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 AC 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, Miguel A. Hernán ; on behalf the INSIGHT START Study Group and the HIV-CAUSAL Collaboration

Correspondence

Dr Sara Lodi, Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, MA02118 Boston, USA. Email: slodi@bu.edu

Phone: +1 (617) 358 2705

Affiliations:

Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA (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, MA, USA (Roger Logan, Miguel A. Hernán); Division of Biostatistics, School of Public Health,

University of Minnesota, Minneapolis, MN, USA (Shweta Sharma; Haitao Chu, James D. Neaton); Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, NC, USA (Jonathan AC Sterne); 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, NJ, USA (Dana Byrne); Medical University of Warsaw, Department for Adult's Infectious Diseases, Warsaw, Poland (Andrzej Horban); Department of Population Health Sciences, University of Bristol, Bristol, United Kingdom (Jonathan AC Sterne); INSERM, Sorbonne Université, Institut Pierre Louis d'Épidémiologie et de Santé Publique (IPLESP), Paris, France , (Dominique Costagliola, Sophie Abgrall); AP-HP, Hôpital Antoine Béclère, Service de Médecine Interne, Clamart, France (Sophie Abgrall); Southern Alberta Clinic, Calgary, Canada (John Gill);; Department of Medicine, University of Calgary, Canada (John Gill); National and Kapodistrian University of Athens, Faculty of Medicine, Dept. of Hygiene, Epidemiology and Medical Statistics, Greece (Giota Touloumi); Programa de Computação Científica, Fundacao Oswaldo Cruz, Rio de Janeiro, Brasil, (Antonio G. Pacheco); Stichting HIV Monitoring, Amsterdam, the Netherlands (Ard van Sighem, Peter Reiss); Amsterdam University Medical Centres, University of Amsterdam, Department of Global Health and Division of Infectious Diseases; Amsterdam Institute for Global Health and Development, and 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 (CEEISCAT), Agència de Salut Pública de Catalunya (ASPC), Badalona, Spain (Alexandra Montoliu Giménez); Centro Nacional de Epidemiología, Instituto de Salud Carlos III, Madrid, Spain (Inmaculada Jarrin); Univ. Bordeaux, ISPED, Inserm, Bordeaux Population Health Research Center, team MORPH3EUS, UMR 1219, CIC-EC 1401, F-33000 Bordeaux, France (Linda Wittkop); CHU de Bordeaux, Pôle de santé publique, Service

d'information médicale, F-33000 Bordeaux, France.Université Paris Sud, UMR 1018, le Kremlin Bicêtre, France (Laurence Meyer); Vall d'Hebrón Research Institute; Barcelona, Spain (Santiago Perez-Hoyos); Yale University School of Medicine, New Haven, CT, US (Amy Justice); Department of Biostatistics, Harvard T.H. Chan School of Public Health (Miguel A. Hernán); Harvard-MIT Division of Health Sciences and Technology; Boston, MA, USA (Miguel A. Hernán);

Funding

Harvard University CFAR grant 5P30AI060354-13; NIH 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 NIH, the Department of Defense, or the Departments of the Army, Navy or Air Force.

Running head

Comparing randomized and observational studies

Abstract

Effect estimates from randomized trials and observational studies may not be directly comparable because of differences in study design, other than randomization, and in data analysis. We propose a three-step procedure to facilitate meaningful comparisons of effect estimates from randomized trials and observational studies: 1) harmonization of the study protocol (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 to the human immunodeficiency virus from the START randomized trial and the observational HIV-CAUSAL Collaboration.

Key words: causal inference; per-protocol; target trial; antiretroviral initiation;

Introduction

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 largely explained by selection bias (because the 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), may 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 re-analysis of the data of both the randomized trials and the observational studies.

Here, we describe a systematic approach to improve the comparison of effect estimates from a randomized trial and an observational study. Our approach has three stages: 1) harmonization of the study protocol 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 INSIGHT Strategic Timing of Antiretroviral Therapy (START) randomized trial (9) and an observational analysis of routinely collected data in the HIV-CAUSAL Collaboration (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 of initiating ART was guided by the CD4 cell count (low levels indicate severe immunosuppression). During the 2000s, a key question was at which CD4 count should ART 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 two 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-naïve individuals with CD4 count>500 cells/mm^3 at the start of the study.

The START trial included 4685 individuals from low, middle and high-income countries. The intention-to-treat hazard ratio for immediate vs. 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 (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 vs. deferred initiation was 0.66 (95% CI 0.56, 0.75) and the risk difference 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 as the two studies presented several key differences summarized in the outer columns of Web Figure 1. In the next sections, we describe a process to harmonize their study design and data analysis.

Stage 1: Harmonization of study protocol

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. In future references to these 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 two CD4 counts>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 two CD4 count>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 ≥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 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 4676 eligible individuals in the actual trial and the 14,595 in the emulated trial after harmonization. Participants in the two 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 2769 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 may take several weeks before treatment is started due to clinical tests and administrative procedures, we defined the grace period to be 1 month. We then defined the two 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 two treatment

strategies, which leads to the expectation of no unmeasured confounding at baseline, i.e., that the two groups are exchangeable at baseline (though not necessarily at later follow-up times) (27). In HIV-CAUSAL, individuals are 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, and 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 months (interquartile range [IQR] 26, 47) in the actual trial and 32 months (IQR 16, 58) 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, serious non-AIDS events and death in START and the earlier of death or any AIDS diagnosis in HIV-CAUSAL. Since 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 [IQR] CD4 counts at which events occurred were 573 cells/mm^3 [444,711] in the actual trial and 560 cells/mm^3 [426,700] 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 post-baseline prognostic factors that predict treatment initiation and loss to follow-up. Because conventional methods cannot appropriately handle post-baseline prognostic factors that affect treatment status and are also affected by past treatment (i.e, treatment-confounder feedback), g-methods like 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 post-baseline values of CD4 count, HIV-RNA, timing of CD4 count and HIV-RNA measurement were sufficient to adjust for post-baseline 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 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 two steps (10, 21, 31). First, we fit separate regression models for each of the post-baseline 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 are used to simulate the outcome risk.

For the first step, in both the actual and emulated trials, we fit 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, and the following baseline variables: CD4 count, HIV-RNA, age, sex, mode of HIV acquisition, calendar year. In addition, models for the trial data were adjusted for income status of country (high versus middle and low income, according to the World Bank definition (33)) and models for the observational data were 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 two 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 Figure 2-4). All analyses were conducted using the publicly available the GFORMULA_RCT and GFORMULA SAS macros (34, 35)

Table 2 shows the estimated per-protocol effects of immediate vs. deferred treatment initiation in the harmonized studies. The estimated 5-year risk of AIDS or death under deferred treatment initiation was 6.0% (95% confidence interval [CI] 4.4,8.1) in the actual trial and 5.1% (4.4,5.7) in the emulated trial. The corresponding estimated risk under immediate treatment initiation was higher in the emulated trial (3.0%; (2.3,3.7)) than in the actual trial (1.8%; (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% (1.1,3.1) percentage points versus a reduction of 4.2% (2.5,6.3) percentage points. The estimated hazard ratio for deferred versus immediate treatment was 3.4 (2.1,6.2) in the actual trial and 1.63 (1.28,2.32) in the emulated trial. The proportions of individuals who had initiated treatment under the deferred treatment initiation were comparable in the two studies (Web Figure 5). Because of the definition of the intervention, corresponding proportions under immediate treatment initiation was 100% in both studies one 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 are conducted to explore the impact on the non-harmonized components on the effect estimates.

For our case study, Table 3 lists components of the protocol that we could not fully harmonize. We now describe the corresponding sensitivity analyses.

Eligibility criteria

The two 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 (the same 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 if the joint distribution of measured covariates would have 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 (95% CI) of AIDS or death in the actual trial was 3.6% (2.7, 4.8) under the immediate treatment strategy and 4.8% (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 may 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<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 two 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 may have differed between the two studies. The proportions of individuals who initiated ART with a protease inhibitor regime and with a non-nucleoside 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%). Since non-nucleoside 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% (0,1.3) lower under immediate initiation (and the 2-year risk 1.7% (1.0,2.7) lower). In the emulated trial, no benefit was estimated in the 1-year risk (and only 0.4% (0,0.9) in 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 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 reran analyses restricted to the four cohorts with lowest follow-up rate (Swiss HIV Cohort Study, CoRIS, PISCIS and French Hospital Database): the risk of AIDS or death was 3.0% (2.1,3.9) for immediate initiation and 4.6% (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 may be the result of including some individuals not fully engaged in HIV care, we conducted analyses excluding individuals who were lost early. The risk of AIDS or death were 3.1% (2.3,3.8) for immediate initiation and 5.2% (4.5,5.7) for deferred initiation after excluding individuals who were lost by 12 months, and 3.3% (2.5,4.1) and 5.5% (4.8,6.0) after excluding individuals who were lost by 12 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 tuberculosis is more common. In the harmonized analyses 28% of outcome events in the actual trial were tuberculosis, but 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 count (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 proposed 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 vs. deferred antiretroviral treatment initiation in HIV-positive individuals with CD4>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 which could not be fully harmonized might explain this difference: (i) 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), (ii) lower adherence to treatment after immediate initiation in the observational study (supported by lower proportion of virological suppression, a proxy for adherence), (iii) higher proportion of tuberculosis events in the randomized trial (the benefit of early initiation might be more pronounced for

tuberculosis), and (v) 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 vs delayed initiation when the outcome was death only (as opposed to AIDS or death in our analysis), the median follow-up was less than 24 months (as opposed to 35 months in our analysis), and 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 two two existing studies and will often result in an imperfect harmonization. For example, in our case study several factors may have also contributed to the observed-randomized difference, as 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 may 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 re-analysis of the randomized trial data. In addition, both observational and randomized estimates need to be adjusted for potential selection bias due to loss to

follow-up, which may also require a re-analysis 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 need to 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

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

Conflict of interest statement: A Phillips has received funding the Bill & Melinda Gates Foundation; HC Bucher and his institution has received honorarium, support to attend conferences or unrestricted research grants from Gilead Sciences, BMS, Viiv Healthcare, Janssen, Abbvie, MSD in the last 3 years preceding the submission date of this manuscript; C Sabin received funding from Gilead Sciences, ViiV Healthcare and Janssen-Cilag for the membership of Data Safety and Monitoring Boards, Advisory Boards, Speaker Panels and for the preparation of educational materials; ; A van Sighem reports grants from Dutch Ministry of Health, Welfare and Sport, during the conduct of the study; grants from European Centre for Disease Prevention and Control, outside the submitted work. G Touloumi has received grants unrelated to this study from Gilead Sciences Europe, UCL, ECDC and EU and National funds; Dr. Costagliola reports grants from Janssen-Cilag (2017-2018), Merck-Sharp & Dohme-Chibret (2015-2017), ViiV (2015), personal fees from Janssen-Cilag (2016) and Merck-Sharp & Dohme-Chibret (2015, 2017) for lectures, personal fees from ViiV (2015) for travel/accomodations/meeting expenses, personal fees from Gilead France from 2011 until december 2015 for French HIV board, personal fees from Innavirvax (2015 and 2016) and Merck Switzerland (2017) for consultancy, outside the submitted work. P Reiss through his institution has received independent scientific grant support 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. No other conflict of interest to report.

List of abbreviations

HIV: human immunodeficiency virus

AIDS: acquired immune deficiency syndrome

ART: antiretroviral treatment

CI: confidence interval

IQR: interquartile range

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):766-79.
2. Hernan MA, Sauer BC, Hernandez-Diaz S, et al. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* 2016;79:70-5.
3. Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med* 2017;377:1391-8.
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):1878-86.
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):155-62.
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):1887-92.
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–e43.
11. DHHS panel on antiretroviral guidelines for adults and adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. 2018. (https://aidsinfo.nih.gov/contentfiles/lvguidelines/adultandadolescentgl.pdf). (Accessed).
12. European AIDS clinical society (EACS). European guidelines for treatment of HIV infected adults in Europe. 2018. (http://www.eacsociety.org/guidelines/eacs-guidelines/eacs-guidelines.html). (Accessed).
13. Temprano Anrs Study Group. A Trial of Early Antiretrovirals and Isoniazid Preventive Therapy in Africa. *N Engl J Med* 2015;373(9):808-22.
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):257-65.
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:S105-18.
17. Cain LE, Logan R, Robins JM, 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):509-15.
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):1815-26.
19. Sterne JA, May M, Costagliola D, et al. Timing of initiation of antiretroviral therapy in AIDS-free HIV-1-infected patients: a collaborative analysis of 18 HIV cohort studies. *Lancet* 2009;373(9672):1352-63.

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):1560-9.
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):119-43.
22. Lodi S, Sharma S, Lundgren JD, et al. The per-protocol effect of immediate versus deferred antiretroviral therapy initiation. *AIDS* 2016;30(17):2659-63.
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):134-41.
24. collaboration HC. Opportunistic infections and AIDS malignancies early after initiating combination antiretroviral therapy in high-income countries. *AIDS* 2014;28(16):2461-73.
25. Collaboration H-C, Ray M, Logan R, et al. The effect of combined antiretroviral therapy on the overall mortality of HIV-infected individuals. *Aids* 2010;24(1):123-37.
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):758-64.
27. Hernán MA, Robins JM. *Causal Inference*. Forthcoming ed. Boca Raton,FL: Chapman & Hall/CRC; 2018.
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. *Mathematical Modelling* 1986;7(9-12):1393-512.
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):1599-611.
32. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York NY: Springer; 2001.
33. Bank W. World Bank Country and Lending Groups. 2015.
34. Harvard Program in 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* 2017;187(4):817-22.
37. Cole SR, Stuart EA. Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *Am J Epidemiol* 2010;172(1):107-15.
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):445-56.
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):207-13.
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):150-8.
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):1180-7.
44. Hernán MA, Robins JM. Early versus deferred antiretroviral therapy for HIV. *N Engl J Med* 2009;361(8):822-3; 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):268-74.
47. MacLehose RF, Kaufman S, Kaufman JS, et al. Bounding causal effects under uncontrolled confounding using counterfactuals. *Epidemiology* 2005;16(4):548-55.

Table 1. Participants' characteristics at baseline after harmonization in START (actual trial) and the HIV-CAUSAL Collaboration (emulated trial)

Characteristics at baseline	**Actual trial (4676)**			**Emulated trial (14,595)**
	Median (interquartile range)	**N**	**%**	**Median [interquartile range]**
CD4 cell count, median [IQR] cells/mm^3 [a]	651 (584,765)			559 (585,779)
Enrollment year, median [IQR]	2012 (2011,2013)			2009 (2007,2011)
Age, median [IQR]	36 (29,44)			36 (29,43)
HIV-RNA, median copies/mL[IQR]	12759 (3019,43391)			17469 (4300,57539)
Females		1253	27%	
HIV acquisition risk group				
MSM		1787	38%	
MSW or WSM		25814	55%	
IDU		64	1%	
Other/Unknown		244	5%	
High-income setting[b]		2769	59%	

MSM: men who have sex with men; MSW: men who have sex with women; WSM: women who have sex with men; IDU: injecting drug use

a. Average of two baseline values

b. Based on World Bank classification

Table 2. Per-protocol effect estimates of the 5-year risk, risk difference and hazard ratios of AIDS or death in START (actual trial) and in HIV-CAUSAL Collaboration (emulated trial) after harmonization.

Treatment strategy	Risk (%)	95% CI	Risk Difference (%) (Deferred-Immediate)	95% CI	Hazard Ratio (Deferred/Immediate)	95% C
Actual trial [a]						
Immediate treatment	1.8	1.1,2.6	Ref	Ref	Ref	Ref
Deferred treatment	6	4.4,8.1	4.2	2.5,6.3	3.4	2.1,6.
Emulated trial [b]						
Immediate treatment	3	2.3,3.7	Ref	Ref	Ref	Ref
Deferred treatment	5.1	4.4,5.7	2.1	1.1,3.1	1.6	1.3,2.

a. N=4,676; 112 events
b. N=14,595; 422 events

Table 3. Possible explanations for differences in estimates from the randomized START and the HIV-CAUSAL Collaboration 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 two 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 n 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	Re-analysis 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 larger risk of tuberculosis than in HIV-CAUSAL	Re-analysis excluding cases of tuberculosis as events (if possible)

ART: antiretroviral therapy;