# How to estimate the effect of treatment duration on survival outcomes using observational data

Miguel A Hernán

Departments of Epidemiology and Biostatistics, Harvard T H Chan School of Public Health, Harvard-MIT Division of Health Sciences of Technology, Boston, Massachusetts, MA 02115, USA

Correspondence to: M A Hernán miguel_hernan@post.harvard.edu

When using observational data, quantifying the effect of treatment duration on survival outcomes is not straightforward because only people who live for a long time can receive treatment for a long time. This problem doesn't apply to randomised trials because people are classified based on the treatment duration they are assigned, rather than the treatment duration that they achieve. This approach accepts that dead people do not deviate from their assigned treatment strategy. By transferring this insight to the analysis of observational data, we can follow three steps to estimate the effect of treatment duration from observational data without the bias of naive comparisons between long term and short term users. The first step is cloning people to assign them to multiple treatment strategies. The second step is censoring clones when they deviate from their assigned treatment strategy. The third step is performing inverse probability weighting to adjust for the potential selection bias introduced by censoring. The procedure can be used to compare any treatment strategies that are sustained over time. Cloning, censoring, and weighting eliminates immortal time bias in the estimates of absolute and relative risk, which helps researchers focus their attention on other biases that may be present in observational analyses and are not so easily eliminated.

## Introduction

Quantifying the effect of treatment duration on survival outcomes is not straightforward because only people who survive for a long time can receive a treatment for a long time. Suppose we want to estimate the effect of statins on the mortality of patients with cancer using a healthcare database.[1] A direct comparison of long term users, short term users, and non-users would be biased because long term users have, by definition, survived for a long time. Several methods can be used to tackle this bias, but some do not enable estimation of absolute risks or appropriate adjustment for time varying confounders. To overcome these limitations, I first review an uncontroversial approach to estimating the effect of treatment duration in randomised trials and then explain how to emulate this approach in observational data analyses.

## Estimating the effect of treatment duration in a randomised trial with full adherence

Let us consider a simple example that encapsulates some key features of the problem. Table 1 shows data from a trial with perfect adherence and no loss to follow-up, in which 12 people are randomly assigned to one of three treatment strategies: no aspirin (durA=0), one year of aspirin (durA=1), or two years of aspirin (durA=2).

Treatment duration did not affect survival at any time in this trial. Under each of the strategies, 25% of people had died by the end of the first year and 75% by the end of the second year. The two year risk ratio for durA=2 compared with durA=0 is 0.75/0.75=1. To avoid statistical considerations, which are not central to our discussion about systematic bias, we will view each person in table 1 as representing a million people with the same data, so that the 95% confidence interval around this null estimate is very narrow.

Suppose that we naively compare the probability of death between people who actually took aspirin for two years (two thirds because there are two deaths among three people (9, 10, 11) after excluding the

## SUMMARY POINTS

- Estimating the absolute and relative risks of treatment duration on survival outcomes requires care because only people who survive a long time can be treated for a long time
- Data from randomised controlled trials with full adherence enable simple analysis of these risks, but data from observational studies do not
- A three step procedure (cloning, censoring, and weighting) that emulates the analysis of randomised trials with full adherence can be used to estimate the effect of treatment duration and of any other treatment strategies that are sustained over time
- Other approaches based on allocating person time and pooling hazard ratios do not enable estimation of absolute risks or appropriately adjust for time varying confounders

Table 1 | Hypothetical randomised trial in which 12 people are randomly assigned to one of three treatment durations of aspirin: no aspirin (durA=0), one year (durA=1), two years (durA=2).

| Person | durA | Aspirin at start of first year | Dead at end of first year | Aspirin at start of second year | Dead at end of second year |
|---|---|---|---|---|---|
| 1 | 0 | No | No | No | No |
| 2 | 0 | No | No | No | Yes |
| 3 | 0 | No | No | No | Yes |
| 4 | 0 | No | Yes | - | Yes |
| 5 | 1 | Yes | No | No | No |
| 6 | 1 | Yes | No | No | Yes |
| 7 | 1 | Yes | No | No | Yes |
| 8 | 1 | Yes | Yes | - | Yes |
| 9 | 2 | Yes | No | Yes | No |
| 10 | 2 | Yes | No | Yes | Yes |
| 11 | 2 | Yes | No | Yes | Yes |
| 12 | 2 | Yes | Yes | - | Yes |

patient who was assigned to two years but died in the first year) and those who did not take aspirin (three quarters because there are three deaths among four people (1, 2, 3, 4)). The ratio (2/3)/(3/4) is <1, even though we know that treatment had no effect. This is not surprising: the average survival is longer in people who received two years of treatment because they were alive for at least two years. By definition, people receiving treatment were "immortal" during those two years, which is why the bias of this naive analysis is referred to as immortal time bias.[2]

This analysis fails to acknowledge a simple fact about the people assigned to two years of treatment who died during the first year: they did not deviate from their assigned treatment strategy, they just happened to die while following their assigned strategy. In a misguided attempt to correct for non-existing non-adherence,[3] the naive analysis introduces bias. By contrast, the valid analysis accepts that dead people necessarily stop receiving treatment, regardless of the treatment duration they were assigned to. We now need to transfer these insights to the analysis of observational data.

### Emulating a randomised trial with full adherence using observational data

Suppose we want to estimate the effect of treatment duration using a healthcare database with 12 people. Table 1 shows the data, but we exclude column durA because observational datasets don't show the treatment duration, if any, that patients were assigned at time zero.[4] For simplicity, we still view each person as representing a million and assume no confounding—people who do and do not receive aspirin at each time have similar prognostic factors.

For the data from a randomised trial, the two year mortality risk ratio for two years of aspirin compared with no aspirin was 1. So it should also be 1 in an unconfounded observational study. But the lack of the variable durA in the observational dataset precludes us from performing the valid analysis we used for the randomised trial. The lack of this variable makes observational analyses susceptible to naive comparisons, such as comparing people who received two years of treatment with those who received no treatment. This comparison was biased in the

randomised trial and will be biased in an observational analysis.

We can emulate the valid analysis of the randomised trial using the observational data in three steps—cloning, censoring, and weighting.

### Cloning: assign people to a treatment strategy at time zero

The solution to the problem created by the lack of the durA variable is surprisingly simple: create it. Person 1 in table 1 did not receive treatment at time zero, so can be assigned to the strategy durA=0. Person 5 did receive treatment at time zero, so could be assigned to either durA=1 or the durA=2. Randomly assigning the person to one of these strategies would be statistically inefficient. Rather, we assign person 5 to both durA=1 and durA=2. Note that looking at the strategy a person ended up following is not a valid way to assign people to strategies at time zero—it will introduce immortal time bias.

We assign each person to all treatment strategies that are compatible with their observed data at time zero. Assigning a person to two strategies simultaneously is equivalent to having two copies (or clones) of the person in the dataset, with each copy assigned to a different strategy.[5] In our example, we create a dataset with two clones of each person who received treatment at time zero. We assign one clone to durA=1 and the other to durA=2. Table 2 shows the expanded population with eight clones in each of these two groups.

### Censoring: ensure that people follow their assigned strategy after time zero

If clones deviate from their assigned strategy during follow-up, we artificially censor them. At one year, clones assigned to durA=1 will be censored if they receive treatment at that time, and clones assigned to durA=2 will be censored if they do not. For our example, three clones in each durA=1 and durA=2 are censored because they deviated from their assigned strategy (table 2).

But comparing the two year risk of death for durA=2 with that for durA=0 among uncensored people is still biased. To see why, look at the five uncensored clones in durA=2 (8b, 9b, 10b, 11b, and 12b). We know from table 1 that the two year risk of death in patients who receive two years of aspirin should be 0.75, but the

| Person/clone | durA | Aspirin at start of first year | Dead at end of first year | Aspirin at start of second year | Dead at end of second year | Inverse probability weight |
|---|---|---|---|---|---|---|
| 1 | 0 | No | No | No | No | 1 |
| 2 | 0 | No | No | No | Yes | 1 |
| 3 | 0 | No | No | No | Yes | 1 |
| 4 | 0 | No | Yes | - | Yes | 1 |
| 5a | 1 | Yes | No | No | No | 2 |
| 6a | 1 | Yes | No | No | Yes | 2 |
| 7a | 1 | Yes | No | No | Yes | 2 |
| 8a | 1 | Yes | Yes | - | Yes | 1 |
| 9a | 1 | Yes | No | Yes | Censored | 0 |
| 10a | 1 | Yes | No | Yes | Censored | 0 |
| 11a | 1 | Yes | No | Yes | Censored | 0 |
| 12a | 1 | Yes | Yes | - | Yes | 1 |
| 5b | 2 | Yes | No | No | Censored | 0 |
| 6b | 2 | Yes | No | No | Censored | 0 |
| 7b | 2 | Yes | No | No | Censored | 0 |
| 8b | 2 | Yes | Yes | - | Yes | 1 |
| 9b | 2 | Yes | No | Yes | No | 2 |
| 10b | 2 | Yes | No | Yes | Yes | 2 |
| 11b | 2 | Yes | No | Yes | Yes | 2 |
| 12b | 2 | Yes | Yes | - | Yes | 1 |

Table 2 | Data from a hypothetical observational study of 12 people

risk in the uncensored is actually 4/5=0.80. The ratio 0.80/0.75 does not equal the true risk ratio of 1. Even though cloning has eliminated the immortal time bias, artificial censoring has introduced selection bias.[6]

### Weighting: adjust for selection bias

To eliminate the selection bias due to artificial censoring, we can use inverse probability weighting.[7] [8] Informally, uncensored individuals receive a weight equal to the inverse of their probability of being uncensored. In other words, people who are censored transfer their weight in the analysis to those who are uncensored. The goal is to construct a hypothetical population in which nobody is censored because everybody followed their assigned strategy.

Clones assigned to the strategy durA=0 are never artificially censored: their probability of being uncensored is 1 and their inverse probability weight is 1/1. Clones assigned to durA=2 are not artificially censored if they died during the first year or survived the first year and received treatment at the start of the second year, because in both cases they adhered to their strategy. That is, the probability of being uncensored is 1 for those who died during the first year (clones 8b, 12b), and 3/6=0.5 for the others (clones 5b, 6b, 7b, 9b, 10b, 11b). For the five uncensored clones assigned to durA=2, the inverse probability weight is 1 if they died during the first year and 1/0.5=2 if they survived the first year.

We can now proceed to carry out the same valid analysis as in the randomised trial, with all nine uncensored clones (four in durA=0 and five in durA=2) weighted by their respective inverse probability weight. The weighted two year mortality risk ratio for durA=2 compared with durA=0 is (6/8)/(3/4)=1. No bias. We are done.

### Conclusions

The three steps described here—cloning, censoring, and weighting—can be used to estimate the effect of treatment duration on survival outcomes when using observational data (fig 1). Cloning is used to assign people to treatment duration strategies at time zero, eliminating immortal time bias.[9] [10] Artificial censoring ensures that the clones follow their assigned strategy through follow-up. It introduces selection bias, which can be eliminated with inverse probability weighting.

Table 1 shows a simple example—two time points, a null causal effect of treatment, no confounding, and no losses to follow-up—to show the immortal time bias arising from a naive observational analysis and how the three step procedure prevents this bias and yields absolute risks. The procedure can be extended to situations with multiple time points and confounding in which people may start and stop treatment and be lost to follow-up. When confounding and other biases (such as selection bias due to losses from follow-up) exist, additional adjustment using inverse probability weighting is required, as has been described in multiple applications in clinical research.[11-13] Validity of the observational estimates relies on the assumption that all time fixed and time varying confounders are correctly measured and adjusted for.

As well as preventing immortal time bias, the three step procedure can estimate absolute risks and can incorporate appropriate adjustment for time varying confounders.[6] [7] Neither of these can be achieved with

| Step | Goal | Method |
|---|---|---|
| 1 | Assign each individual to every strategy consistent with her data at time zero | Cloning |
| 2 | End follow-up when an individual's data stop being consistent with her assigned strategy | Censoring |
| 3 | Adjust for selection bias introduced by the previous step | IP weighting |

Fig 1 | Three step procedure to compare sustained treatment strategies using observational data. IP=inverse probability

**Box 1: Applications of the three step procedure for comparing sustained treatment strategies**

**Antiretroviral therapy initiation in patients with HIV[16]**

The method was used to compare clinical strategies for starting antiretroviral therapy when CD4 cell count first fell below a threshold ranging between 200 and 500 cells/µL. Delaying initiation (low CD4 thresholds) was estimated to increase the risk of AIDS or death. A similar approach was later used to compare several antiretroviral switching strategies.[17]

**Epoetin dosing in people with end stage renal disease[18]**

The method was used to compare two sustained strategies for intravenous epoetin-α administration over time. One to achieve and maintain hematocrit values between 34.5% and 39.0%, and the other to values between 30.0% and 34.5%. No meaningful differences in survival or cardiovascular risk at six months were found.

**Timing of first line treatment in men with advanced prostate cancer[19]**

The method was used to compare immediate versus deferred initiation of androgen deprivation therapy in men with rising prostate specific antigen as the only sign of relapse of prostate cancer. The 10 year survival was similar under both strategies.

other methods of eliminating immortal time bias that are based on reallocating person time and estimating a weighted average of the time varying hazard ratios.[2] The procedure described here can also be used to estimate the effect of treatment duration in randomised trials with incomplete adherence[3]—cloning may not be necessary because we know the strategy to which each person was assigned.

More generally, the three step procedure can be used to compare any treatment strategies that are sustained over time,[14] of which treatment duration strategies (in which we explicitly specify the duration of treatment) are a simple class. In clinical practice, we often consider sustained strategies in which treatment decisions at each time depend on the patient's evolving clinical history; for example, "increase the dose of epoetin by 10% if haemoglobin drops below 10 g/dL." The three step procedure has been used in these more complex settings (box 1). The underlying principle is that an observational analysis needs to explicitly emulate a (hypothetical) target trial in which eligible people are assigned to different strategies at time zero.[15]

An alternative to cloning, censoring, and weighting that eliminates immortal time bias, estimates absolute risks, and adequately handles treatment confounder feedback is Robins's g formula.[4] Unlike the g formula, however, the three step method can be easily implemented using standard statistical software, even for longitudinal data. The data management required for cloning and censoring can be accomplished with a few lines of code, and inverse probability weighting is typically based on the probabilities predicted by a standard logistic regression model. By contrast, applying the g formula with time varying confounders requires some programming and the fitting of multiple models.

In summary, cloning, censoring, and weighting eliminates immortal time bias in the estimates of absolute and relative risk, which helps researchers focus their attention on other biases that may be present in observational analyses and are not so easily eliminated.

1   Emilsson L, García-Albéniz X, Logan RW, Caniglia EC, Kalager M, Hernán MA. Examining bias in studies of statin treatment and survival in patients with cancer. *JAMA Oncol* 2017; (forthcoming).

2   Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ* 2010;340:b5087. doi:10.1136/bmj.b5087

3   Hernán MA, Robins JM. Per-protocol analyses of pragmatic trials. *N Engl J Med* 2017;377:1391-8. doi:10.1056/NEJMsm1605385

4   Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period — Application to the healthy worker survivor effect[published errata appear in Mathl Modelling 1987;14:917-21]. *Math Model* 1986;7:1393-512doi:10.1016/0270-0255(86)90088-6.

5   Orellana L, Rotnitzky A, Robins JM. Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: main content. *Int J Biostat* 2010;6:8.

6   Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004;15:615-25. doi:10.1097/01.ede.0000135174.63482.43

7   Hernán MA, Robins JM. *Causal Inference.* Chapman & Hall/CRC, 2018, (forthcoming).

8   Mansournia MA, Altman DG. Inverse probability weighting. *BMJ* 2016;352:i189. doi:10.1136/bmj.i189

9   Hernán MA. Counterpoint: epidemiology to guide decision-making: moving away from practice-free research. *Am J Epidemiol* 2015;182:834-9. doi:10.1093/aje/kwv215

10  García-Albéniz X, Hsu J, Hernán MA. The value of explicitly emulating a target trial when using real world evidence: an application to colorectal cancer screening. *Eur J Epidemiol* 2017;32:495-500. doi:10.1007/s10654-017-0287-2

11  Cole SR, Hernán MA, Anastos K, Jamieson BD, Robins JM. Determining the effect of highly active antiretroviral therapy on changes in human immunodeficiency virus type 1 RNA viral load using a marginal structural left-censored mean model. *Am J Epidemiol* 2007;166:219-27. doi:10.1093/aje/kwm047

12  Sterne JAC, Hernán MA, Ledergerber B, et al, Swiss HIV Cohort Study. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *Lancet* 2005;366:378-84. doi:10.1016/S0140-6736(05)67022-5

13  Choi HK, Hernán MA, Seeger JD, Robins JM, Wolfe F. Methotrexate and mortality in patients with rheumatoid arthritis: a prospective study. *Lancet* 2002;359:1173-7. doi:10.1016/S0140-6736(02)08213-2

14  Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J Clin Epidemiol* 2016;79:70-5. doi:10.1016/j.jclinepi.2016.04.014

15  Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016;183:758-64. doi:10.1093/aje/kwv254

16  Cain LE, Logan R, Robins JM, et al, HIV-CAUSAL Collaboration. 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;154:509-15. doi:10.7326/0003-4819-154-8-201104190-00001

17  Cain LE, Saag MS, Petersen M, et al, Antiretroviral Therapy Cohort Collaboration, the Centers for AIDS Research Network of Integrated Clinical Systems, and the HIV-CAUSAL Collaboration. Using observational data to emulate a randomized trial of dynamic treatment-switching strategies: an application to antiretroviral therapy. *Int J Epidemiol* 2016;45:2038-49.

18  Zhang Y, Thamer M, Kaufman J, Cotter D, Hernán MA. Comparative effectiveness of two anemia management strategies for complex elderly dialysis patients. *Med Care* 2014;52(Suppl 3):S132-9. doi:10.1097/MLR.0b013e3182a53ca8

19  Garcia-Albeniz X, Chan JM, Paciorek A, et al. Immediate versus deferred initiation of androgen deprivation therapy in prostate cancer patients with PSA-only relapse. An observational follow-up study. *Eur J Cancer* 2015;51:817-24. doi:10.1016/j.ejca.2015.03.003