American Journal of Epidemiology

Vol. 188, No. 8
DOI: 10.1093/aje/kwz100
Advance Access publication:
May 7, 2019

Practice of Epidemiology

Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples With Apples

Sara Lodi*, Andrew Phillips, Jens Lundgren, Roger Logan, Shweta Sharma, Stephen R. Cole, Abdel Babiker, Matthew Law, Haitao Chu, Dana Byrne, Andrzej Horban, Jonathan A. C. Sterne, Kholoud Porter, Caroline Sabin, Dominique Costagliola, Sophie Abgrall, John Gill, Giota Touloumi, Antonio G. Pacheco, Ard van Sighem, Peter Reiss, Heiner C. Bucher, Alexandra Montoliu Giménez, Inmaculada Jarrin, Linda Wittkop, Laurence Meyer, Santiago Perez-Hoyos, Amy Justice, James D. Neaton, and Miguel A. Hernán, on behalf the INSIGHT START Study Group and the HIV-CAUSAL Collaboration

* Correspondence to Dr. Sara Lodi, Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, MA 02118 (e-mail: slodi@bu.edu).

Initially submitted October 8, 2018; accepted for publication April 17, 2019.

Effect estimates from randomized trials and observational studies might not be directly comparable because of differences in study design, other than randomization, and in data analysis. We propose a 3-step procedure to facilitate meaningful comparisons of effect estimates from randomized trials and observational studies: 1) harmonization of the study protocols (eligibility criteria, treatment strategies, outcome, start and end of follow-up, causal contrast) so that the studies target the same causal effect, 2) harmonization of the data analysis to estimate the causal effect, and 3) sensitivity analyses to investigate the impact of discrepancies that could not be accounted for in the harmonization process. To illustrate our approach, we compared estimates of the effect of immediate with deferred initiation of antiretroviral therapy in individuals positive for the human immunodeficiency virus from the Strategic Timing of Antiretroviral Therapy (START) randomized trial and the observational HIV-CAUSAL Collaboration.

antiretroviral initiation; causal inference; per-protocol effect; target trial

Abbreviations: AIDS, acquired immune deficiency syndrome; ART, antiretroviral treatment; CI, confidence interval; HIV, human immunodeficiency virus; START, Strategic Timing of Antiretroviral Therapy.

Randomized trials and observational studies are used to estimate the comparative effectiveness and safety of clinical strategies. When a randomized trial and an observational study address a similar question, discrepancies between their effect estimates tend to be attributed to uncontrolled confounding (due to imbalance of prognostic factors between the treatment groups) in the observational study. However, such discrepancies can also be explained by differences in study design and data analysis.

For example, the randomized-observational discrepancy for the effect of postmenopausal estrogen plus progestin therapy on coronary heart disease was explained largely by selection bias (because follow-up in the observational study started some time after initiation of therapy), whereas unmeasured confounding seemed to play a lesser role (1, 2). As another example, randomized trials tend to use intention-to-treat estimates that quantify the effect of being assigned to treatment, regardless of whether treatment is actually received, whereas many observational studies quantify the effect of the treatment that was actually received (3).

If differences other than randomization are not explicitly taken into account, randomized-observational comparisons, as commonly undertaken in meta-analyses (4–8), might be hard to interpret because they generally compare "apples with oranges" rather than "apples with apples." Informative comparisons between randomized and observational estimates will often require a careful reanalysis of the data of both the randomized trials and the observational studies.

We describe a systematic approach to improve the comparison of effect estimates from a randomized trial and an observational study. Our approach has 3 stages: 1) harmonization of the study protocols to ensure that the studies target the same causal effect, 2) harmonization of the data analysis to target a common estimand, and 3) sensitivity analyses to investigate the impact of any remaining discrepancies.

We illustrate our systematic approach through a case study: a comparison of the International Network for Strategic Initiatives in Global HIV Trials (INSIGHT) Strategic Timing of Antiretroviral Therapy (START) randomized trial (9) and an observational analysis of routinely collected data in the HIV-CAUSAL Collaboration (https://www.hsph.harvard.edu/causal/hiv/) (10). Both studies compared the effectiveness of strategies for initiation of antiretroviral treatment in human immunodeficiency virus (HIV)-positive individuals. Both studies found that immediate initiation of antiretroviral therapy was beneficial, but the magnitude of the estimated benefit appeared to differ.

CASE STUDY: INITIATION OF ANTIRETROVIRAL THERAPY IN HIV-POSITIVE INDIVIDUALS

Antiretroviral therapy (ART) is a life-long treatment for HIV-positive individuals (11, 12). Historically, the decision to initiate ART was guided by the CD4-cell count (low levels indicate severe immunosuppression). During the 2000s, a key question was the CD4 count at which ART should be initiated. Results from randomized trials (9, 13–15) and observational studies (6, 9, 10, 13–21) led to the now widely accepted conclusion that ART should be initiated as soon as possible after diagnosis of HIV infection. The 2 most recent studies, the randomized START trial (9) and the observational HIV-CAUSAL Collaboration (10), compared the effectiveness of immediate initiation regardless of CD4 count versus deferred initiation until CD4 count dropped below 350 cells/mm^3 or acquired immunodeficiency syndrome (AIDS) was diagnosed in HIV-positive, AIDS-free, and treatment-naive individuals with CD4 counts of >500 cells/mm^3 at the start of the study.

The START trial included 4,685 individuals from low-, middle-, and high-income countries. The intention-to-treat hazard ratio for immediate versus delayed initiation for the primary outcome (the earliest of any serious AIDS-related event, serious non-AIDS-related event, or death) was 0.43 (95% confidence interval (CI): 0.30, 0.62), and the per-protocol hazard ratio was 0.34 (95% CI: 0.21, 0.52) (22).

The HIV-CAUSAL study included 17,612 individuals from cohorts in 9 countries in Europe and the Americas (23–25). All cohorts record routinely collected clinical data on patient characteristics, ART use, CD4 count, HIV-RNA, AIDS-defining illnesses, and deaths. The 7-year risk ratio of AIDS or death for immediate versus deferred initiation was 0.66 (95% CI: 0.56, 0.75), and the risk difference was 2.5% (95% CI: 1.8, 3.2).

At a first glance, the estimated effect of immediate initiation appeared more beneficial in the randomized trial than the observational study. However, the effect estimates were not directly comparable, because the 2 studies presented several key differences, which are summarized in the outer columns of Web Figure 1 (available at https://academic.oup.com/aje). In the next sections, we describe a process to harmonize their study design and data analysis.

STAGE 1: HARMONIZATION OF STUDY PROTOCOLS

The first stage of our systematic approach requires an explicit description of the protocol of a pragmatic randomized trial that is as similar as possible to the original trial and that the observational analysis will attempt to emulate—the target trial (26). The key components of the protocol of the target trial that need to be specified are eligibility criteria, outcome, treatment strategies, start/end of follow-up, causal contrast, and statistical analysis.

In our case study, we defined the target trial protocol for HIV-CAUSAL to closely resemble the protocol of START. The central columns of Web Figure 1 summarize the harmonization of the protocols of START and of the target trial emulated by HIV-CAUSAL. Hereafter, we refer to them as the "actual" and "emulated" trials. The harmonization resulted in close, but not identical, protocols. For several components of the protocol, we had to find a reasonable compromise, as described below.

Eligibility criteria

The START trial required 2 CD4 counts of >500 cells/mm^3 at least 14 days apart within 60 days before randomization. In clinical practice, CD4 count is typically measured every 90–180 days, and measurements 14–60 days apart are rare. As a compromise, the protocol of the emulated trial was modified to include individuals with at least 2 CD4 counts of >500 cells/mm^3 within 90 days of each other. Baseline was defined as the randomization date in the actual trial and as the date of the second CD4 count of >500 cells/mm^3 in the emulated trial. We excluded 9 START participants with no baseline HIV-RNA measurement within 60 days before randomization.

START recruited participants from clinics in high-, middle-, and low-income countries in 2009–2013, while HIV-CAUSAL included data from mostly high-income countries in 2000–2013. Restricting the actual trial to high-income countries would have resulted in too few events, so we did not impose geographic constraints in either study (and we added Brazil to the emulated trial). Restricting to 2009–2013 was not possible in the emulated trial because of the substantial reduction in follow-up, so we restricted the emulated trial to 2005–2013 as a compromise that resulted in comparable average follow-up between studies.

Table 1 displays baseline characteristics of the 4,676 eligible individuals in the actual trial and the 14,595 in the emulated trial after harmonization. Participants in the 2 studies had similar distributions of baseline CD4 count, HIV-RNA, and age. The actual trial included a larger proportion of women and heterosexuals. The distribution of sex and risk group in the subset of 2,769 START participants in high-income countries was comparable to that in the emulated trial (Web Table 1).

Treatment strategy

Before harmonization, the definition of the treatment strategies differed slightly between the actual and emulated trial protocols, and the grace period during which an individual should initiate treatment was not specifically defined in the original START trial, while it was 6 months in the observational study. Because in practice it might take several weeks before treatment is started due to clinical tests and administrative procedures,

Table 1. Participants' Characteristics at Baseline After Harmonization in Strategic Timing of Antiretroviral Therapy (Actual Trial) and the HIV-CAUSAL[a] (Emulated Trial), Using Data From Multiple Countries, 2005–2013

Characteristic	Actual Trial (n = 4,676)			Emulated Trial (n = 14,595)		
	Median (IQR)	No.	%	Median (IQR)	No.	%
CD4 cell count[b], cells/mm^3	651 (584–765)			559 (585–779)		
Enrollment year	2012 (2011–2013)			2009 (2007–2011)		
Age, years	36 (29–44)			36 (29–43)		
HIV-RNA, copies/mL	12,759 (3,019–43,391)			17,469 (4,300–57,539)		
Female sex		1,253	27		2,079	14
HIV-acquisition risk group						
MSM		1,787	38		9,204	63
MSW or WSM		2,581	55		3,306	23
IDU		64	1		391	3
Other/Unknown		244	5		1,694	11
High-income setting[c]		2,769	59		14,212	97

Abbreviations: HIV, human immunodeficiency virus; IDU, injecting drug use; IQR, interquartile range; MSM, men who have sex with men; MSW, men who have sex with women; WSM, women who have sex with men.

[a] https://www.hsph.harvard.edu/causal/hiv/.

[b] Average of 2 baseline values.

[c] Based on World Bank classification (33).

we defined the grace period to be 1 month. We then defined the 2 treatment initiation strategies to be identical in the actual and emulated trial. In both studies, the strategies did not prescribe a particular pattern of treatment adherence after initiation. Predictors of protocol deviation in the START trial are described elsewhere (22). In the emulated trial, individuals who initiated ART within 1 month of baseline had similar characteristics to those who initiated ART later or never initiated ART (Web Table 2).

Randomized assignment

Randomized assignment to treatment strategies is the fundamental distinction between randomized and observational studies. In START, individuals were randomly allocated to one of the 2 treatment strategies, which leads to the expectation of no unmeasured confounding at baseline (i.e., that the 2 groups are exchangeable at baseline, although not necessarily at later follow-up times) (27). In HIV-CAUSAL, individuals were not randomly allocated so we assumed no unmeasured confounding at baseline conditional on measured prognostic factors that influence the timing of treatment such as CD4 count, HIV-RNA, age, sex, mode of HIV acquisition, and calendar year. This assumption cannot be empirically verified.

Follow-up

In the actual and emulated trials, follow-up started at baseline and ended at the earliest of outcome occurrence, loss to follow-up, or end of the study. Because the estimation of the per-protocol effect in both studies requires adjustment for post-baseline CD4 count and HIV-RNA, loss to follow-up was defined in the actual and emulated trial as 12 months without one of these measurements. After harmonization, the median follow-up was 35 (interquartile range, 26–47) months in the actual trial and 32 (interquartile range, 16–58) months in the emulated trial. The proportion of individuals lost to follow-up in the first 5 years was 8% in the actual trial and 35% in the emulated trial.

Outcome

The original outcome was a composite endpoint encompassing serious AIDS-related event, serious non-AIDS events, and death in START and the earlier of death or any AIDS diagnosis in HIV-CAUSAL. Because information on non-AIDS events was not available in HIV-CAUSAL, the harmonized outcome definition was the same as in the original HIV-CAUSAL study (28). Because the START outcomes were restricted to adjudicated events only (9, 22), but AIDS events were not adjudicated in HIV-CAUSAL, we further defined the outcome to include any AIDS or death event regardless of adjudication. After harmonization, there were 112 outcome events over 14,196 person-years in the actual trial and 422 cases over 41,262 person-years in the emulated trial. The median CD4 counts at which events occurred were 573 (interquartile range, 444–711) cells/mm^3 in the actual trial and 560 (interquartile range, 426–700) cells/mm^3 in the emulated trial.

Causal contrast

The original studies used different causal contrasts. The original analysis of START estimated the intention-to-treat effect (9), whereas the HIV-CAUSAL study estimated the observational analog of the per-protocol effect: the effect that would have been observed under perfect adherence to the protocol (29). Because the magnitude of the intention-to-treat effect depends on the study-specific degree of adherence

to the protocol, we chose to estimate the per-protocol effect in both the actual and emulated trials.

STAGE 2: HARMONIZATION OF DATA ANALYSIS

The second stage of our systematic approach requires a reanalysis of both studies under the common target trial protocol. Specifically, for both studies, valid estimation of the per-protocol effect requires adjustment for baseline and postbaseline prognostic factors that predict treatment initiation and loss to follow-up. Because conventional methods cannot appropriately handle postbaseline prognostic factors that affect treatment status and are also affected by past treatment (i.e., treatment-confounder feedback), g-methods such as inverse probability weighting or the g-formula should be used instead (27).

In our case study, we assumed that the baseline variables in Table 1 and the postbaseline values of CD4 count, HIV-RNA, and the timing of CD4 count and HIV-RNA measurement were sufficient to adjust for postbaseline confounding. We used the parametric g-formula to estimate the per-protocol effect in both the actual and emulated trials. This method was used in the original analysis of HIV-CAUSAL (10) and in an analysis of START conducted after the primary paper was published (22).

The parametric g-formula is a generalization of standardization for time-varying treatments and confounders (30). It can be used to estimate the risk of the outcome that would have been observed if all individuals in the study had adhered to a particular treatment strategy and none had been lost to follow-up, under the assumptions of no residual confounding and selection bias, no measurement error, and no model misspecification. Briefly, the estimation procedure has 2 steps (10, 21, 31). First, we fitted separate regression models for each of the postbaseline variables and for the outcome variable at each month as a function of previous treatment and covariate history and of baseline covariates. Second, for each treatment strategy, these models were used to simulate the outcome risk.

For the first step, in both the actual and emulated trials, we fitted separate logistic regression models for time-varying indicators of measurement of HIV-RNA, measurement of CD4 count, ART initiation, and the outcome and linear regression models for CD4 count and HIV-RNA on the natural logarithm scale. All models included as covariates restricted cubic splines with 5 knots (32) of the most recent value of CD4 count, HIV-RNA, and time since last CD4 count and HIV-RNA measurements, as well as the following baseline variables: CD4 count, HIV-RNA, age, sex, mode of HIV acquisition, and calendar year. In addition, models for the trial data adjusted for income status of country (high- vs. middle- and low-income, according to the World Bank definition (33)), and models for the observational data adjusted for geographical origin and cohort. All models included a product ("interaction") term for number of months since treatment initiation. The placement of the knots of the splines and the thresholds for the baseline categories differed in the 2 studies. Nonparametric bootstrapping based on 500 samples was used to compute 95% confidence intervals based on the percentiles of the bootstrap distribution.

To explore the validity of our parametric assumptions, in both studies we compared the observed means of the outcome and time-varying covariates with those predicted by our models. The time-varying means predicted by our models under observed ART initiation were similar to the observed means in the original data (Web Figures 2–4). All analyses were conducted using the publicly available the GFORMULA_RCT and GFORMULA macros for SAS (SAS Institute, Inc., Cary, North Carolina) (34, 35)

Table 2 shows the estimated per-protocol effects of immediate versus deferred treatment initiation in the harmonized studies. The estimated 5-year risk of AIDS or death under deferred treatment initiation was 6.0% (95% CI: 4.4, 8.1) in the actual trial and 5.1% (95% CI: 4.4, 5.7) in the emulated trial. The corresponding estimated risk under immediate treatment initiation was higher in the emulated trial (3.0%, 95% CI: 2.3, 3.7) than in the actual trial (1.8%, 95% CI: 1.1, 2.6). As a consequence, the emulated trial estimated a smaller benefit of immediate initiation than the actual trial: a 5-year reduction in the absolute scale of 2.1 (95% CI: 1.1, 3.1) percentage points versus a reduction of 4.2 (95% CI: 2.5, 6.3) percentage points. The estimated hazard ratio for deferred versus immediate treatment was 3.4 (95% CI: 2.1, 6.2) in the actual trial and 1.63 (95% CI: 1.3, 2.3) in the emulated trial. The proportions of individuals who had initiated treatment under the deferred treatment initiation were comparable in the 2 studies (Web Figure 5). Because of the

Table 2. Per-Protocol Effect Estimates of the 5-Year Risk, Risk Difference and Hazard Ratios of AIDS or Death in Strategic Timing of Antiretroviral Therapy (Actual Trial) and in HIV-CAUSAL[a] (Emulated Trial) After Harmonization, Using Data From Multiple Countries, 2005–2013

Treatment Strategy	Risk (%)	95% CI	RD (%) (Deferred − Immediate)	95% CI	HR (Deferred ÷ Immediate)	95% CI
Actual trial[b]						
Immediate treatment	1.8	1.1, 2.6	0	Referent	1.0	Referent
Deferred treatment	6.0	4.4, 8.1	4.2	2.5, 6.3	3.4	2.1, 6.2
Emulated trial[c]						
Immediate treatment	3.0	2.3, 3.7	0	Referent	1.0	Referent
Deferred treatment	5.1	4.4, 5.7	2.1	1.1, 3.1	1.6	1.3, 2.3

Abbreviations: CI, confidence interval; HR, hazard ratio; RD, risk difference.
[a] https://www.hsph.harvard.edu/causal/hiv/.
[b] n = 4,676; 112 events.
[c] n = 14,595; 422 events.

definition of the intervention, corresponding proportions under immediate treatment initiation were 100% in both studies 1 month after baseline. These results were robust to the choice of placement of the spline knots.

STAGE 3: SENSITIVITY ANALYSES TO INVESTIGATE REMAINING DISCREPANCIES

The final stage of our systematic approach identifies components of the protocol that could not be fully harmonized and that, therefore, might explain the differences in effect estimates between the actual and emulated trials. Then a set of sensitivity analyses were conducted to explore the impact on the nonharmonized components on the effect estimates. For our case study, Table 3 lists components of the protocol that we could not fully harmonize.

Eligibility criteria

The 2 studies might differ in the distribution of treatment effect modifiers. For example, the actual trial included a larger proportion of women, heterosexual individuals, and individuals with baseline date on or after 2009 than the emulated trial. Because subgroup analyses are unfeasible given the large number of strata defined by these characteristics, we equalized (standardized) the distribution of the measured baseline factors between studies via the g-formula (this can also can be achieved via inverse-probability weighting (36, 37)). We simulated the risk in the subset of individuals in high-income countries in the actual trial as if the joint distribution of measured covariates had been that of the emulated trial. The procedure is illustrated in the flowchart in Web Figure 6. Any discrepancy between the original and standardized g-formula estimates can be attributed to differences in baseline characteristics.

After standardization to the randomized trial's baseline distribution, the estimated 5-year risk of AIDS or death in the actual trial was 3.6% (95% CI: 2.7, 4.8) under the immediate treatment strategy and 4.8% (95% CI: 4.3, 5.5) under the deferred treatment strategy. The similarity of these estimates with those reported in Table 2 suggests that the observed discrepancy cannot be fully explained by differences in the distribution of the measured factors at baseline.

Treatment strategies

The implementation of the treatment strategies in the actual and emulated trials might have differed if the pattern of treatment discontinuation after treatment initiation varied between the studies. Because data on adherence after initiation was not available in HIV-CAUSAL, we compared the proportion of individuals with virological suppression (HIV-RNA of <50 copies/mL), a proxy of adherence to ART, between both studies up to 5 years for each month after baseline. This proportion was similar in the 2 studies under deferred treatment initiation, but it was lower in the emulated trial under immediate initiation (Web Figure 7). Therefore, differential adherence after initiation might, in part, explain the discrepancy.

The composition of ART regimens might have differed between the 2 studies. The proportions of individuals who initiated ART with a protease inhibitor regime and with a nonnucleoside reverse transcriptase inhibitor regime were 20% and 73% in the actual trial and 35% and 58% in the emulated trial. The proportion of individuals who initiated ART with an integrase inhibitor regime was similar in both studies (8% and 7%). Because nonnucleoside reverse transcriptase inhibitor and protease inhibitor regimes are, in general, similarly effective at controlling viral replication (38, 39), differences in initial ART regimes are unlikely to explain the discrepancy.

Assignment procedure

While the presence of unmeasured confounding cannot be empirically shown, there are indirect ways to explore this issue. For example, a difference in effect estimates early in the follow-up between the studies is suggestive of unmeasured confounding. In the actual trial, the 1-year risk was 0.6% (95% CI: 0, 1.3) lower under immediate initiation (and the 2-year risk was 1.7% (95% CI: 1.0, 2.7) lower). In the emulated trial, no benefit was estimated for the 1-year risk (and only 0.4% (95% CI: 0, 0.9) for 2-year risk). See Web Table 3 for more detailed results. This difference suggests that some individuals who started treatment early in the emulated trial might have had a worse prognosis in a way that was not captured in the data. This confounding might also partly explain why the effect estimates are attenuated in the emulated trial compared with the actual trial.

Follow-up

The proportion of individuals lost to follow-up at 24 months was 8% in the actual trial and 20% in the emulated trial. Because loss to follow-up in the emulated trial ranged between 14% and 41% across cohorts, we repeated analyses restricted to the 4 cohorts with lowest follow-up rate (Swiss HIV Cohort Study, CoRIS, PISCIS, and French Hospital Database—members of the HIV-CAUSAL Collaboration (https://www.hsph.harvard.edu/causal/hiv/)): risk of AIDS or death was 3.0% (95% CI: 2.1, 3.9) for immediate initiation and 4.6% (95% CI: 3.8, 5.5) for deferred initiation, which are similar to those estimated in Table 2 (3.0% and 5.0%). Also, because the higher loss to follow-up in the emulated trial might be the result of including some individuals not fully engaged in HIV care, we conducted analyses excluding individuals who were lost early. The risks of AIDS or death were 3.1% (95% CI: 2.3, 3.8) for immediate initiation and 5.2% (95% CI: 4.5, 5.7) for deferred initiation after excluding individuals who were lost by 12 months, and 3.3% (95% CI: 2.5, 4.1) and 5.5% (95% CI: 4.8, 6.0) after excluding individuals who were lost by 24 months. In summary, differences in loss to follow-up seem unlikely to explain the discrepancy.

Outcome

The emulated trial included only centers from high-income countries where tuberculosis is rare, while the actual trial included data from middle- and low-income countries where

Table 3. Possible Explanations for Differences in Estimates From the Strategic Timing of Antiretroviral Therapy (Randomized) and the HIV-CAUSAL[a] (Observational Study) After Harmonization of Study Design and Statistical Analysis

Protocol Components	Potential Remaining Differences	Examples	Proposed Sensitivity Analyses
Eligibility criteria	Differences in the patient mix	START included more women and heterosexual individuals.	Standardization
Treatment strategy	Differences in treatment uptake	Individuals in START might be more adherent than individuals in HIV-CAUSAL.	Compare treatment adherence (or a proxy) in the observed and emulated trials at 1, 2, 3, 4, and 5 years.
		Individuals in the 2 studies might have received ART combinations with different efficacy.	Compare distribution of initial ART combination.
Assignment procedures	Confounding by indication	Individuals who started ART with high CD4 count in HIV-CAUSAL might have worse prognosis.	Estimate and compare treatment effects in the two studies at 1, 2, 3, 4 and 5 years.
Follow-up	Differential loss to follow-up	In HIV-CAUSAL, individuals who are lost to follow-up tend to have high CD4 count.	Reanalysis excluding individuals who were lost to follow-up in the first 12 or 24 months since baseline
Outcome	Differences in baseline risk for the outcome	Individuals in START might have a higher risk of tuberculosis than in HIV-CAUSAL.	Reanalysis excluding cases of tuberculosis as events (if possible)

Abbreviations: ART, antiretroviral therapy; START, Strategic Timing of Antiretroviral Therapy.
[a] https://www.hsph.harvard.edu/causal/hiv/.

tuberculosis is more common. In the harmonized analyses, 28% of outcome events in the actual trial were tuberculosis, but this rate was only 5% in the emulated trial. Unlike other opportunistic infections, tuberculosis can occur early in the course of the HIV infection and at high CD4 cell counts (40, 41). An obvious sensitivity analysis would have been to redefine the outcome excluding tuberculosis, but the small number of outcome events in the immediate initiation group of START prevented us from doing this. A differential effect of the compared treatment strategies on tuberculosis compared with other conditions might partly explain the discrepancy.

DISCUSSION

We propose a 3-step approach for the comparison of effect estimates from existing randomized trials and observational studies based on routinely collected data: 1) harmonization of the causal question, 2) harmonization of data analysis, and 3) sensitivity analyses to examine the impact of any remaining discrepancies that could not be accounted for in the harmonization process. We applied this general approach to comparison of the effect of immediate versus deferred antiretroviral treatment initiation in HIV-positive individuals with CD4 counts of >500 cells/mm^3. After harmonization, the 5-year risk difference of AIDS or death for deferred versus immediate treatment was 4.2% in START and 2.1% in HIV-CAUSAL. These results reinforce the current recommendations of initiating treatment as early as possible in HIV-positive individuals. Our 3-step approach can be applied to any randomized-observational comparison.

The harmonized risk under deferred treatment was similar in the randomized trial and in the observational study, but the harmonized risk under immediate initiation was higher in the observational study. Four differences that could not be fully harmonized might explain this difference: 1) residual confounding in the observational study (supported by the poor prognosis of individuals who started treatment soon after baseline even after adjusting for the measured prognostic factors); 2) lower adherence to treatment after immediate initiation in the observational study (supported by lower proportion of virological suppression, a proxy for adherence); 3) higher proportion of tuberculosis events in the randomized trial (the benefit of early initiation might be more pronounced for tuberculosis); and 5) overestimation of the beneficial effect of immediate treatment initiation in the randomized trial due to early stopping after an interim analysis indicating a benefit (42, 43).

Whatever the remaining differences, the harmonized estimates from the randomized trial and observational study were in the same neighborhood. In contrast, a previous observational analysis (18), which did not appropriately emulate a target trial (44), yielded an implausible hazard ratio estimate of 0.51 for immediate versus delayed initiation when the outcome was death only (as opposed to AIDS or death, as in our analysis), the median follow-up was less than 24 months (as opposed to 35 months in our analysis), and there was a 6-month grace period (as opposed to 1 month in our analysis).

In contrast with previous comparisons of observational studies based on meta-analysis (4–8) and within study comparison (45), our approach requires the reanalysis of 2 existing studies and will often result in an imperfect harmonization. For example, in our case study, several factors might have also contributed to the observed-randomized difference, in that we could not fully harmonize some eligibility criteria, the definition of outcome, and the clinical setting. Future work can extend our general framework to incorporate quantitative assessments of the impact of imperfect harmonization on the randomized-observational discrepancies (46, 47).

A reanalysis is often required because the primary inferential target of most randomized trials is the intention-to-treat

effect (e.g., the effect of treatment assignment, regardless of adherence to the treatment), which might not be directly transportable to populations outside of the study with different adherence patterns. Therefore, we used a per-protocol approach to compare the randomized and observational estimates, which required a reanalysis of the randomized trial data. In addition, both observational and randomized estimates must be adjusted for potential selection bias due to loss to follow-up, which might also require a reanalysis of the randomized trial data. The validity of per-protocol effect estimates from both the randomized trial and the observational study relies on untestable assumptions of no unmeasured confounding (3).

In summary, comparisons of randomized trials and observational studies must explicitly consider differences in components of the study design and statistical analysis. Our approach provides a structured framework to compare effect estimates from randomized trials and observational studies and to reduce the number of reasons that can explain discrepancies in those effect estimates.

ACKNOWLEDGMENTS

Author affiliations: Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts (Sara Lodi); Institute for Global Health, University College London, United Kingdom (Andrew Phillips, Kholoud Porter, Caroline Sabin); Department of Infectious Diseases, Rigshospitalet, University of Copenhagen, Denmark (Jens Lundgren); Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Roger Logan, Miguel A. Hernán); Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota (Shweta Sharma, Haitao Chu, James D. Neaton); Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina (Stephen R. Cole); Medical Research Council Clinical Trials Unit in University College London, London, United Kingdom (Abdel Babiker); The Kirby Institute, Sidney, Australia (Matthew Law); Division of Infectious Diseases, Department of Medicine, Cooper University Hospital, Cooper Medical School at Rowan University, Camden, New Jersey (Dana Byrne); Department for Adult's Infectious Diseases, Medical University of Warsaw, Warsaw, Poland (Andrzej Horban); Department of Population Health Sciences, University of Bristol, Bristol, United Kingdom (Jonathan A. C. Sterne); Institut Pierre Louis d'Épidémiologie et de Santé Publique, Institut National de la Santé et de la Recherche Médicale, Sorbonne Université, Paris, France (Dominique Costagliola, Sophie Abgrall); Service de Médecine Interne, Hôpital Antoine Béclère, AP-HP, Clamart, France (Sophie Abgrall); Southern Alberta Clinic, Calgary, Canada (John Gill); Department of Medicine, Faculty of Medicine, University of Calgary, Canada (John Gill); Department of Hygiene, Epidemiology and Medical Statistics, Faculty of Medicine, National and Kapodistrian University of Athens, Greece (Giota Touloumi); Programa de Computação Científica, Fundacao Oswaldo Cruz, Rio de Janeiro, Brazil (Antonio G. Pacheco); Stichting HIV Monitoring, Amsterdam, the Netherlands (Ard van Sighem, Peter Reiss); Department of Global Health, Amsterdam University Medical Center, University of Amsterdam, Amsterdam, the Netherlands (Peter Reiss); Division of Infectious Diseases, University of Amsterdam, Amsterdam, the Netherlands (Peter Reiss); Amsterdam Institute for Global Health and Development, Amsterdam, the Netherlands (Peter Reiss); Amsterdam Public Health Research Institute, Amsterdam, the Netherlands (Peter Reiss); Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, University of Basel, Switzerland (Heiner C. Bucher); Centre for Epidemiological Studies on HIV/STI in Catalonia, Agència de Salut Pública de Catalunya, Badalona, Spain (Alexandra Montoliu Giménez); Centro Nacional de Epidemiología, Instituto de Salud Carlos III, Madrid, Spain (Inmaculada Jarrin); Bordeaux Population Health Research Center, Team MORPH3EUS, Unité Mixte de Recherche 1219, CIC-EC 1401, Institut de Santé Publique, d'Epidémiologie et de Développement, Institut National de la Santé et de la Recherche Médicale, University of Bordeaux, Bordeaux, France (Linda Wittkop); Service D'information Médicale, Pôle de Santé Publique, Centre Hospitalier Universitaire de Bordeaux, Bordeaux, France (Laurence Meyer); Unité Mixte de Recherche 1018, le Kremlin Bicêtre, Université Paris Sud, Paris, France (Laurence Meyer); Vall d'Hebrón Research Institute, Barcelona, Spain (Santiago Perez-Hoyos); Department of Internal Medicine, Yale University School of Medicine, New Haven, Connecticut (Amy Justice); Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts (Miguel A. Hernán); and Harvard-MIT Division of Health Sciences and Technology, Boston, Massachusetts (Miguel A. Hernán).

This work received funding from Harvard University Center for AIDS Research (grant 5P30AI060354-13) and the National Institutes of Health (grants AI102634, UL1TR001079, UM1-AI068641 and UM1-AI120197).

The funding organizations had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. The views expressed are those of the author(s) and do not necessarily reflect the official views of the Uniformed Services University of the Health Sciences, the National Institutes of Health, the Department of Defense, or the Departments of the Army, Navy or Air Force.

A complete list of contributors to the HIV-CAUSAL Collaboration and of the INSIGHT START group can be found in Web Appendices 1 and 2.

A.R. has received funding from the Bill & Melinda Gates Foundation; H.R. and his institution have received honoraria, support to attend conferences, or unrestricted research grants from Gilead Sciences, BMS, ViiV Healthcare, Janssen, Abbvie, and MSD in the 3 years preceding the submission date of this manuscript; C.S. received funding from Gilead Sciences, ViiV Healthcare, and Janssen-Cilag for membership in Data Safety and Monitoring Boards, Advisory Boards, and Speaker Panels and for the preparation of educational materials; A.v.S. reports grants from Dutch Ministry of Health, Welfare and

Sport, during the conduct of the study, and grants from European Centre for Disease Prevention and Control, outside the submitted work. G.T. has received grants unrelated to this study from Gilead Sciences Europe, UCL, ECDC, and EU and National funds; D.C. reports grants from Janssen-Cilag, Merck-Sharp & Dohme-Chibret, and ViiV, as well as personal fees from Janssen-Cilagand Merck-Sharp & Dohme-Chibret for lectures, personal fees from ViiV for travel/accommodations/meeting expenses, personal fees from Gilead France for French HIV board, and personal fees from Innavirvax and Merck Switzerland for consultancy, outside the submitted work. P.R. has received independent scientific grant support through his institution from Gilead Sciences, Janssen Pharmaceuticals Inc., Merck & Co., and ViiV Healthcare; he has served on scientific advisory boards for Gilead Sciences, ViiV Healthcare, Merck & Co., Teva Pharmaceutical Industries, and on a data-safety monitoring committee for Janssen Pharmaceuticals Inc., for which his institution has received remuneration. The other authors report no conflicts.

REFERENCES

1. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology*. 2008;19(6):766–779.
2. Hernán MA, Sauer BC, Hernández-Díaz S, et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol*. 2016;79: 70–75.
3. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med*. 2017;377(14):1391–1398.
4. Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *Cochrane Database Syst Rev*. 2014;(4):MR000034.
5. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*. 2000; 342(25):1878–1886.
6. Edwards JP, Kelly EJ, Lin Y, et al. Meta-analytic comparison of randomized and nonrandomized studies of breast cancer surgery. *Can J Surg*. 2012;55(3):155–162.
7. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*. 2000;342(25):1887–1892.
8. Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. *BMJ*. 2016;352:i493.
9. INSIGHT START Study Group. Initiation of Antiretroviral Therapy in early asymptomatic HIV infection. *N Engl J Med*. 2015;373(9):795–807.
10. Lodi S, Phillips A, Logan R, et al. Comparative effectiveness of strategies for antiretroviral treatment initiation in HIV-positive individuals in high-income countries: an observational cohort study of immediate universal treatment versus CD4-based initiation. *Lancet HIV*. 2015;2(8):e335–e343.
11. DHHS Panel on Antiretroviral Guidelines for Adults and Adolescents—A Working Group of the Office of AIDS Research Advisory Council. Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV. 2018. https://aidsinfo.nih.gov/contentfiles/lvguidelines/adultandadolescentgl.pdf. Accessed January 19, 2019.
12. European AIDS Clinical Society. European guidelines for treatment of HIV infected adults in Europe. 2018. http://www.eacsociety.org/guidelines/eacs-guidelines/eacs-guidelines.html. Accessed January 19, 2019.
13. TEMPRANO ANRS Study Group. A trial of early antiretrovirals and isoniazid preventive therapy in Africa. *N Engl J Med*. 2015;373(9):808–822.
14. Severe P, Juste MA, Ambroise A, et al. Early versus standard antiretroviral therapy for HIV-infected adults in Haiti. *N Engl J Med*. 2010;363(3):257–265.
15. Cohen MS, Chen YQ, McCauley M, et al. Prevention of HIV-1 infection with early antiretroviral therapy. *N Engl J Med*. 2011; 365(6):493–505.
16. Anglemyer A, Rutherford GW, Easterbrook PJ, et al. Early initiation of antiretroviral therapy in HIV-infected adults and adolescents: a systematic review. *AIDS*. 2014;28(suppl 2):S105–S118.
17. HIV-CAUSAL Collaboration, Cain LE, Logan R, et al. When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries: an observational study. *Ann Intern Med*. 2011;154(8):509–515.
18. Kitahata MM, Gange SJ, Abraham AG, et al. Effect of early versus deferred antiretroviral therapy for HIV on survival. *N Engl J Med*. 2009;360(18):1815–1826.
19. When To Start Consortium, Sterne JA, May M, et al. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet*. 2009;373(9672):1352–1363.
20. Writing Committee for the CASCADE Collaboration. Timing of HAART initiation and clinical outcomes in human immunodeficiency virus type 1 seroconverters. *Arch Intern Med*. 2011;171(17):1560–1569.
21. Young JG, Cain LE, Robins JM, et al. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Stat Biosci*. 2011;3(1):119–143.
22. Lodi S, Sharma S, Lundgren JD, et al. The per-protocol effect of immediate versus deferred antiretroviral therapy initiation. *AIDS*. 2016;30(17):2659–2663.
23. Caniglia EC, Cain LE, Justice A, et al. Antiretroviral penetration into the CNS and incidence of AIDS-defining neurologic conditions. *Neurology*. 2014;83(2):134–141.
24. HIV-CAUSAL Collaboration. Opportunistic infections and AIDS malignancies early after initiating combination antiretroviral therapy in high-income countries. *AIDS*. 2014; 28(16):2461–2473.
25. HIV-CAUSAL Collaboration, Ray M, Logan R, et al. The effect of combined antiretroviral therapy on the overall mortality of HIV-infected individuals. *AIDS*. 2010;24(1):123–137.
26. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol*. 2016;183(8):758–764.
27. Hernán MA, Robins JM. *Causal Inference*. Forthcoming ed. Boca Raton, FL: Chapman & Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/. Accessed May 23, 2019.
28. Ancelle-Park R. Expanded European AIDS case definition. *Lancet*. 1993;341(8842):441.
29. Hernán MA, Hernandez-Diaz S. Beyond the intention-to-treat in comparative effectiveness research. *Clin Trials*. 2012;9(1): 48–55.

30. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period: application to the healthy worker survivor effect. *Math Model*. 1986;7(9–12):1393–1512.
31. Taubman SL, Robins JM, Mittleman MA, et al. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol*. 2009;38(6):1599–1611.
32. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer; 2001.
33. The World Bank. World Bank Country and Lending Groups. 2015. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups. Accessed May 6, 2019.
34. Harvard Program on Causal Inference: Software. http://www.hsph.harvard.edu/causal/software/. Accessed April 12, 2019.
35. Harvard Program on Causal Inference: sascode_Lodi_AJE19. https://www.hsph.harvard.edu/causal/sascode_lodi_aje19/. Accessed April 12, 2019.
36. Hong JL, Jonsson Funk M, LoCasale R, et al. Generalizing randomized clinical trial results: implementation and challenges related to missing data in the target population. *Am J Epidemiol*. 2018;187(4):817–827.
37. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am J Epidemiol*. 2010;172(1):107–115.
38. Wang Q, Young J, Bernasconi E, et al. Virologic and immunologic responses in treatment-naive patients to ritonavir-boosted atazanavir or efavirenz with a common backbone. *HIV Clin Trials*. 2014;15(3):92–103.
39. Daar ES, Tierney C, Fischl MA, et al. Atazanavir plus ritonavir or efavirenz as part of a 3-drug regimen for initial treatment of HIV-1. *Ann Intern Med*. 2011;154(7):445–456.
40. Lodi S, del Amo J, d'Arminio Monforte A, et al. Risk of tuberculosis following HIV seroconversion in high-income countries. *Thorax*. 2013;68(3):207–213.
41. Sonnenberg P, Glynn JR, Fielding K, et al. How soon after infection with HIV does the risk of tuberculosis start to increase? A retrospective cohort study in South African gold miners. *J Infect Dis*. 2005;191(2):150–158.
42. Guyatt GH, Briel M, Glasziou P, et al. Problems of stopping trials early. *BMJ*. 2012;344:e3863.
43. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*. 2010;303(12):1180–1187.
44. Hernán MA, Robins JM. Early versus deferred antiretroviral therapy for HIV. *N Engl J Med*. 2009;361(8):822–824; author reply 3–4.
45. Wong VC, Steiner PM. Designs of empirical evaluations of nonexperimental methods in field settings. *Eval Rev*. 2018;42(2):176–213.
46. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. *Ann Intern Med*. 2017;167(4):268–274.
47. MacLehose RF, Kaufman S, Kaufman JS, et al. Bounding causal effects under uncontrolled confounding using counterfactuals. *Epidemiology*. 2005;16(4):548–555.