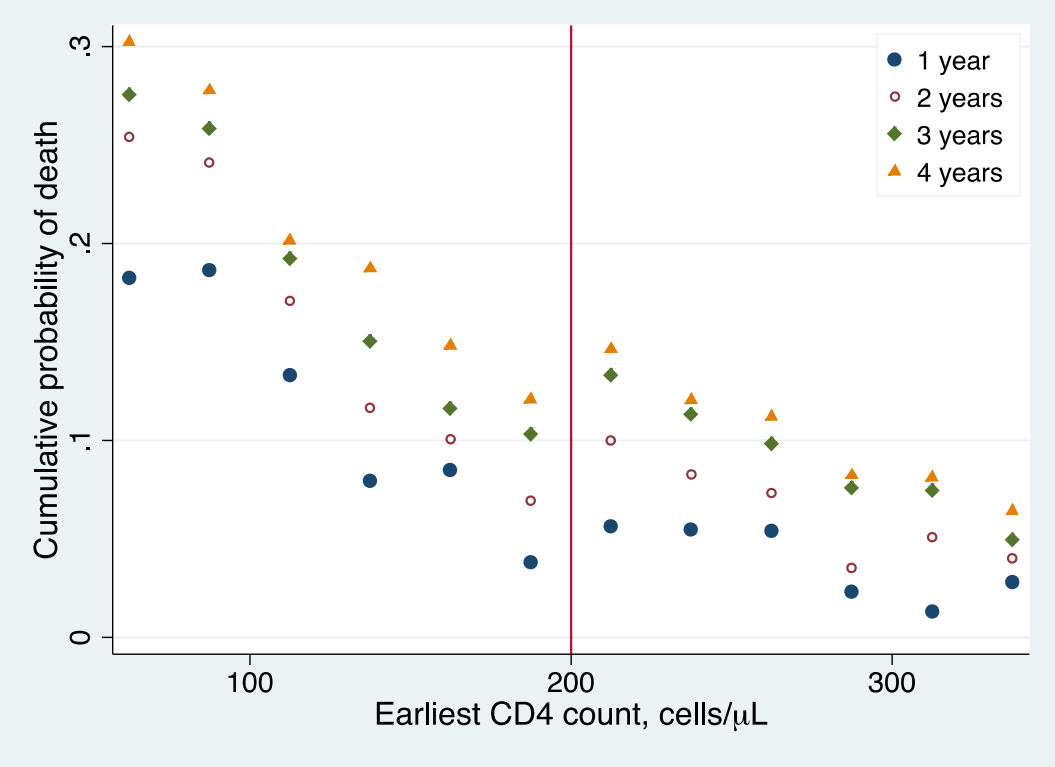
Each hazard ratio is estimated in its own regression. Models control for separate linear terms in earliest CD4+ count on either side of the threshold. Models in rows 6-9 present hazard ratios conditional on individual-level frailties (random effects). Models in rows 10-13 display hazard models for HIV-related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



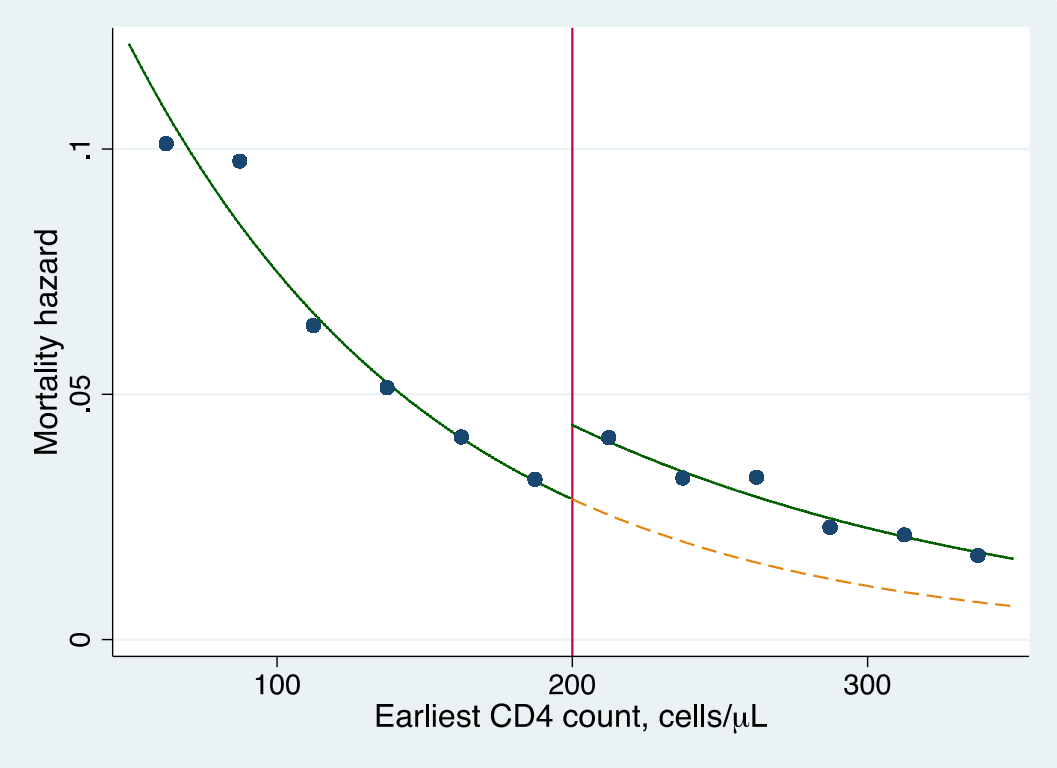
Note: Kaplan-Meier estimates of the probability of initiation among survivors. Follow-up time was censored at date of death or last survey visit.

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



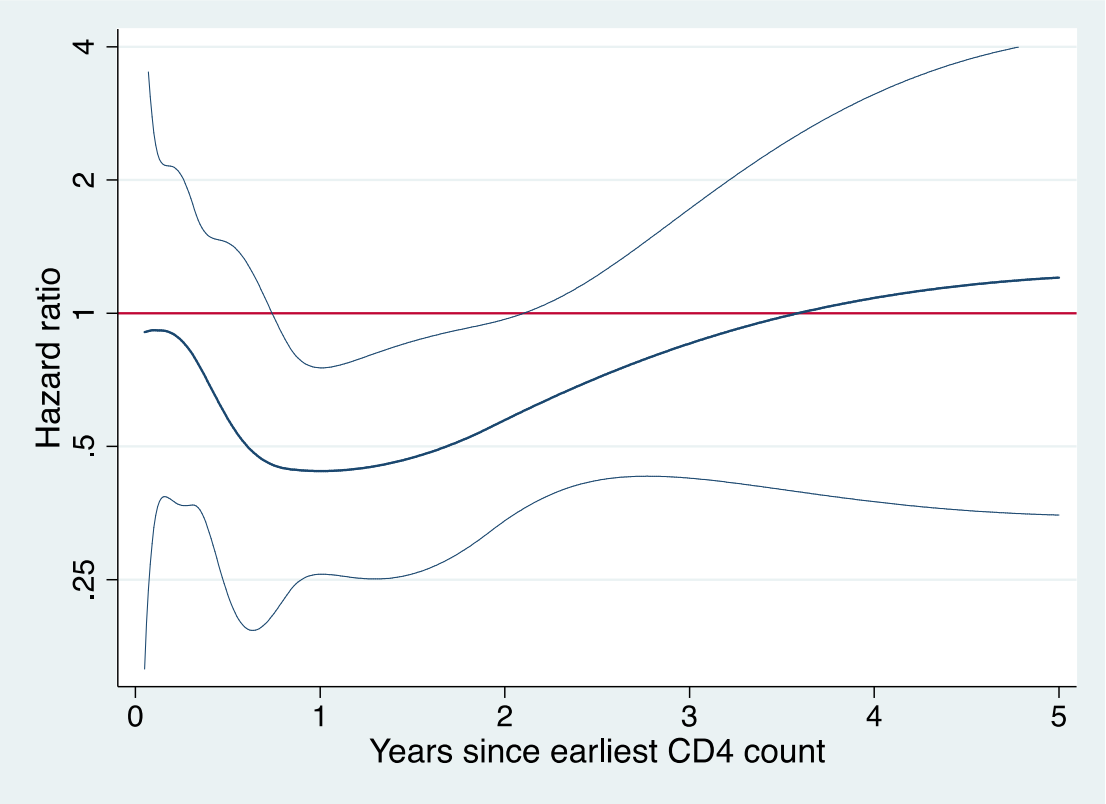
Note: Kaplan-Meier estimates of the probability of death. Follow-up time was censored at date of death or last survey visit.

Figure S3. Baseline CD4+ count and hazard of death



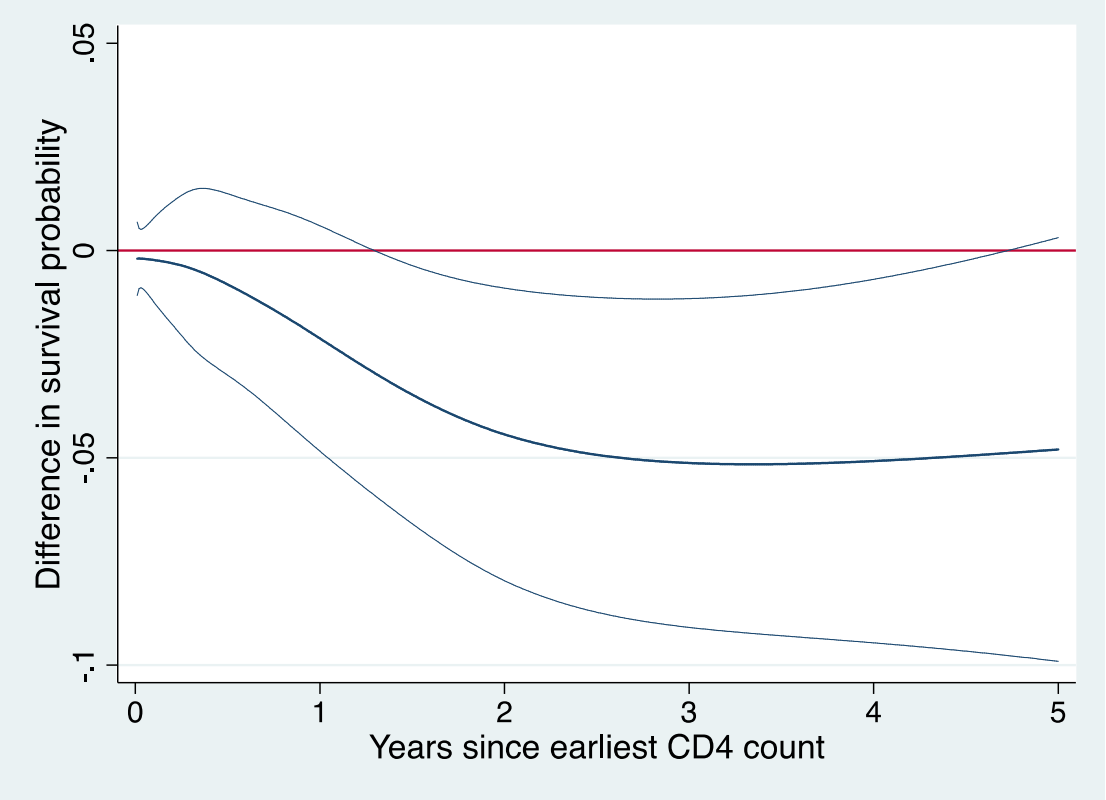
Predictions from exponential hazard model.

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



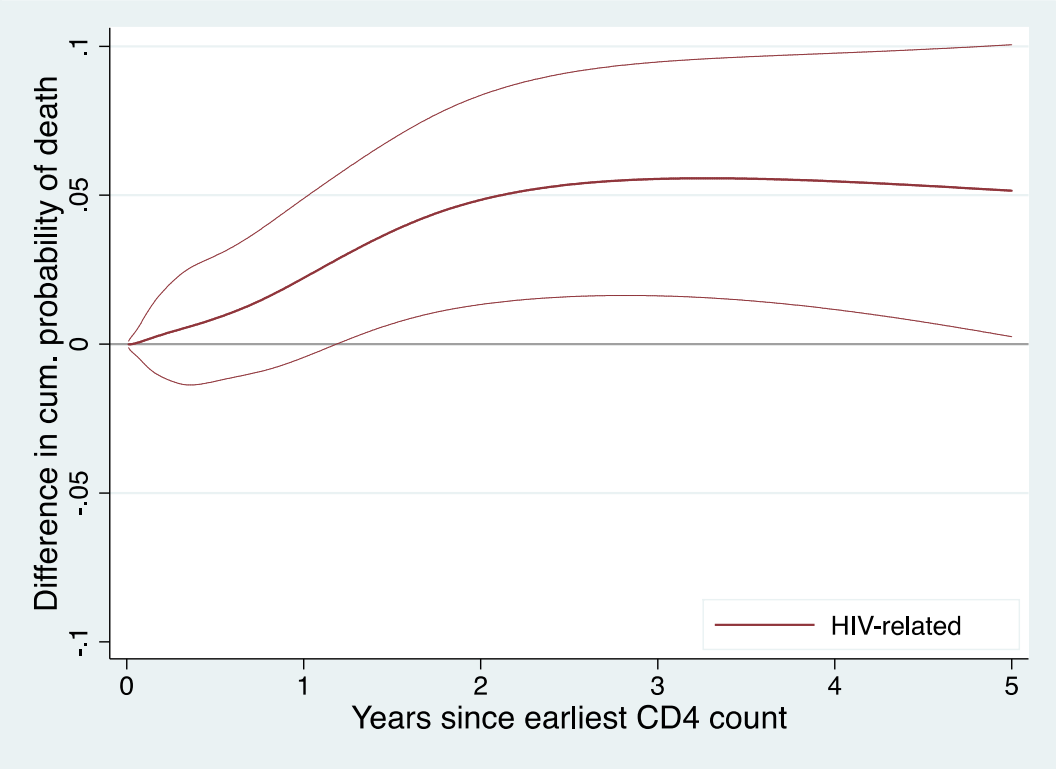
Predictions from flexible parametric survival model.

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



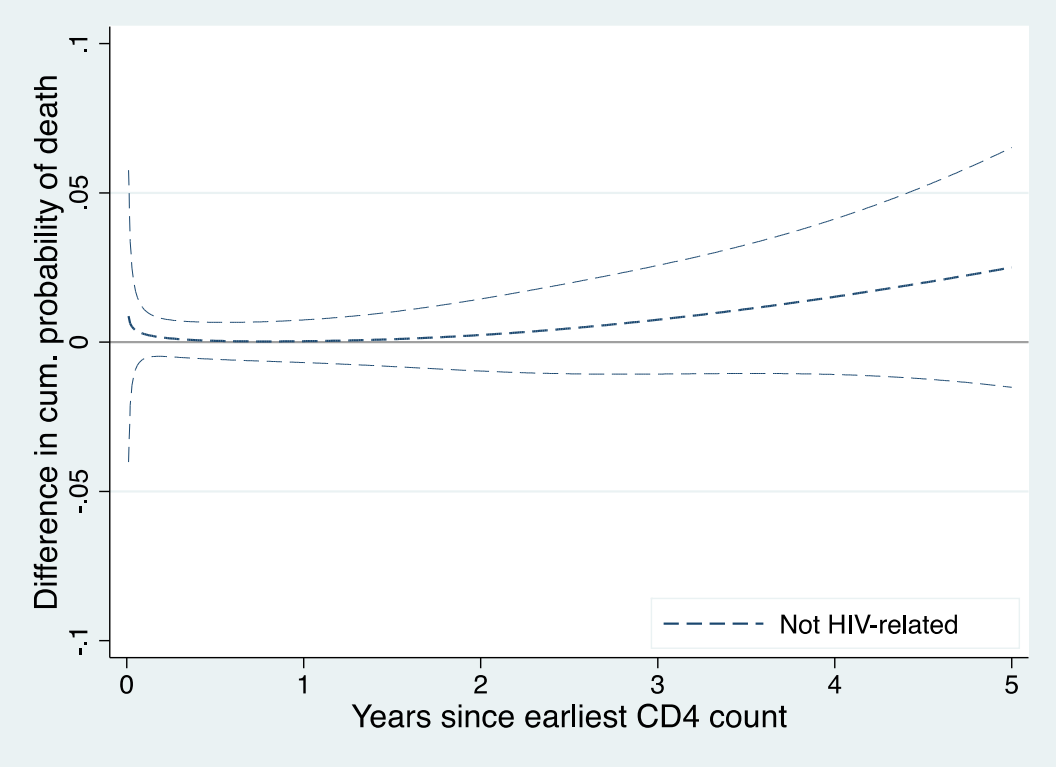
Predictions from flexible parametric survival model.

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

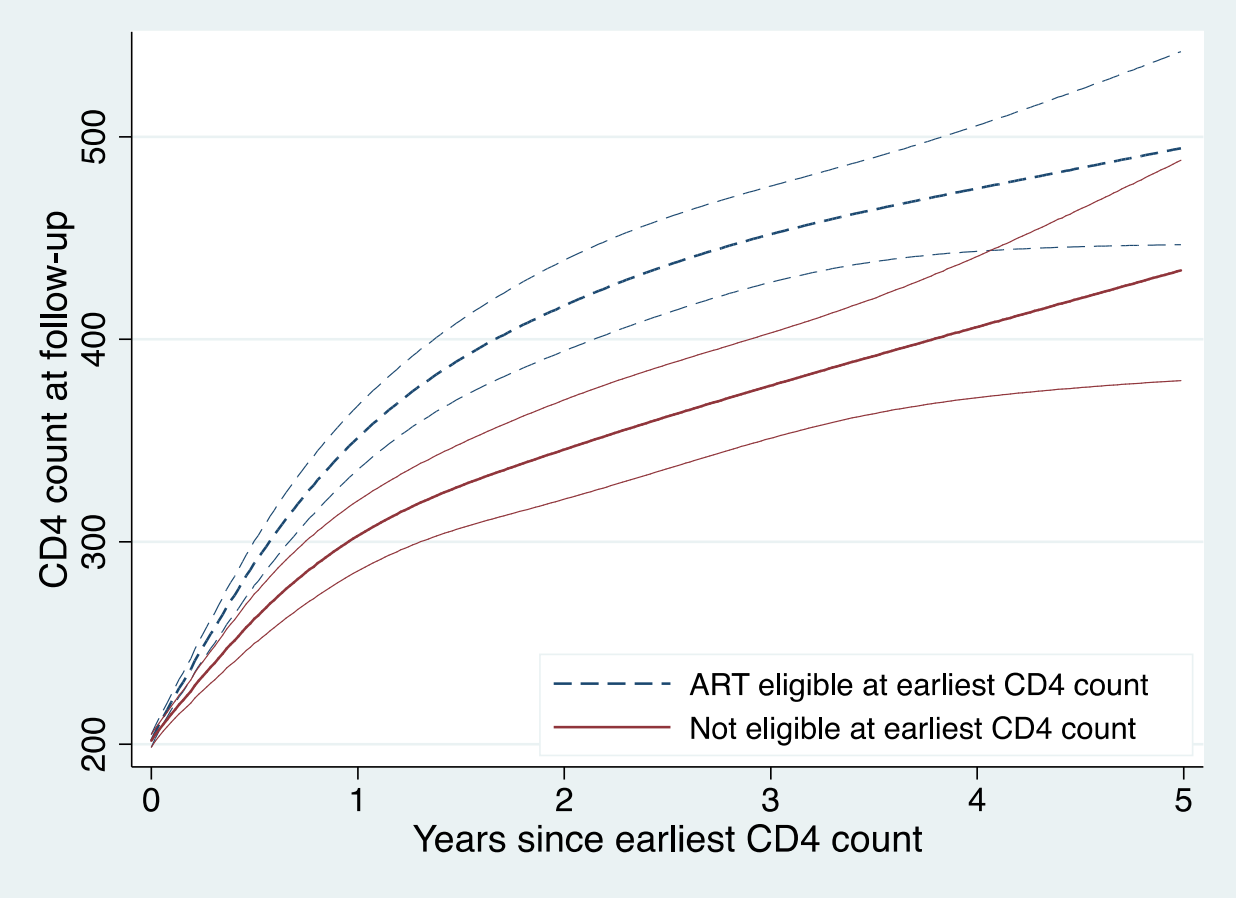


Figure displays predicted mean CD4 counts over time for patients presenting with an initial CD4 count just below (eligible) vs. just above (not eligible) the 200-cell threshold. Linear mixed-effects regression-discontinuity models were estimated with the effect of time modeled as a cubic spline, and 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.