American Journal of Epidemiology
© The Author 2009. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

Vol. 171, No. 1
DOI: 10.1093/aje/kwp329
Advance Access publication:
November 24, 2009

Practice of Epidemiology

Using Marginal Structural Measurement-Error Models to Estimate the Long-term Effect of Antiretroviral Therapy on Incident AIDS or Death

Stephen R. Cole*, Lisa P. Jacobson, Phyllis C. Tien, Lawrence Kingsley, Joan S. Chmiel, and Kathryn Anastos

* Correspondence to Dr. Stephen R. Cole, University of North Carolina Gillings School of Global Public Health, MacGavran-Greenberg Hall, CB#7435, Chapel Hill, NC 27599-7435 (e-mail: cole@unc.edu).

Initially submitted March 25, 2009; accepted for publication September 16, 2009.

To estimate the net effect of imperfectly measured highly active antiretroviral therapy on incident acquired immunodeficiency syndrome or death, the authors combined inverse probability-of-treatment-and-censoring weighted estimation of a marginal structural Cox model with regression-calibration methods. Between 1995 and 2007, 950 human immunodeficiency virus–positive men and women were followed in 2 US cohort studies. During 4,054 person-years, 374 initiated highly active antiretroviral therapy, 211 developed acquired immunodeficiency syndrome or died, and 173 dropped out. Accounting for measured confounders and determinants of dropout, the weighted hazard ratio for acquired immunodeficiency syndrome or death comparing use of highly active antiretroviral therapy in the prior 2 years with no therapy was 0.36 (95% confidence limits: 0.21, 0.61). This association was relatively constant over follow-up ($P = 0.19$) and stronger than crude or adjusted hazard ratios of 0.75 and 0.95, respectively. Accounting for measurement error in reported exposure using external validation data on 331 men and women provided a hazard ratio of 0.17, with bias shifted from the hazard ratio to the estimate of precision as seen by the 2.5-fold wider confidence limits (95% confidence limits: 0.06, 0.43). Marginal structural measurement-error models can simultaneously account for 3 major sources of bias in epidemiologic research: validated exposure measurement error, measured selection bias, and measured time-fixed and time-varying confounding.

acquired immunodeficiency syndrome; bias (epidemiology); cohort studies; confounding factors (epidemiology); epidemiologic measurements; HIV; pharmacoepidemiology; selection bias

Abbreviations: AIDS, acquired immunodeficiency syndrome; CL, confidence limits; HAART, highly active antiretroviral therapy; HIV, human immunodeficiency virus; IPTC, inverse probability of treatment and censoring; MACS, Multicenter AIDS Cohort Study; SE, standard error.

Incident acquired immunodeficiency syndrome (AIDS) is a central clinical event in the progression of human immunodeficiency virus (HIV) infection. Because of a demonstrated (1–3) strong, immediate protective effect of highly active antiretroviral therapy (HAART), randomized evidence bearing on the long-term effectiveness of HAART remains unavailable.

Observational analyses of prospective cohorts to estimate the effectiveness of HAART are difficult because one must measure and account for known and unknown time-fixed and time-varying confounders (4, 5). Standard adjustment or stratification for known time-varying confounders fails to consistently estimate the net (i.e., direct and indirect) effect of HAART on incident AIDS (6) and allows possible selection bias (7–9). Prior observational analyses accounting for known time-varying confounders (10, 11) have 1) assumed once initiated on HAART, individuals continue to use therapies (10, 11); 2) taken reported HAART use as measured without error (10); 3) followed participants for 6.5 years (10) to 7.5 years (11); and 4) estimated discrepant effects of HAART (hazard ratios ranged from 0.14 (11) to 0.54 (10)), depending in part on the specification of the comparison group.

In the present paper, we use data from ongoing observational cohort studies of adult women and men to estimate the net effect of HAART use, versus no antiretroviral therapy use, on AIDS-free survival over a period of more than 11 years. We allow individuals to start and stop therapy at each semiannual study visit and correct for exposure measurement error in reported HAART use based on pooled external validation data. We combine use of inverse probability-of-treatment-and-censoring (IPTC) weighted estimation of a marginal structural Cox model (12, 13) with regression calibration (14–17), which together enable consistent estimation of the net effect of HAART exposure under the assumptions of no unmeasured confounding, no informative censoring, no residual measurement error, and correct specification of the models used to estimate the IPTC weights.

MATERIALS AND METHODS

Study population

This analysis used information from the Multicenter AIDS Cohort Study (MACS) (18), which beginning in 1984 enrolled 6,972 homosexual and bisexual men in Baltimore, Maryland; Chicago, Illinois; Pittsburgh, Pennsylvania; and Los Angeles, California, and from the Women's Interagency HIV Study (19), which beginning in 1994 enrolled 3,772 women in New York, New York; Chicago; Los Angeles; San Francisco, California; and Washington, DC. Every 6 months, participants in both studies completed a physical examination and an extensive interviewer-administered questionnaire with information on antiretroviral therapy use and provided a blood sample for the determination of CD4 cell count and HIV-1 viral load. Positive enzyme-linked immunoabsorbent assays with confirmatory Western blots were used to determine HIV-1 seropositivity. Institutional review boards approved all protocols and informed consent forms, which were completed by study participants in both cohorts.

Analyses presented here include the 950 men and women who were alive, HIV seropositive, and not using antiretroviral therapies in April 1995 before HAART became available (first regimen approved by the US Food and Drug Administration on December 6, 1995). Each participant contributed a maximum of 24 study visits beginning with the first semiannual visit after April 1995 (the baseline visit) and ending with the last visit that he or she was seen alive without clinical AIDS, before initiation of non-HAART antiretroviral therapy, at the second consecutive missed visit (i.e., dropout), or at the date of analysis in September 2007, whichever occurred first. For participants missing baseline data on any time-varying covariate, baseline was redefined to be the first visit with complete data. This approach is analogous to late entries in survival analysis (20) and assumes that late entry is noninformative (21).

AIDS and death ascertainment

The endpoints of interest were first diagnosis of clinical AIDS or death from any cause. The 1993 Centers for Disease Control and Prevention clinical conditions criteria were used to define clinical AIDS (22). Therefore, participants were not considered to have clinical AIDS if they had only a CD4 count of <200 cells/mm^3 or a CD4 percentage of <14 but no clinical AIDS-defining condition. A description of outcomes ascertainment has been published elsewhere (18, 23). Briefly, physician or hospital records confirmed reported clinical AIDS cases among the cohort of men, whereas the cohort of women self-reported clinical AIDS. Deaths were ascertained by using death certificate abstractions upon notification and National Death Registry searches.

Assessment of HAART

The primary exposure was use of HAART in the prior 2 years versus no antiretroviral therapy, because this comparison is of current clinical interest (11). We also explored recent and long-term HAART use (as defined below). The definition of HAART was based on the US Department of Health and Human Services panel guidelines (24), as previously published (10). Typical HAART regimens consisted of 2 or more nucleoside or nucleotide reverse transcriptase inhibitors in combination with 1 or more protease inhibitor or 1 nonnucleoside reverse transcriptase inhibitor.

Assessment of covariates

A number of time-fixed and time-varying covariates were recorded. T-lymphocyte subsets were determined by immunofluorescence using flow cytometry (Becton Dickinson, Mountain View, California) (25). The HIV-1 RNA viral load, in copies per milliliter of plasma, was measured by using an isothermal nucleic acid sequence-based amplification method for women (bioMérieux, Boxtel, the Netherlands) and a reverse transcription polymerase chain reaction amplification assay for men (Roche Molecular Systems, Branchburg, New Jersey). Missing time-varying covariate information after the baseline visit (9%) was carried forward from the most recent prior observed value.

Statistical methods

Let uppercase letters denote random variables and lowercase letters possible realizations. Let $Y_{ij} = 1$ indicate incident AIDS or death during the visit interval $(j, j + 1]$, 0 otherwise, for participant $i = 1$ to N and visit $j = 0$ to $J_i - 1$. Let J_i be the minimum of the last study visit or the (planned) visit subsequent to incident AIDS, death, or censoring. Therefore, the timescale is time on study. Let $C_{ij} = 1$ indicate dropout or initiation of non-HAART therapy during the visit interval $(j, j + 1]$, 0 otherwise. Let $X_{ij} = 1$ indicate reported use of HAART in the visit interval $(j - 1, j]$, 0 otherwise. Let $Z_{ij} = 1$ indicate actual use of HAART in the same visit interval. Finally, let L_{ij} denote time-varying covariates measured at visit j. Denote the history of a time-varying variable using overbars, so that $\bar{X} = \bar{X}_{ij} = \{X_{i0}, X_{i1}, \ldots, X_{ij}\}$ is the history of exposure to HAART through visit j.

A marginal structural (12) pooled logistic regression (26) model is

$$\log\left\{\frac{\Pr(Y_{ij}^{\bar{x},\bar{c}=0}=1)}{1-\Pr(Y_{ij}^{\bar{x},\bar{c}=0}=1)}\right\} = b_{0j} + b_1 g(\bar{x}),$$

where $Y_{ij}^{\bar{x},\bar{c}=0}$ is a time-varying indicator of incident AIDS or death in the visit interval $(j, j+1]$ if the participant had followed treatment history $\bar{x}, \bar{c} = 0$, and $g(\bar{x}) = \min(4, j+1)^{-1} \sum_{k=\max(0,j-3)}^{j} x_{ik}$ is the proportion of reported HAART use over the prior 4 study visits. Our estimand is a discrete-time hazard ratio for incident AIDS or death, $\exp(b_1)$, comparing treatment with HAART over the prior 4 visits (approximately 2 years) against no antiretroviral therapy. We also considered recent exposure, $g(\bar{x}) = x_{ij}$, and exposure over the entire follow-up period, $g(\bar{x}) = j^{-1} \sum_{k=0}^{j} x_{ik}$.

We estimate $b = \{b_{0j}, b_1\}$ as $\beta = \{\beta_{0j}, \beta_1\}$ by maximizing a weighted version of the Bernoulli likelihood function

$$L(\beta) = \prod_{i=1}^{N}\prod_{j=0}^{J_i-1} p_{ij}^{Y_{ij}\hat{W}_{ij}} \times \left(1-p_{ij}\right)^{\left(1-Y_{ij}\right)\hat{W}_{ij}},$$

where $p_{ij} = 1/\left(1+\exp\left\{-\left[\beta_{0j} + \beta_1 g\left(\bar{X}_{ij}\right)\right]\right\}\right)$, $\hat{W}_{ij}$ are estimated time-varying IPTC weights (described below) and β_{0j} are visit-specific intercepts fit by using a restricted cubic spline with 4 knots at the 5th, 33rd, 67th, and 95th percentiles. The discrete-time hazard ratio well approximates the continuous time hazard ratio when the risk of AIDS or death in any interval is less than 10%, which held in our example because the largest event proportion in any visit interval was 6%.

To account for time-varying confounding of HAART use and for right censoring by dropout or non-HAART antiretroviral therapy initiation, we fit the above pooled logistic model using stabilized IPTC weights of the form $W_{ij} = W_{ij}^{X} \times W_{ij}^{C}$, where

$$W_{ij}^{X} = \prod_{k=0}^{j} f[X_{ik} | \bar{X}_{ik-1}, \bar{C}_{ik-1} = 0] \\ / f[X_{ik} | \bar{X}_{ik-1}, \bar{C}_{ik-1} = 0, \bar{L}_{ik-1}]$$

and

$$W_{ij}^{C} = \prod_{k=0}^{j} \Pr[C_{ik} = 0 | \bar{C}_{ik-1} = 0, \bar{X}_{ik}] \\ / \Pr[C_{ik} = 0 | \bar{C}_{ik-1} = 0, \bar{X}_{ik}, \bar{L}_{ik-1}],$$

where $f[\cdot|\cdot]$ is the conditional density function evaluated at the observed covariate values for a given participant.

Baseline covariates L_{i0} were measured at the semiannual study visit immediately prior to the baseline visit and included age, sex, CD4 count categories (i.e., <200, 200–350, 351–500, >500 cells/mm^3), and viral load categories (i.e., <4,001, 4,001–10,000, and >10,000 copies/mL). The IPTC weights were stabilized by past HAART exposure, namely $\{X_{ik-1}, X_{ik-2}, X_{ik-3}\}$, to represent $\bar{X}_{ik-1}$. The time-varying covariate histories $\bar{L}_{ik-1}$ were specified as restricted cubic splines (with 4 knots located at the same percentiles as given above) for CD4 cell count and $\log_{10}$ HIV viral load both measured at visit $k-1$. We estimated the components of W_{ij} using pooled logistic regression models, as previously described (27). If confounding by unmeasured factors is absent and censoring is ignorable, the IPTC weighted estimates β of the pooled logistic model approximate the parameters b of the marginal structural model. Formal definitions of unmeasured confounding and ignorable censoring are given in Hernán et al. (13).

Regression calibration (14–17) using external validation data was applied to the IPTC weighted data. In the weighted data, the relation of the measured confounders $\bar{L}_{ij-1}$ and the misclassified exposure $\bar{X}_{ij}$ is removed, but the relation between the misclassified $\bar{X}_{ij}$ and true exposure $\bar{Z}_{ij}$ persists (28). Therefore, given a mapping (i.e., calibration) of the relation between the misclassified $\bar{X}_{ij}$ and true exposure $\bar{Z}_{ij}$ in the nonconfounded weighted data, one is able to correct for misclassification of exposure by using regression calibration. Details of the validation data, which include studies of 126 MACS men and 205 patients enrolled in the University of North Carolina Center for AIDS Research clinic cohort, are provided in Appendix 1; details of regression calibration are given in Appendix 2. A limited Monte Carlo simulation demonstrating some finite sample properties of the proposed approach is provided in Appendix 3.

On the basis of prior research (10, 27, 29), interactions between HAART and sex and between HAART and baseline CD4 cell count categories were explored. We also explored the constancy of the hazard ratio over time on study using both a product between treatment and (continuous) time and a split at 2 years, which is approximately the median event time. To explore the variance traded to account for possible bias due to time-varying confounding, we truncated the IPTC weights from below and above at percentiles 1 and 99, respectively (30). In addition to marginal structural models, we fit standard pooled logistic regression models with the same time-varying exposure and covariates for comparison, as described previously (10). All analyses were conducted with SAS version 9 software (SAS Institute, Inc., Cary, North Carolina), using robust variance estimates (31) to calculate confidence limits and P values for the marginal structural models (refer to the Appendix in Cole et al. (27)).

RESULTS

At study entry, the 950 participants were a median age of 38 (quartiles: 33, 44) years and had a CD4 count of 453 (quartiles: 303, 641) cells/mm^3 and a viral load of 4.5 (quartiles: 4, 4.9) $\log_{10}$ copies/mL for the 73% with detectable values. Sixty-one percent were women, and 41% were Caucasian (Table 1).

The 950 participants contributed 4,054 person-years under observation. The median length of follow-up was 2.2 (quartiles: 0.5, 8.6) years over the follow-up period between September 1995 and September 2007. The CD4 count averaged over follow-up was 40 cells/mm^3 higher and the HIV viral load was 0.4 $\log_{10}$ copies/mL lower than at baseline. Two hundred eleven (22%) participants developed AIDS (n = 180) or died (n = 31) during follow-up,

Table 1. Baseline and Follow-up Characteristics of 950 Men and Women Infected With HIV Type 1, Multicenter AIDS Cohort Study and Women's Interagency HIV Study, 1995–2007

Characteristic	Baseline (1995) (*n* = 950 Persons)			Follow up (1995–2007) (*n* = 9,172 Person-Visits[a])		
	%	Mean (SD)	No.	%	Mean (SD)	No.
Age, years		39 (8)			47 (9)	
Female sex	61		578	53		4,825
African American	59		560	53		4,839
Use of HAART	0		0	39		3,556
Category of CD4 cell count (no. of cells/mm^3)						
<200	11		109	9		793
200–350	21		195	20		1,793
351–500	26		251	24		2,234
>500	42		395	47		4,352
CD4 cell count (no. of cells/mm^3)		498 (279)			538 (299)	
Category of HIV-1 RNA level (no. of copies/mL)						
<4,001	34		324	58		5,378
4,001–10,000	13		120	11		985
>10,000	53		506	31		2,809
Log$_{10}$ HIV-1 RNA level (no. of copies/mL)		4.5 (0.7)[b]			4.1 (0.7)[b]	

Abbreviations: AIDS, acquired immunodeficiency syndrome; HAART, highly active antiretroviral therapy; HIV, human immunodeficiency virus; SD, standard deviation.

[a] 950 people contributed 9,172 person-visits, with 4,054 person-years of follow-up.

[b] Among 691 and 5,040 detectable measurements at baseline and follow-up, respectively.

307 (32%) completed follow-up alive, 259 (27%) were censored because of initiation of non-HAART antiretroviral therapy, and the remaining 173 (18%) were censored because of study dropout.

Among the 950 participants, 374 (39%) initiated HAART during follow-up. Two thousand ten of 4,054 person-years (50%) were contributed prior to HAART exposure, 1,773 person-years (44%) were contributed while participants were exposed to HAART, and 270 person-years were contributed by participants following discontinuation of HAART. Concerning person-years of exposure to HAART, 2,098 of 4,054 person-years (52%) were fully unexposed in the prior 2 years and 1,151 of 4,054 person-years (28%) were fully exposed; the remaining 20% of 4,054 person-years were partially exposed in the prior 2 years (i.e., 275, 244, and 286 person-years were attributed to 25%, 50%, and 75% HAART exposure, respectively). The predicted probability of HAART use since the prior visit conditional on measured covariates ranged from 1.1×10^{-8} to 0.99, with a mean of 0.40. After stabilization, the IPTC weights ranged from 0.15 to 23.6, with a mean of 0.97; the 1st and 99th percentiles were 0.28 and 2.87, and quartiles were 0.68, 0.90, and 1.08, respectively. As expected, for participants not using HAART at the prior visit, the odds of HAART use increased by a factor of 1.34 (95% confidence limits (CL): 1.28, 1.41) for each decrement of 100 CD4 cells/mm^3 and by a factor of 1.40 (95% CL: 1.07, 1.83) for a detectable viral load.

The unadjusted hazard of AIDS or death was 0.75 (95% CL: 0.49, 1.17) for those using HAART during the prior 2 years relative to those not using antiretroviral therapy (Table 2). The adjusted hazard ratio was 0.95 (95% CL: 0.58, 1.56), so the unadjusted hazard ratio of 0.75 was 1.27 times weaker after adjustment for time-varying CD4 cell count and viral load. The adjusted estimate was slightly less precise, with a 1.14-times larger standard error (for the log of the hazard ratio, 0.25 vs. 0.22).

The hazard of AIDS or death from the marginal structural model (i.e., weighted) was 0.36 (95% CL: 0.21, 0.61) for HAART use in the prior 2 years relative to not using therapy (Table 2). This weighted estimate was 1.8 and 2.6 times stronger than the unadjusted and adjusted estimates, respectively. However, the weighted estimate was less precise than the unadjusted estimate, with a 1.2-times larger standard error (for the log of the hazard ratio, 0.27 vs. 0.22). Truncating the IPTC weights at the 1st and 99th percentiles yielded a similar estimate (hazard ratio = 0.42, 95% CL: 0.26, 0.67) and precision (a 0.89-times smaller standard error for the log of the hazard ratio, 0.24 vs. 0.27). The effect of recent HAART use (prior 6 months) and use over the entire follow-up period, rather than use in the prior 2 years, yielded weighted hazard ratios of 0.43 (95% CL: 0.28, 0.67) and 0.25 (95% CL: 0.13, 0.48), respectively.

The effect of HAART during the prior 2 years appeared stronger among men (hazard ratio = 0.31, 95% CL: 0.13, 0.75) than women (hazard ratio = 0.64, 95% CL: 0.35, 1.20), but this difference may have been due to chance (P homogeneity = 0.16). The effect of HAART appeared stronger at lower levels of baseline CD4 count (hazard ratios

Table 2. Effect of HAART on Incident AIDS or Death Among 950 Men and Women Infected With HIV Type 1, Multicenter AIDS Cohort and Women's Interagency HIV Studies, 1995–2007

Model	Exposure	Hazard Ratio	95% CL[a]
Unadjusted	No ART	1	
	HAART[b]	0.75	0.49, 1.17
Adjusted[c]		0.95	0.58, 1.56
Weighted[c]		0.36	0.21, 0.61
Weighted and calibrated		0.17	0.06, 0.43[d]

Abbreviations: AIDS, acquired immunodeficiency syndrome; ART, antiretroviral therapy; CL, confidence limits; HAART, highly active antiretroviral therapy; HIV, human immunodeficiency virus.

[a] Robust for weighted models.

[b] Use during the prior 2 years.

[c] Controlled for time-varying prior CD4 cell count and HIV-1 RNA level by using restricted cubic splines.

[d] Confidence limits obtained by the delta method using robust variance.

for <200, 200–350, 351–500, and >500 cells/mm^3 at baseline were 0.11, 0.24, 0.40, and 0.47, respectively; *P* homogeneity = 0.12). The effect of HAART appeared stronger at earlier time on study, with hazard ratios of 0.14 (95% CL: 0.03, 0.54) before 2 years and 0.40 (95% CL: 0.23, 0.69) after 2 years from study entry, but this difference may have been due to chance (*P* homogeneity = 0.12). Moreover, a test of the proportional hazards assumption not categorizing time yielded *P* homogeneity = 0.19.

When the pooled external validation data were used, the estimated calibration slope, $\hat{\gamma}_1$, was 0.57 (95% CL: 0.49, 0.65). In the pooled external validation data, the percentage of participants using HAART according to medical records, given no reported use, was 35% (standard error (SE), 5) and the percentage using HAART according to medical records, given reported use, was 92% (SE, 2). The hazard of AIDS or death from regression calibration of the marginal structural model (i.e., weighted and calibrated) was 0.17 ($= \exp[\ln(0.36)/0.57]$, 95% CL: 0.06, 0.43) for HAART use during the prior 2 years relative to not using therapy (Table 2). This calibrated estimate was 2.1 times stronger than the weighted estimate. The weighted and calibrated estimate was less precise than the weighted estimate, with a 1.8-times larger standard error (for the log of the hazard ratio, 0.48 vs. 0.27). When we restricted the external validation data to MACS (hazard ratio = 0.29, 95% CL: 0.15, 0.55) or University of North Carolina (hazard ratio = 0.12, 95% CL: 0.04, 0.39) as a sensitivity analysis, results that bounded the estimate using combined data were in concordance with expectations given the characteristics of the validation data and provided a range within which the true result likely resides.

DISCUSSION

Using a marginal structural Cox proportional hazards model and regression calibration, we estimated that, relative to not using antiretroviral therapy, HAART use during the prior 2 years decreases the hazard of AIDS or death by 83% (range of the 95% CL across sensitivity analysis: 45, 96), and this effect appears to persist for more than 10 years. This dramatic protective effect was attenuated by half when efforts were not made to account for misclassification of reported HAART use. Moreover, the protective effect was further attenuated when standard statistical methods were used to account for the time-varying confounding.

Our results show a stronger effect of HAART on disease progression than the results reported by the trials described by Hammer et al. in 1997 (1) and Cameron et al. in 1998 (3) comparing an early HAART regimen with a combination-therapy regimen. We would expect to see stronger effects than these trials demonstrated because 1) our comparison group reflected the absence of antiretroviral therapy rather than a combination therapy; 2) the HAART regimens used over the course of follow-up have improved, whereas both trials used single, early HAART regimens; and 3) both trials reported noncompliance with assigned therapies, the magnitude of which could notably null-bias the intent-to-treat trial results (32–34).

Our results are also stronger than the findings of Detels et al. (35), who used calendar period as an instrumental variable (36) for HAART exposure in a subset of 536 MACS men for whom seroconversion dates were known and reported a hazard ratio for incident AIDS or death of 0.35 (95% CL: 0.20, 0.61) in a comparison of the time period following HAART introduction with the time period of monotherapy. We would expect to find stronger effects than Detels et al. (35) because, akin to noncompliance in a trial, use of calendar period as an instrument for therapies is subject to information bias if the use of therapies is not a step function across the calendar periods, which it is not, as shown in Figure 1 of Detels et al. and in Cain et al. (37). Indeed, an instrumental variable correction for misclassification of the Detels et al. result yielded a rate ratio of 0.2 (38), which is close to our estimate.

Finally, prior observational analyses accounting for time-varying confounding using marginal structural models (10, 11) have assumed that, once initiating HAART, individuals continue to use therapies (10, 11); have taken reported HAART use as measured without error (10); and have estimated the effect of HAART versus no therapy (hazard ratio = 0.14, 95% CL: 0.07, 0.29) (11), mono- or combination therapy (hazard ratio = 0.51, 95% CL: 0.29, 0.87) (10), combination therapy alone (hazard ratio = 0.49, 95% CL: 0.31, 0.79) (11), or any non-HAART therapy, including no therapy (hazard ratio = 0.54, 95% CL: 0.38, 0.78) (10). As expected, the exposure-misclassification-corrected hazard ratio of 0.17 (95% CL: 0.06, 0.43) presented here coheres better with a previously reported hazard ratio of 0.14 (95% CL: 0.07, 0.29), where marginal structural models were applied to data with HAART use obtained by medical records (11), than the estimate assuming no misclassification of reported HAART use (i.e., 0.36, 95% CL: 0.21, 0.61).

Past work using marginal structural models has largely omitted discussion of the choice of the final structural model, with few exceptions (39). Such model choice largely centers on the functional form of the exposure effect on the outcome, for which there is a broad literature. However, a synthesis of that literature in the context of choice of the final structural model is needed but is beyond the scope of this paper. For instance, the choice of timescale to be used affects the meaning of survival curves and may affect the value of a summary hazard ratio. In addition, the choice of how to represent exposure (e.g., 2-year window) may alter results and have implications for clinical practice or public policy.

The present results should be interpreted with consideration of the following limitations. First, as is true of all observational analyses, the estimates have a causal interpretation only under the assumption of no unmeasured confounding. This assumption likely holds approximately here because the most important clinical and laboratory information used by physicians as indications for HAART was collected and used in the models to estimate the weights (40). As described previously (30), numerous additional functional forms for the weight models were explored (e.g., longer covariate histories, more flexible splines), as well as a broader set of covariates (e.g., age; race; body mass index; HIV-related symptoms; *Pneumocystis jiroveci* pneumonia prophylaxis therapy use; and red blood, platelet, CD3, and CD8 cell counts), but such alternative model specifications did not appreciably alter the results. If the assumption of no unmeasured confounders is correct and the model used to create the treatment weights is correctly specified, then weighting creates a pseudo-population in which the probability of HAART initiation is not a function of the time-varying covariates (i.e., no confounding exists), but the effect of HAART initiation on AIDS or death is the same as in the actual study population.

Second, interpretable causal contrasts require that the consistency assumption be met (41). The consistency assumption is likely to hold approximately in the present setting, where the exposure is a treatment (42, 43).

Third, valid use of IPTC weights requires that there not be a probability of 0 or 1 that participants are exposed at any level of the confounders among the uncensored (30). This assumption was met in theory in our study and appeared to be met in practice (notwithstanding wide-ranging predicted probabilities of exposure) because some participants with high CD4 counts and low viral load initiated HAART, while others with low CD4 counts and high viral loads did not. Refer to Cole and Hernán (30) for a more detailed discussion of positivity in these data.

Fourth, and as in all prospective analyses with right censoring, the results are based on the assumption that right censoring is ignorable conditional on measured covariates. Neither the present analyses nor past analyses (10, 27, 29, 44) suggest there is notable selection bias due to measured variables in these data.

Fifth, the results rely on the assumption that time to AIDS or death is measured without error. Moreover, we assume that the semiannual data on covariates obtained from these interval cohorts (45) are frequent enough so that the information used by clinicians and participants to decide on therapies is not overly coarsened. A prior report comparing interval and clinical cohorts in HIV found similar inferences (46). Furthermore, the calibrated results rely on the assumption that the external validation data for HAART use are accurate. Beyond the simple accuracy of reported HAART use, there is the issue of transportability of the validation data to the main study. The provided corrections apply to the extent that the validation data proxy well for the main study participants. Note that compatibility in participant characteristics is neither necessary nor sufficient for the validation data to proxy well, but similarities in characteristics, as found between our validation and main study data, do provide some reassurance. Little prior work has been published on how measurement error may affect results from structural models for complex longitudinal data (47, 48). For instance, random exposure measurement error may not always lead to null bias because measured exposure at time j may also act as a measured proxy for an unmeasured causal confounder for the effect of treatment at future times (47), which would occur if the treatments Z_{ij} and Z_{ij+1} had an unmeasured common cause or treatment Z_{ij+1} was determined in part by treatment Z_{ij}. In such cases, when the IPTC weights account for the history of measured treatment X_{ij+1}, we may not completely eliminate the confounding (in either direction) by past actual treatment Z_{ij}.

One could account for measurement error in other ways. For instance, with internal validation data, one could use multiple imputation for measurement error correction (49). Alternatively, one could use Bayesian or approximate Bayesian (50, 51) methods with extensions to allow for unidentified bias parameters (52). With any such misclassification correction method, an apparent loss of precision with corrected estimates can be viewed as a movement of the systematic error from a bias in the point estimate to a less-biased estimate with increased uncertainty.

In conclusion, the observed association of HAART with incident AIDS or death appears to persist for more than a decade at a level stronger than observed using standard statistical methods or marginal structural models assuming no misclassification of reported HAART use. Without data from fully compliant randomized trials that follow patients with widely varying risk profiles for prolonged periods, prospective observational studies with repeated assessments of exposure and detailed collection of clinical and laboratory information provide the best evidence available for estimating risk-group-specific, long-term therapeutic effects.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina (Stephen R. Cole); Department of Epidemiology, The Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland (Lisa P. Jacobson); Department of Medicine, University of California, San Francisco, California (Phyllis C. Tien); San Francisco Veterans Affairs Medical Center, San Francisco, California (Phyllis C. Tien); Department of Infectious Diseases and Microbiology, University of Pittsburgh, Pittsburgh, Pennsylvania (Lawrence Kingsley); Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois (Joan

S. Chmiel); and Montefiore Medical Center and Albert Einstein College of Medicine, Bronx, New York (Kathryn Anastos).

Dr. Stephen Cole's work was partially supported by National Institutes of Health grants R03-AI-071763, R01-AA-017594, and P30-AI-50410. The Women's Interagency HIV Study is funded by the National Institute of Allergy and Infectious Diseases (grants UO1-AI-35004, UO1-AI-31834, UO1-AI-34994, UO1-AI-34989, UO1-AI-34993, and UO1-AI-42590) and by the National Institute of Child Health and Human Development (grant UO1-HD-32632). The study is cofunded by the National Cancer Institute, the National Institute on Drug Abuse, and the National Institute on Deafness and Other Communication Disorders. Funding is also provided by the National Center for Research Resources (UCSF-CTSI grant UL1 RR024131). MACS is funded by grants UO1-AI-35042, 5-MO1-RR-00722 (GCRC), UO1-AI-35043, UO1-AI-37984, UO1-AI-35039, UO1-AI-35040, UO1-AI-37613, and UO1-AI-35041 from the National Institutes of Health.

The authors thank Drs. Lauren E. Cain and Emily S. Brouwer for assistance with the validation data. The authors also thank Drs. Miguel A. Hernán, Daniel Westreich, Sander Greenland, and James Robins for their expert advice.

Data in this manuscript were collected by the Women's Interagency HIV Study Collaborative Study Group with centers at the New York City/Bronx Consortium (Kathryn Anastos, Principal Investigator); Brooklyn, New York (Howard Minkoff, Principal Investigator); Washington, DC Metropolitan Consortium (Mary Young, Principal Investigator); The Connie Wofsy Study Consortium of Northern California (Ruth Greenblatt, Principal Investigator); Los Angeles County/Southern California Consortium (Alexandra Levine, Principal Investigator); Chicago Consortium (Mardge Cohen, Principal Investigator); and Data Coordinating Center (Stephen Gange, Principal Investigator).

MACS includes the following: *Baltimore, Maryland*—The Johns Hopkins Bloomberg School of Public Health: Joseph B. Margolick (Principal Investigator), Haroutune Armenian, Barbara Crain, Adrian Dobs, Homayoon Farzadegan, Joel Gallant, John Hylton, Lisette Johnson, Shenghan Lai, Ned Sacktor, Ola Selnes, James Shepard, and Chloe Thio. *Chicago, Illinois*—Howard Brown Health Center, Feinberg School of Medicine, Northwestern University, and Cook County Bureau of Health Services: John P. Phair (Principal Investigator), Joan S. Chmiel (Co-Principal Investigator), Sheila Badri, Bruce Cohen, Craig Conover, Maurice O'Gorman, David Ostrow, Frank Palella, Daina Variakojis, and Steven M. Wolinsky. *Los Angeles, California*—University of California, UCLA Schools of Public Health and Medicine: Roger Detels (Principal Investigator), Barbara R. Visscher (Co-Principal Investigator), Aaron Aronow, Robert Bolan, Elizabeth Breen, Anthony Butch, Thomas Coates, Rita Effros, John Fahey, Beth Jamieson, Otoniel Martínez-Maza, Eric N. Miller, John Oishi, Paul Satz, Harry Vinters, Dorothy Wiley, Mallory Witt, Otto Yang, Stephen Young, and Zuo Feng Zhang. *Pittsburgh, Pennsylvania*—University of Pittsburgh, Graduate School of Public Health: Charles R. Rinaldo (Principal Investigator), Lawrence Kingsley (Co-Principal Investigator), James T. Becker, Robert W. Evans, John Mellors, Sharon Riddler, and Anthony Silvestre. *Data Coordinating Center, Baltimore, Maryland*—The Johns Hopkins Bloomberg School of Public Health: Lisa P. Jacobson (Principal Investigator), Alvaro Muñoz (Co-Principal Investigator), Keri Althoff, Christopher Cox, Gypsyamber D'Souza, Stephen J. Gange, Elizabeth Golub, Janet Schollenberger, Eric C. Seaberg, and Sol Su. *National Institutes of Health, Bethesda, Maryland*—National Institute of Allergy and Infectious Diseases: Robin E. Huebner; National Cancer Institute: Geraldina Dominguez; and National Heart, Lung, and Blood Institute: Cheryl McDonald. Website located at http://www.statepi.jhsph.edu/macs/macs.html.

The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Conflict of interest: none declared.

REFERENCES

1. Hammer SM, Squires KE, Hughes MD, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team. *N Engl J Med*. 1997;337(11):725–733.
2. Gulick RM, Mellors JW, Havlir D, et al. Treatment with indinavir, zidovudine, and lamivudine in adults with human immunodeficiency virus infection and prior antiretroviral therapy [see comments]. *N Engl J Med*. 1997;337(11):734–739.
3. Cameron DW, Heath-Chiozzi M, Danner S, et al. Randomised placebo-controlled trial of ritonavir in advanced HIV-1 disease. The Advanced HIV Disease Ritonavir Study Group. *Lancet*. 1998;351(9102):543–549.
4. Miettinen OS. The need for randomization in the study of intended effects. *Stat Med*. 1983;2(2):267–271.
5. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*. 1986;15(3):413–419.
6. Robins J. The control of confounding by intermediate variables. *Stat Med*. 1989;8(6):679–701.
7. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3(2):143–155.
8. Cole SR, Hernán MA. Fallibility in estimating direct effects (with discussion). *Int J Epidemiol*. 2002;31(1):163–165.
9. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15(5):615–625.
10. Cole SR, Hernán MA, Robins JM, et al. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *Am J Epidemiol*. 2003;158(7):687–694.
11. Sterne JA, Hernán MA, Ledergerber B, et al. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *Lancet*. 2005;366(9483):378–384.
12. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–560.

13. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of non-randomized treatments. *J Am Stat Assoc.* 2001;96(454):440–448.
14. Stefanski LA, Carroll RJ. Covariate measurement error in logistic regression. *Ann Stat*. 1985;13:1335–1351.
15. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med*. 1989;8(9):1051–1069; discussion 1071–1073.
16. Spiegelman D, McDermott A, Rosner B. Regression calibration method for correcting measurement-error bias in nutritional epidemiology. *Am J Clin Nutr.* 1997;65(4 suppl): 1179S–1186S.
17. Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariate misclassification and measurement error, in main study/validation study designs. *J Am Stat Assoc.* 2000;95:51–61.
18. Kaslow RA, Ostrow DG, Detels R, et al. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *Am J Epidemiol*. 1987; 126(2):310–318.
19. Barkan SE, Melnick SL, Preston-Martin S, et al. The Women's Interagency HIV Study. WIHS Collaborative Study Group. *Epidemiology.* 1998;9(2):117–125.
20. Lamarca R, Alonso J, Gómez G, et al. Left-truncated data with age as time scale: an alternative for survival analysis in the elderly population. *J Gerontol A Biol Sci Med Sci.* 1998;53(5): M337–M343.
21. Wang MC, Brookmeyer R, Jewell NP. Statistical models for prevalent cohort data. *Biometrics*. 1993;49(1):1–11.
22. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *MMWR Recomm Rep*. 1992;41(RR-17): 1–19.
23. Hessol NA, Schwarcz S, Ameli N, et al. Accuracy of self-reports of acquired immunodeficiency syndrome and acquired immunodeficiency syndrome-related conditions in women. *Am J Epidemiol*. 2001;153(11):1128–1133.
24. Panel on Clinical Practices for Treatment of HIV Infection. *Guidelines for the Use of Antiretroviral Agents in HIV-1-infected Adults and Adolescents*. (http://aidsinfo.nih.gov/Guidelines/). (Accessed November 3, 2008).
25. Mellors JW, Muñoz A, Giorgi JV, et al. Plasma viral load and CD4+ lymphocytes as prognostic markers of HIV-1 infection. *Ann Intern Med*. 1997;126(12):946–954.
26. Abbott RD. Logistic regression in survival analysis. *Am J Epidemiol*. 1985;121(3):465–471.
27. Cole SR, Hernán MA, Anastos K, et al. Determining the effect of highly active antiretroviral therapy on changes in human immunodeficiency virus type 1 RNA viral load using a marginal structural left-censored mean model. *Am J Epidemiol*. 2007;166(2):219–227.
28. Hernán MA, Cole SR. Invited commentary: Causal diagrams and measurement bias. *Am J Epidemiol*. 2009;170(8):959–962.
29. Cole SR, Hernán MA, Margolick JB, et al. Marginal structural models for estimating the effect of highly active antiretroviral therapy initiation on CD4 cell count. *Am J Epidemiol*. 2005; 162(5):471–478.
30. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656–664.
31. White HA. A heteroskedasticity-consistent covariance matrix estimator and a direct test of heteroskedasticity. *Econometrica*. 1980;48:817–838.
32. Mark SD, Robins JM. A method for the analysis of randomized trials with compliance information: an application to the Multiple Risk Factor Intervention Trial. *Control Clin Trials*. 1993;14(2):79–97.
33. Cole SR, Chu H. Effect of acyclovir on herpetic ocular recurrence using a structural nested model. *Contemp Clin Trials*. 2005;26(3):300–310.
34. Greenland S, Lanes S, Jara M. Estimating effects from randomized trials with discontinuations: the need for intent-to-treat design and G-estimation. *Clin Trials*. 2008;5(1): 5–13.
35. Detels R, Muñoz A, McFarlane G, et al. Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. *JAMA*. 1998;280(17): 1497–1503.
36. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol*. 2000;29(4):722–729.
37. Cain LE, Cole SR, Chmiel JS, et al. Effect of highly active antiretroviral therapy on multiple AIDS-defining illnesses among male HIV seroconverters. *Am J Epidemiol*. 2006;163 (4):310–315.
38. Cain LE, Cole SR, Greenland S, et al. Effect of highly active antiretroviral therapy on incident AIDS using calendar period as an instrumental variable. *Am J Epidemiol*. 2009;169(9): 1124–1132.
39. Neugebauer R, Van der Laan M. Nonparametric causal effects based on marginal structural models. *J Stat Plan Inference*. 2007;137:419–434.
40. Ahdieh L, Gange SJ, Greenblatt R, et al. Selection by indication of potent antiretroviral therapy use in a large cohort of women infected with human immunodeficiency virus. *Am J Epidemiol*. 2000;152(10):923–933.
41. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology.* 2009; 20(1):3–5.
42. Hernán MA. Invited commentary: Hypothetical interventions to define causal effects—afterthought or prerequisite? *Am J Epidemiol*. 2005;162(7):618–620; discussion 621–622.
43. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obes (Lond)*. 2008;32(suppl 3): S8–S14.
44. Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11(5): 561–570.
45. Lau B, Gange SJ, Moore RD. Interval and clinical cohort studies: epidemiological issues. *AIDS Res Hum Retroviruses*. 2007;23(6):769–776.
46. Lau B, Gange SJ, Kirk GD, et al. Evaluation of human immunodeficiency virus biomarkers: inferences from interval and clinical cohort studies. *Epidemiology.* 2009;20(5):664–672.
47. Robins JM. General methodological considerations. *J Econometrics*. 2003;112:89–106.
48. Goetghebeur E, Vansteelandt S. Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure. *Stat Methods Med Res.* 2005;14(4):397–415.
49. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *Int J Epidemiol*. 2006;35(4): 1074–1081.
50. Greenland S. Bayesian perspectives for epidemiological research: I. Foundations and basic methods. *Int J Epidemiol*. 2006;35(3):765–775.

51. Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. *Int J Epidemiol*. 2007;36(1): 195–202.
52. Greenland S. Multiple-bias modelling for analysis of observational data. *J R Stat Soc (A)*. 2005;168:267–306.
53. Cain LE. *Bridging the Gap Between Observational and Randomized Evidence: HAART and AIDS* [dissertation]. Baltimore, MD: Johns Hopkins Bloomberg School of Public Health; 2008.
54. Brouwer ES, Napravnik S, Corbett AH, et al. A validation study of self reported antiretroviral use within a clinical cohort of HIV positive patients. *Pharmacoepidemiol Drug Saf*. 2008; 17(S1):S3.
55. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. *J Am Stat Assoc*. 1990;85:652–663.
56. Pearl J. Causal diagrams for empirical research. *Biometrika*. 1995;82:669–710.
57. Pearl J. *Causality*. New York, NY: Cambridge; 2000.

APPENDIX 1

Validation Data

The mapping between the misclassified and true HAART exposure is taken from 2 reports of validation data (53, 54). In MACS, as reported by Cain (53), the validity of self-reported use of HAART was assessed for 126 HIV-positive adult men who were seen at least once during the period from October 1, 2004, to April 1, 2008, at the Moore Clinic in Baltimore; the Whitman Walker Clinic in Washington, DC; or the Northwestern Clinic in Chicago. For each man, the most recent single clinic visit that met the following criteria was selected: it occurred between October 1, 2003, and April 1, 2008, with a subsequent MACS study visit between October 1, 2004, and April 1, 2008, such that the difference between clinic and study visits was less than 1 year (with a single exception of 1.16 years). In MACS (and the Women's Interagency HIV Study), information on use of each antiretroviral medication is elicited from the participant through interviewer-administered questionnaires with the assistance of photo-medication cards. In the validation substudy, data of all antiretroviral medications that the participant was continuing or starting at the clinic visit were abstracted. The MACS algorithm was used to classify each participant's drug combination as HAART or non-HAART. When medical-record-abstracted HAART was used as a ‘‘gold standard,'' the numbers of true-positive, true-negative, false-positive, and false-negative participants were 101, 18, 4, and 3, respectively. Therefore, the sensitivity, specificity, and positive and negative predictive values of self-reported HAART use were 97 (SE, 2), 82 (SE, 8), 96 (SE, 2), and 86 (SE, 8), respectively. This validation sample of 126 men was aged 49 years (standard deviation, 8), and 55% were African American.

In the University of North Carolina Center for AIDS Research HIV Cohort study, as reported by Brouwer et al. (54), the validity of self-reported use of HAART was assessed for a random sample of 205 HIV-infected adult men and women who completed a 90-question clinical and sociodemographic survey when seen between January 2001 and May 2006 at the University of North Carolina HIV clinic in Chapel Hill. In the University of North Carolina clinic cohort, each antiretroviral medication that the participant was using at a clinic visit was abstracted from medical records. In the validation substudy, participants were asked on the survey to report current use of each antiretroviral medication. When medical-record-abstracted therapies were used as a gold standard, a participant using antiretroviral therapy who correctly reported use of all antiretroviral therapies was considered a true positive; conversely, a participant not using antiretroviral therapy who correctly reported no use was considered a true negative. During the calendar period during which the study took place (i.e., 2001–2006), the vast majority of HIV patients using antiretroviral therapies would have been on a HAART regimen. The numbers of true-positive, true-negative, false-positive, and false-negative participants were 103, 53, 14, and 35, respectively. Therefore, the sensitivity, specificity, and positive and negative predictive values of self-reported therapy use were 75 (SE, 4), 79 (SE, 5), 88 (SE, 3), and 60 (SE, 5), respectively. This validation sample of 205 participants was a median age of 42 years (quartiles: 36, 47), 66% male, and 71% African American.

The pooled MACS and University of North Carolina data yielded a sensitivity and specificity of 84 (SE, 2) and 80 (SE, 4), respectively. There was little heterogeneity in the specificity between studies (chi-squared $= 0.08$, $P = 0.78$), but there was notable heterogeneity in the sensitivities (chi-squared $= 22.6$, $P < 0.01$). Therefore, in addition to using the pooled sensitivity and specificity, we present results using the University of North Carolina and MACS validation data separately.

APPENDIX 2

Regression Calibration on IPTC Weighted Data

Regression calibration proceeds with an estimate of the discrete-time log hazard ratio $\hat{\beta}_1$ between the misclassified exposure and clinical AIDS or death, which is given by the pooled logistic model described in the Statistical Methods section of the text.

Second, a linear calibration model, $E(Z_m|X_m) = \gamma_0 + \gamma_1 X_m$, is fit to the pooled external validation data (refer to Appendix 1), with random errors assumed $\varepsilon_m \sim N(0, \sigma)$, for $m = 1$ to 331 ($= 126 + 205$). Theoretical (55) and simulation (17, 49) evidence supports the use of the linear approximation with a misclassified dichotomous exposure.

Third, a misclassification-corrected log hazard ratio is obtained as $\hat{\theta}_1 = \hat{\beta}_1/\hat{\gamma}_1$, with 95% confidence limits for $\hat{\theta}_1$ obtained by using a variance of $\hat{V}(\hat{\theta}_1) = \hat{\gamma}_1^{-2}\hat{V}(\hat{\beta}_1) + (\hat{\beta}_1^2/\hat{\gamma}_1^4)\hat{V}(\hat{\gamma}_1)$, which is a first-order approximation using the delta method but with a robust variance taken for the discrete-time log hazard ratio.

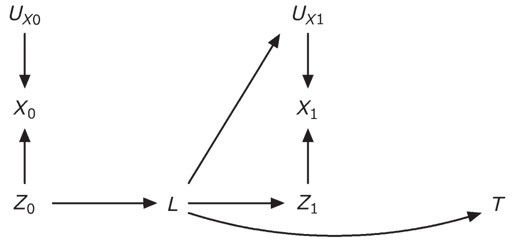
APPENDIX 3

Monte Carlo Simulation of Bias and Confidence Limit Coverage

Data are simulated from the causal diagram (56, 57) illustrated in Appendix Figure 1. Reading from Appendix Figure 1, treatment has no direct causal effect on the time to event. However, treatment at time 0, Z_0, has an indirect causal effect mediated through the time-varying covariate L. Therefore, in this setting, the total causal effect equals the indirect effect of initial treatment. Moreover, the time-varying covariate is a confounder of the association between subsequent treatment Z_1 and events, hence we will subsequently refer to L as a time-varying confounder. Treatment Z_j is nondifferentially and independently misclassified as X_j.

A simulated data record comprises a value for Z_0, X_0, L, Z_1, X_1, T; we drew 1,000 simulated data records for each of 10,000 simulation data sets. First, a Bernoulli random variable was generated with marginal probability P for treatment at time 0, Z_0. Second, a Bernoulli random variable was generated with marginal probability 0.5 for the time-varying confounder, L, conditional on the realized value of treatment at time 0, z_0, as $1/[1 + \exp(\alpha_0 + \alpha_1 z_0)]$. Third, a Bernoulli random variable was generated with marginal probability P for treatment at time 1, Z_1, conditional on the realized value of the time-varying confounder, ℓ, as $1/[1 + \exp(\beta_0 + \beta_1 \ell)]$. Fourth, a Weibull random variable was generated conditional on the realized value of the time-varying confounder ℓ with shape parameter $\lambda = \exp(-\gamma_0 - \gamma_1 \ell)$ and scale parameter κ, as $\kappa\lambda t^{\kappa-1}\exp(-\lambda t^{\kappa})$. The Weibull-distributed times were administratively censored such that, in expectation, about 15% of simulated subjects incurred events during follow-up. Finally, X_0 and X_1 were generated such that there was a sensitivity of 0.9 and specificity of 0.8. Calibration was conducted with external validation sample size equal to the study size.

We examined the scenario defined by $P = 1/2$, $\kappa = 2$, $\alpha_1 = \log(5)$, $\beta_1 = \log(5)$, and $\gamma_1 = \log(5)$, which we term the alternative hypothesis scenario because the total causal effect is nonnull (i.e., the expected causal hazard ratio is 1.7), and the scenario where $\alpha_1 = \log(1)$, which we term the null hypothesis scenario because the total causal effect is null.



Appendix Figure 1. Causal diagram depicting simulation data. *T* is time to acquired immunodeficiency syndrome or death, Z_j is the true exposure to highly active antiretroviral therapy for $j = \{0,1\}$, X_j is the measured exposure to highly active antiretroviral therapy, *L* are the time-varying confounders, and *U* are unmeasured determinants of the subscripted variable.

To compare the estimates, we calculated simulated bias, computed as the estimated log hazard ratio minus the true log hazard ratio; simulated standard error, computed as the average of the estimated standard errors; and simulated confidence limits coverage, computed as the proportion of times that the confidence limits contain the true hazard ratio. Simulation results are subject to Monte Carlo error; on the basis of the 10,000 simulations, the 95% confidence limits coverage estimates have a simulation standard error of approximately 0.2%.

For each of 10,000 simulation trials, we conducted 2 analyses estimating the association between cumulative average treatment and time to event. First, we estimated the association obtained from a standard marginal structural Cox proportional hazards model, as detailed in the main text as model 1. Second, we estimated the association obtained from a marginal structural measurement-error Cox proportional hazards model, as detailed in Appendix 2. Both results were compared with the total causal effect obtained as the indirect effect of initial treatment under the diagram shown in Figure 1.

All simulations converged. Appendix Table 1 shows that, under the alternative hypothesis, the standard marginal structural model is null biased but the marginal structural measurement-error model provides an unbiased estimate of the total causal effect, as well as appropriate confidence limits coverage, at a cost of reduced precision. Under the null hypothesis, as expected, both the standard and measurement-error models provided type 1 error rates within 2 simulation standard errors of the expected 5%.

Appendix Table 1. Simulated Bias, Robust Standard Error, and Confidence Limit Coverage for Standard Marginal Structural and Marginal Structural Measurement-Error Cox Models Under an Alternative and Null Hypothesis and Nondifferential and Independent Misclassification of Treatment, 10,000 Samples of Size 1,000

Marginal Structural Cox Model	Bias (MCSE)	SE (MCSE)	CL Coverage % (MCSE)[a]
Alternative hypothesis			
Standard	−0.162 (0.002)	0.213 (0.001)	88.4 (0.3)
Measurement error	−0.003 (0.003)	0.296 (0.001)	95.2 (0.2)
Null hypothesis			
Standard	0.001 (0.002)	0.215 (0.001)	4.6 (0.2)
Measurement error	−0.001 (0.003)	0.292 (0.001)	5.4 (0.2)

Abbreviations: CL, confidence limits; MCSE, Monte Carlo simulation error; SE, standard error.

[a] Type 1 error rate under null hypothesis.