Using marginal structural measurement-error models to estimate the long-term effect of antiretroviral therapy on incident AIDS or death

Stephen R. Cole¹, Lisa P. Jacobson², Phyllis C. Tien³, Lawrence Kingsley⁴, Joan S. Chmiel⁵, Kathryn Anastos⁶

¹Department of Epidemiology, University of North Carolina, Chapel Hill, NC
²Department of Epidemiology, Bloomberg School of Public Health, Baltimore, MD
³Department of Medicine, University of California, San Francisco, and San Francisco Veterans Affairs Medical Center, San Francisco, CA
⁴Department of Infectious Diseases and Microbiology, University of Pittsburgh, Pittsburgh, PA
⁵Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL
⁶Montefiore Medical Center and Albert Einstein College of Medicine, Bronx, NY

Word (abstract) count: 4045 (191)
<table>
<thead>
<tr>
<th>Acronyms</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDS</td>
<td>acquired immunodeficiency syndrome</td>
</tr>
<tr>
<td>HAART</td>
<td>highly active antiretroviral therapy</td>
</tr>
<tr>
<td>CD4</td>
<td>T-lymphocyte subtypes</td>
</tr>
<tr>
<td>CL</td>
<td>confidence limits</td>
</tr>
<tr>
<td>HR</td>
<td>hazard ratio</td>
</tr>
<tr>
<td>HIV-1</td>
<td>human immunodeficiency virus, type 1</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>IPTC</td>
<td>inverse probability-of-treatment-and-censoring</td>
</tr>
</tbody>
</table>

**Correspondence**

Dr. Stephen R. Cole, UNC School of Public Health  
MacGavran-Greenberg Hall, CB#7435  
Chapel Hill, NC 27599-7435  
Telephone (919) 966-7415, Fax (919) 966-2089, E-mail cole@unc.edu
Abstract

To estimate the net effect of imperfectly measured highly active antiretroviral therapy (HAART) on incident AIDS or death, the authors combined inverse probability-of-treatment-and-censoring (IPTC) weighted estimation of a marginal structural Cox model with regression-calibration methods. 950 HIV positive men and women were followed in two cohort studies between 1995 and 2007. During 5011 person-years, 417 initiated HAART, 248 developed AIDS or died, and 124 dropped out. Accounting for measured confounders and determinants of drop out, the weighted hazard rate ratio (HR) for AIDS or death of 0.30 (95% confidence limits [CL]: 0.18, 0.50) was relatively constant over follow up (P = 0.53), and twice as strong as crude or adjusted HRs of 0.64 and 0.72, respectively. Accounting for measurement error in reported HAART exposure using external validation data on 331 men and women provided a HR of 0.12; with bias shifted from the point estimate to the estimate of precision as seen by the 1.6-fold wider CL (95% CL: 0.06, 0.26). Such marginal structural measurement-error models can simultaneously account for three major sources of bias in epidemiologic research: validated exposure measurement error, measured selection bias, and measured time-fixed and time-varying confounding.

Key Words: Bias, confounding; Bias, measurement; Bias, selection; Causal inference; Cohort study; HIV/AIDS; Pharmacoepidemiology
Incident acquired immunodeficiency syndrome (AIDS) is a central clinical event in the progression of human immunodeficiency virus (HIV) infection. Due to a demonstrated (1-3) strong immediate protective effect of 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 (a) assumed once initiated on HAART individuals continue to use therapies (10, 11), (b) taken reported HAART use as measured without error (10), (c) followed participants for 6.5 (10) to 7.5 (11) years, and (d) 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 greater than 11 years. We allow individuals to move on and off 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 allow for 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, Chicago, Pittsburgh and Los Angeles; as well as the Women’s Interagency HIV Study (WIHS) (19), which beginning in 1994 enrolled 3,772 women in New York, Chicago, Los Angeles, San Francisco and Washington DC. Every six months, participants in both studies completed a physical examination, 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 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 FDA approved 6 December 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 he or she was seen alive without clinical AIDS, before initiation of non-HAART antiretroviral therapy or dropout (i.e., defined as two consecutive missed visits), or at the date of analysis in September 2007, whichever came 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 non-informative (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 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 < 200 cells/mm³ or CD4 percent < 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, while the cohort of women self-reported clinical AIDS. Deaths were ascertained 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, as 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 Department of Health and Human Services panel guidelines (24) as previously published (10). Typical HAART regimens consisted of 2+ nucleoside or nucleotide reverse transcriptase inhibitors in combination with 1+ protease inhibitor or one non-nucleoside 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 ml of plasma, was measured using an isothermal nucleic acid sequence-based amplification method for women (bioMérieux, Boxtel, NL) 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 (13 percent) 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 $J-1$. Therefore the time scale is time-on-study. Let $C_{ij} = 1$ indicate drop out 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 $ar{X} = ar{X}_j = \{X_{i0}, X_{i1}, \ldots, X_{ij}\}$ is the history of exposure to HAART through visit $j$.

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

$$\log \left\{ \frac{\Pr(Y_{ij}^\bar{X}, \bar{z} = 0) = 1}{1 - \Pr(Y_{ij}^\bar{X}, \bar{z} = 0) = 1} \right\} = b_{0j} + b_1 \bar{g}(\bar{x}),$$
where $Y_{ij}^{x,c=0}$ is a time-varying indicator of incident AIDS or death in the visit interval $(j, j+1]$ had the participant followed treatment history $x, c = 0$ 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 $\beta = \{\beta_{0j}, \beta_1\}$ by maximizing a weighted version of the Bernoulli likelihood function

$$L(\beta) = \prod_{i=1}^{N} \prod_{j=0}^{j-1} p_{ij}^{y_{ij}^\text{W}} \times (1-p_{ij})^{(1-y_{ij})^\text{W}_j},$$

where $p_{ij} = 1/(1 + \exp(-[\beta_{0j} + \beta_1 g(\bar{x}_{ij})]))$, $\text{W}_j$ are estimated time-varying IPTC weights (described below); and $\beta_{0j}$ are visit-specific intercepts fit using a restricted cubic spline with 4 knots at the 5, 33, 67 and 95 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 percent, which held in our example as the largest event proportion in any visit interval was 6 percent.

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

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

and
Marginal structural measurement-error model

\[ W_{ij}^C = \prod_{k=0}^j \frac{Pr[C_{ik} = 0 | C_{ik-1} = 0, X_{ik}]}{Pr[C_{ik} = 0 | C_{ik-1} = 0, X_{ik}, L_{ik-1}]} , \]

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., < 4,000; 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 \( \beta \) 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 reference (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 non-confounded weighted data, one is able to correct for misclassification of exposure using regression calibration. Details of the validation data are provided in Appendix 1; and details of regression calibration are provided in Appendix.
2. A limited Monte Carlo simulation demonstrating some finite sample properties of the proposed approach is provided in Appendix 3.

Based on 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 using SAS version 9 (Statistical Analysis System Inc, Cary, North Carolina), using robust variance estimates (31) to calculate confidence limits (CL) and P values for the marginal structural models (Appendix, (27)).

RESULTS

At study entry, the 950 participants had median (quartiles) age of 38 (33, 44) years, CD4 count of 453 (303, 641) cells/mm³, viral load of 4.5 (4, 4.9) log₁₀ copies/ml for the 73 percent with detectable values; 61 percent were women, and 41 percent were Caucasian (Table 1).

The 950 participants contributed 5,011 person-years under observation. The median (quartiles) length of follow up was 3.9 (0.7, 10.8) years over the follow up period between September 1995 and September 2007. The CD4 count averaged over follow up was 32 cells/mm³ higher and the HIV viral load was 0.4 log₁₀ copies/ml lower, than at baseline. 248 (26 percent) of participants developed AIDS (N = 213) or died (N = 35) during follow up, 316 (33 percent)
completed follow up alive, 262 (28 percent) were censored due to initiation of non-HAART antiretroviral therapy, and the remaining 124 (13 percent) were censored due to dropout. 417 of 950 participants (44 percent) initiated HAART during follow up. 2,412 of 5,011 person-years (48 percent) were contributed prior to HAART exposure, 2,262 person-years (45 percent) were contributed while exposed to HAART, and 336 person-years were contributed by participants following discontinuation of HAART. 2,536 of 5,011 person-years (51 percent) were fully unexposed to HAART in the prior 2 years and 1,540 of 5,011 person-years (31 percent) were fully exposed to HAART in the prior 2 years; the remaining 18 percent of 5,011 person-years were partially exposed to HAART in the prior 2 years (i.e., 313, 281 and 341 person-years were exposed to 25%, 50% and 75% HAART, respectively). The predicted probability of HAART use since the prior visit conditional on measured covariates ranged from $5.9 \times 10^{-9}$ to 0.99 with a mean of 0.38. After stabilization, the IPTC weights ranged from 0.17 to 18.54, with a mean of 0.98, 1st and 99th percentiles at 0.28 and 3.08 and quartiles at 0.67, 0.90, 1.10, respectively. As expected, among participants not using HAART at the prior visit, the odds of HAART use increased by a factor of 1.33 (95 percent CL: 1.28, 1.39) for each decrement of 100 CD4 cells/mm³ and by a factor of 1.29 (95 percent CL: 1.00, 1.66) for a detectable viral load.

The unadjusted hazard of AIDS or death was 0.64 (95 percent CL: 0.43, 0.94) among those using HAART during the prior 2 years, relative to those not using antiretroviral therapy (Table 2). The unadjusted hazard ratio of 0.64 was 1.13 times weaker after adjustment for time-varying CD4 cell count and viral load. The adjusted estimate was slightly less precise with a 1.11 times larger standard error (for the log of the hazard ratio, 0.220 versus 0.199).

The hazard of AIDS or death from the marginal structural model (i.e., weighted) was 0.30 (95 percent CL: 0.18, 0.50) for HAART use in the prior 2 years, relative to not using
antiretroviral therapy (Table 2). This weighted estimate was 2.13 and 2.4 times stronger than the unadjusted and adjusted estimates, respectively. However, the weighted estimate was less precise than the unadjusted estimate with a 1.28 times larger standard error (for the log of the hazard ratio, 0.255 versus 0.199). Truncating the IPTC weights at the 1st and 99th percentiles yielded a similar estimate (hazard ratio = 0.36; 95 percent CL: 0.23, 0.55) and precision (0.88 times smaller standard error for the log of the hazard ratio, 0.225, versus 0.255). 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 0.38 (95 percent CL: 0.25, 0.57) and 0.25 (95 percent CL: 0.15, 0.43), respectively.

The effect of HAART during the prior 2 years appeared stronger among men (hazard ratio = 0.26; 95 percent CL: 0.11, 0.61) than women (hazard ratio = 0.50; 95 percent CL: 0.28, 0.90), but this difference may have been due to chance (P homogeneity = 0.18). The effect of HAART was stronger at lower levels of baseline CD4 count (hazard ratios for <200, 200-350, 351-500, and >500 cells/mm³ at baseline were 0.14, 0.15, 0.34, and 0.46, respectively; P homogeneity = 0.05). The effect of HAART appeared stronger at earlier time-on-study, with a hazard ratio of 0.13 (95 percent CL: 0.03, 0.50) before 2 years and 0.33 (95 percent CL: 0.20, 0.55) after 2 years from study entry, but this difference may have been due to chance (P homogeneity = 0.19). Moreover, a test of the proportional hazards assumption not categorizing time yielded a P homogeneity = 0.53.

Using the pooled external validation data, the estimated calibration slope, $\hat{\gamma}_1$, was 0.57 (95 percent CL: 0.49, 0.65). In the pooled external validation data, the percent (± standard error) using HAART by medical records given no reported use was 35±5 percent and the percent using HAART by medical records given reported use was 92±2 percent. The hazard of AIDS or death
from regression-calibration of the marginal structural model (i.e., weighted and calibrated) was
0.12 ( = \exp\left[\frac{\ln(0.30)}{0.57}\right]; 95\% \text{ CL: 0.06, 0.26}) for HAART use during the prior 2
years relative to not using antiretroviral therapy (Table 2). This calibrated estimate was 2.5 times
stronger than the weighted estimate. The weighted and calibrated estimate was less precise than
the weighted estimate with a 1.44 times larger standard error (for the log of the hazard ratio,
0.368 versus 0.255). Restricting the external validation data to the MACS (hazard ratio = 0.23;
95\% \text{ CL: 0.14, 0.38}) or UNC (hazard ratio = 0.09; 95\% \text{ CL: 0.03, 0.22}) 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 provide a range within
which the true result likely resides.

DISCUSSION

Using a marginal structural Cox proportional hazards model and regression-calibration,
the authors estimated that relative to not using antiretroviral therapy, HAART use during the
prior 2 years decreases the hazard of AIDS or death by 89 percent (range of 95\% CL across
sensitivity analysis: 62, 97), and this effect appears to persist for over 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 to a combination therapy regimen. We would expect to see
stronger effects than these trials because (a) our comparison group is the absence of antiretroviral
therapy rather than a combination therapy, (b) the HAART regimens used over the course of follow up have improved while the trials both used single early HAART regimens, and finally, (c) 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 percent CL: 0.20, 0.61) in a comparison of the time period following HAART introduction with the time period of mono-therapy. We would expect to see 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 can be seen in Figures 1 of references (35, 37). Indeed, an instrumental variable correction for misclassification of the Detels et al result yielded a rate ratio of 0.2 (38), which is closer to our estimate. Finally, prior observational analyses accounting for time-varying confounding using marginal structural models (10, 11) have assumed once initiated on HAART individuals continue to use therapies (10, 11), have taken reported HAART use as measured without error (10), and estimated either the effect of HAART versus no therapy (hazard ratio = 0.14; 95 percent CL: 0.07, 0.29) (11), mono- or combination therapy (hazard ratio = 0.51, 95 percent CL: 0.29, 0.87) (10), combination therapy alone (hazard ratio = 0.49; 95 percent CL: 0.31, 0.79) (11), or any non-HAART therapy, including no therapy (hazard ratio = 0.54; 95 percent CL: 0.38, 0.78) (10). As expected, the exposure misclassification corrected hazard ratio of 0.12 (95 percent CL: 0.06, 0.26) presented here coheres well with previously reported hazard ratio of 0.14 (95 percent 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 percent CL: 0.24, 0.55).

Past work in marginal structural models has largely omitted discussion of the choice of
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 the choice of the final structural model is
needed and beyond the scope of this paper. For instance, the choice of time scale to be used
affects the meaning of survival curves and may affect the value of a summary hazard ratio. Also,
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, like all observational analyses, the estimates have a causal interpretation only under the
assumption of no unmeasured confounding. This assumption likely holds approximately here as
the most important clinical and laboratory information used by physicians as indications for
HAART were collected and used in the models for the estimation of 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*
 pneumona 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) as some participants with high CD4 counts and low viral load initiated HAART while others with low CD4 and high viral loads did not. Please see reference (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 the semiannual data on covariates obtained from these interval cohorts (45) is 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). Further, the calibrated results rely on the assumption that the external validation data for HAART use is accurate. Beyond the simple accuracy of the 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 proxies...
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 does provide some
reassurance. There is little prior published work 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), this 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 null-bias in the point estimate to a less biased corrected estimate with
increased uncertainty.

In conclusion, the observed association of HAART with incident AIDS or death appears
to persist for greater 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
Marginal structural measurement-error model 18

of exposure and detailed collection of clinical and laboratory information provide the best
evidence available for the estimation of risk group-specific, long-term therapeutic effects.
APPENDIX 1: Validation Data

The mapping between the misclassified and true HAART exposure is taken from two reports of validation data (53, 54). In the 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 1 October 2004 to 1 April 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 to meet the following criteria was selected: occurred between 1 October 2003 and 1 April 2008, with a subsequent MACS study visit between 1 October 2004 and 1 April 2008, such that the difference between clinic and study visits was less than one year (with a single exception of 1.16 years). In the MACS (and WIHS), 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 sub study, all antiretroviral medications that the participant was continuing or starting at the clinic visit were abstracted. Using the MACS algorithm, each participant’s drug combination was classified as HAART or non-HAART. Using medical record abstracted HAART 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, positive and negative predictive values (±standard error) of self-reported HAART use were 97±2, 82±8, 96±2, and 86±8, respectively. This validation sample of 126 men was aged 49±8y and 55 percent 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 socio-demographic survey when seen between January 2001 and May 2006 at the UNC HIV clinic in Chapel Hill. In the UNC clinic cohort, each antiretroviral medication that the participant is using at a clinic visit was abstracted from medical records. In the validation sub study, participants were asked on the survey to report current use of each antiretroviral medication. Using medical record abstracted therapies as a gold standard, a participant using antiretroviral therapy who correctly reported use of all their 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 when the study took place (i.e., 2001 to 2006), the vast majority of HIV patients using antiretroviral therapies would be 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, positive and negative predictive values of self-reported therapy use were 75±4, 79±5, 88±3, and 60±5, respectively. This validation sample of 205 participants had a median (quartiles) age of 42 (36, 47) years, were 66 percent male, and 71 percent African-American.

The pooled MACS and UNC data yielded sensitivity and specificity of 84±2 and 80±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 UNC and MACS validation data separately.
Regression calibration proceeds with an estimate of the discrete-time log hazard ratio $\hat{\beta}$ between the misclassified exposure and clinical AIDS or death, which is given by the pooled logistic model described in the Statistical Analysis section.

Second, a linear calibration model, $E(Z_m | X_m) = \gamma_0 + \gamma_1 X_m$, is fit to the pooled external validation data (see 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 $\hat{\theta}_i = \hat{\beta}_i / \hat{\gamma}_1$, with 95 percent CL for $\hat{\theta}_i$ obtained using a variance of $\hat{V}(\hat{\theta}_i) = \hat{\gamma}_1^{-2} \hat{V}(\hat{\beta}_i) + \left(\hat{\beta}_i / \hat{\gamma}_1^2\right) \hat{V}(\hat{\gamma}_1)$, which is a first-order approximation using the delta method but with a robust variance taken for the discrete-time log hazard ratio.
Data are simulated from the un-weighted 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, $\ell$, 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 $\ell$ with shape parameter $\lambda = \exp(-\gamma_0 - \gamma_1 \ell)$ and scale parameter $\kappa$, as $\kappa \lambda \ell^{\kappa-1} \exp(-\lambda \ell^\kappa)$. The Weibull-distributed times were administratively-censored such that, in expectation, about 15 percent of simulated subjects incurred events during follow up. Finally, $X_0$ and $X_1$ were generated such that there was 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 non-null (i.e., 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.

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 CL coverage, computed as the proportion of times the confidence limits contain the true hazard ratio. Simulation results are subject to Monte Carlo error; based on the 10,000 simulations, the 95 percent CL coverage estimates have a simulation standard error of \( \approx 0.2 \) percent.

For each of 10,000 simulation trials, we conducted two 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 to the total causal effect obtained as the indirect effect of initial treatment under the diagram shown in Figure 1.

All simulations converged. In Appendix Table 1 we see, under the alternative hypothesis, that 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 CL 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 percent.
# TABLE 1: Baseline and follow-up characteristics of 950 men and women infected with human immunodeficiency virus type 1, Multicenter AIDS Cohort Study and Women’s Interagency HIV Study, 1995-2007

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Baseline (1995) (n = 950 persons)</th>
<th>Follow up (1995-2007) (n = 11,102 person-visits&lt;sup&gt;a&lt;/sup&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%</td>
<td>Mean</td>
</tr>
<tr>
<td>Mean age (years)</td>
<td></td>
<td>39 (8)&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Female sex</td>
<td>61</td>
<td>578</td>
</tr>
<tr>
<td>African American</td>
<td>59</td>
<td>560</td>
</tr>
<tr>
<td>Use of HAART</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Category of CD4 cell count (no. cells/mm&lt;sup&gt;3&lt;/sup&gt;):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 200</td>
<td>11</td>
<td>109</td>
</tr>
<tr>
<td>200 – 350</td>
<td>21</td>
<td>195</td>
</tr>
<tr>
<td>351 – 500</td>
<td>26</td>
<td>251</td>
</tr>
<tr>
<td>&gt; 500</td>
<td>42</td>
<td>395</td>
</tr>
<tr>
<td>Mean CD4 cell count (no. cells/mm&lt;sup&gt;3&lt;/sup&gt;)</td>
<td>498 (279)</td>
<td>530 (298)</td>
</tr>
<tr>
<td>Category of HIV-1 RNA level (no. copies/ml):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 4,001</td>
<td>34</td>
<td>324</td>
</tr>
<tr>
<td>4,001 – 10,000</td>
<td>13</td>
<td>120</td>
</tr>
<tr>
<td>&gt; 10,000</td>
<td>53</td>
<td>506</td>
</tr>
<tr>
<td>Mean log&lt;sub&gt;10&lt;/sub&gt; HIV-1 RNA level (co. copies/ml)</td>
<td>4.5 (0.7)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>4.1 (0.7)&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

<sup>a</sup> 950 people contributed 11,102 person-visits with 5010.8 person-years of follow up

<sup>b</sup> Number in parentheses, standard deviation

<sup>c</sup> Among 691 and 6,123 detectable measurements at baseline and follow up, respectively
TABLE 2: Effect of highly active antiretroviral therapy on incident AIDS or death, 950 men and women infected with human immunodeficiency virus type 1, Multicenter AIDS Cohort and Women’s Interagency HIV Studies, 1995-2007

<table>
<thead>
<tr>
<th>Models</th>
<th>Exposure</th>
<th>Hazard Ratio</th>
<th>95% CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unadjusted</td>
<td>No ART</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>HAARTa</td>
<td>0.64</td>
<td>0.43, 0.94</td>
</tr>
<tr>
<td>Adjusted c</td>
<td></td>
<td>0.72</td>
<td>0.47, 1.11</td>
</tr>
<tr>
<td>Weighted c</td>
<td></td>
<td>0.30</td>
<td>0.18, 0.50</td>
</tr>
<tr>
<td>Weighted &amp; Calibrated</td>
<td></td>
<td>0.12</td>
<td>0.06, 0.26 d</td>
</tr>
</tbody>
</table>

a Use during the prior 2 years  
b Confidence limits, CL; robust for weighted models  
c Adjusted and weighted models control for time-varying prior CD4 cell count and HIV-1 RNA level using restricted cubic splines  
d CL obtained by from delta method using robust variance
APPENDIX FIGURE 1: Causal diagram depicting simulation data. T is time to AIDS or death, Z_j is the true exposure to HAART for j = \{0,1\}, X_j is the measured exposure to HAART, L are the time varying confounders, and U are unmeasured determinants of the subscripted variable.
APPENDIX TABLE 1: Simulated bias, robust standard error (SE) and confidence limit (CL) 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, 10000 samples of size 1000

<table>
<thead>
<tr>
<th>Marginal Structural Cox Models</th>
<th>Bias</th>
<th>SE</th>
<th>CL coverage, % a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative hypothesis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>-0.162±0.002</td>
<td>0.213±0.001</td>
<td>88.4±0.3%</td>
</tr>
<tr>
<td>Measurement-error</td>
<td>-0.003±0.003</td>
<td>0.296±0.001</td>
<td>95.2±0.2%</td>
</tr>
<tr>
<td>Null hypothesis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard</td>
<td>0.001±0.002</td>
<td>0.215±0.001</td>
<td>4.6±0.2%</td>
</tr>
<tr>
<td>Measurement-error</td>
<td>-0.001±0.003</td>
<td>0.292±0.001</td>
<td>5.4±0.2%</td>
</tr>
</tbody>
</table>

a Type 1 error rate under null hypothesis
ACKNOWLEDGEMENTS

Dr. Stephen Cole was partially support by NIH grants R03-AI-071763, R01-AA-017594 and P30-AI-50410. The authors would like to thank Dr. Lauren E. Cain and Dr. Emily S. Brouwer for assistance with the validation data and Dr. Miguel A. Hernán, Dr. Daniel Westreich and Dr. Sander Greenland for expert advice.

Data in this manuscript were collected by the Women's Interagency HIV Study (WIHS) Collaborative Study Group with centers (Principal Investigators) at New York City/Bronx Consortium (Kathryn Anastos); Brooklyn, NY (Howard Minkoff); Washington DC Metropolitan Consortium (Mary Young); The Connie Wofsy Study Consortium of Northern California (Ruth Greenblatt); Los Angeles County/Southern California Consortium (Alexandra Levine); Chicago Consortium (Mardge Cohen); Data Coordinating Center (Stephen Gange). The WIHS is funded by the National Institute of Allergy and Infectious Diseases (U01-AI-35004, U01-AI-31834, U01-AI-34994, U01-AI-34989, U01-AI-34993, and U01-AI-42590) and by the National Institute of Child Health and Human Development (U01-HD-32632). The study is co-funded 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 Number UL1 RR024131). 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.

The Multicenter AIDS Cohort Study (MACS) includes the following: Baltimore: The Johns Hopkins University 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,
REFERENCES


