Statistics in Medicine

A tutorial on dealing with time-varying eligibility for treatment: Comparing the risk of major bleeding with DOACs versus warfarin

Journal:	*Statistics in Medicine*
Manuscript ID	SIM-20-0205.R1
Wiley - Manuscript type:	Tutorial in Biostatistics
Date Submitted by the Author:	05-Jun-2020
Complete List of Authors:	Schnitzer, Mireille; Universite de Montreal, Faculté de pharmacie Platt, Robert; McGill University, Department of Epidemiology and Biostatistics Durand, Madeleine; Centre Hospitalier de L'Universite de Montreal, Department of Medicine
Keywords:	causal inference, contraindication, inverse probability weighting, longitudinal analysis, positivity assumption

TITLE: A tutorial on dealing with time-varying eligibility for treatment: Comparing the risk of major bleeding with DOACs versus warfarin

AUTHORS: Mireille E. Schnitzer[1], Robert W. Platt[2], and Madeleine Durand[3]

AFFILIATIONS

[1]Faculty of Pharmacy and Department of Social and Preventative Medicine, Université de Montréal, Montreal, Canada
[2]Departments of Pediatrics and of Epidemiology, Biostatistics, and Occupational Health, McGill University, and the Research Institute of the McGill University Health Centre, Montreal, Canada
[3]Internal Medicine Service, Department of Medicine, Centre Hospitalier de l'Université de Montréal, and the Research Center of Centre Hospitalier de l'Université de Montréal, Montréal, Canada

ABSTRACT

In this tutorial, we focus on the problem of how to define and estimate treatment effects when some patients develop a contraindication and are thus ineligible to receive a treatment of interest during follow-up. We first describe the concept of positivity, which is the requirement that all subjects in an analysis be eligible for all treatments of interest conditional on their baseline covariates, and the extension of this concept in the longitudinal treatment setting. We demonstrate using simulated datasets and regression analysis that under violations of longitudinal positivity, typical associational estimates between treatment over time and the outcome of interest may be misleading depending on the data-generating structure. Finally, we explain how one may define "treatment strategies," such as "treat with medication unless contraindicated," to overcome the problems linked to time-varying eligibility. Finally, we show how contrasts between the expected potential outcomes under these strategies may be consistently estimated with inverse probability weighting methods. We provide R code for all the analyses described.

Keywords: anticoagulants, causal inference, contraindication, inverse probability weighting, longitudinal analysis, longitudinal exposure, positivity assumption.

1. INTRODUCTION

When evaluating the relative safety and effectiveness of different medications, understanding a causal inference approach based on counterfactual outcomes may avoid pitfalls resulting in misleading effect estimates. In this tutorial, we focus on the problem of how to define and estimate treatment effects when some patients develop a contraindication and are thus ineligible to receive a treatment of interest during follow-up. [1-4] Under these circumstances, the construction of inverse probability weights for the estimation of the effect of sustained treatment becomes impossible as the probability of receiving treatment will be zero for contraindicated patients. While tutorials exist for estimating the parameters of marginal structural models using IPW [5,6] we focus on the issues that arise when eligibility for treatment is time-varying and standard implementations no longer apply. Guidance on this topic is currently lacking in the literature, both for observational studies and randomized controlled trials (RCTs), [7] even though it is frequently seen in clinical settings.

We motivate the problem through a hypothetical cohort analysis that compares the safety of Direct-Acting Oral Anticoagulants (DOACs) to that of warfarin. We then cover the topics of the positivity assumption often required to define an effect of interest, *treatment strategies* [3,8] (also known as dynamic treatment regimes [9]), and inverse probability of treatment weighting (IPW), [10] and present R code to both numerically illustrate the problem and demonstrate one possible solution. Specifically, we show using simulated data that several naive regression approaches may result in biased estimates while a modified IPW approach based on treatment strategies does not. We conclude that in such complex settings, an investigator may choose to contrast mean outcomes under different treatment strategies rather than different fixed exposures.

2. COMPARING THE SAFETY AND EFFECTIVENESS OF DOACS VERSUS WARFARIN

Atrial fibrillation is the most frequent arrhythmia. It is characterized by an irregular heart rate, and increased risk of cardio-embolic stroke due to embolization of blood clots forming in the heart cavities. To decrease the risk of stroke, most patients with atrial fibrillation have an indication for lifelong anticoagulation. Currently, both warfarin (vitamin K antagonist) and DOACs are recommended for anticoagulation in non-valvular atrial fibrillation. [11] While the safety and effectiveness of warfarin are well understood, the level of anticoagulation achieved for patients on warfarin requires close monitoring of International Normalized Ratio (INR) and frequent dose adjustments due to numerous drug interactions. [12] Frequent blood monitoring imposes a burden on the patient and healthcare system. [13] DOACs, which have been available in Canada since 2010, require no such routine monitoring and are given as fixed doses based on patient

characteristics. Multiple studies, including phase 3 RCTs [14-17] and observational studies, [18] have thus contrasted the effectiveness (stroke and mortality prevention) and safety (risk of bleeding) of DOACs versus the mainstay warfarin for patients with atrial fibrillation. Notably, severe renal failure (less than 15-30 ml/min creatinine clearance, depending on the type of DOAC) is one of several contraindications to DOACs (which also include mechanical heart valve or valvular disease) but is not a contraindication for warfarin. [19,20] Since DOACs entered the market, treatment switches between warfarin and DOACs have been frequent. Typically, warfarin-treated patients will be switched to DOACs for treatment simplification, while DOAC-treated patients will be switched to warfarin if they develop contraindications to DOACs.

As a motivating example for this tutorial, we consider an analysis where the goal is to evaluate the relative risks of major bleeding using a prospective longitudinal cohort study. One complication in such a cohort is the development of renal failure post-baseline. Upon developing such a condition, a patient would no longer be eligible to take a DOAC but would still be eligible to take warfarin. Renal failure is a well-known risk factor for cardiovascular disease, bleeding and death. [21,22] Therefore, as renal failure develops, higher risk individuals are censored from the DOAC cohort and move into the warfarin cohort (as they remain eligible for warfarin treatment).

In the following sections, we explain and demonstrate why standard regression methods of dealing with such a situation may result in incorrect conclusions and we present a method that allows for clinically interpretable results.

3. POSITIVITY IN OBSERVATIONAL STUDIES

Suppose that we are interested in comparing the effect of sustained exposure to DOACs versus sustained exposure to warfarin. In the context of observational studies, where treatments are not assigned by the study design, the assumptions of positivity, no unmeasured confounding, [10] the absence of interference, [23] and exposures corresponding to well-defined interventions [24] are required to equate a statistical association to a “causal” relation. If medication usage is the exposure of interest, positivity means that given the values of their confounding variables, every subject in the analysis was eligible to receive any of the medications being compared in the analysis. [25]

In the comparison of DOACs and warfarin, a patient with severe renal failure at baseline will not be eligible for treatment with a DOAC and should therefore be excluded from the analysis and from the population in which we are estimating a treatment effect. Note that the same exclusion criteria would be present in an RCT, and the effect of

interest would only be estimated in the population eligible for both treatment options, as this is the population where the treatment comparison is clinically relevant. As contraindication for DOACs prohibits treatment, any average population-level effect will be due to the effect in those eligible to receive a DOAC.

Intuitively, estimation of average causal effects requires positivity because patients who were not eligible to receive a medication may be entirely incomparable to those who did. In our example, patients ineligible for a DOAC will always get warfarin and may be incomparable to patients who were eligible for either. We will have no information on how ineligible patients would have fared under the contraindicated medication, thereby requiring an extrapolation of the observed effects using patients who were actually eligible in order to estimate an overall effect [26] which may give invalid results.

In Figure 1(a), we give an extreme illustration of the violation of this condition with a single continuous baseline covariate, L, a single binary exposure, and a continuous outcome, Y, for which larger values indicate improved health. The exposure to warfarin or a DOAC is indicated by the plotting labels (either "0" or "1", corresponding to exposure to warfarin or a DOAC, respectively). Suppose that subjects with a value of L below a certain threshold will deterministically not be exposed, and subjects with a value beyond this threshold will be exposed. Clearly, L confounds the relationship between the exposure and outcome as its value affects which treatment is received and, given the plot, appears to affect the outcome as well. However, even knowing L, one cannot be certain whether the differences between exposed and unexposed subjects are due to the treatment or L itself. In Figure 1(b), we plotted the simple regression lines fit with just the unexposed observations (dotted line) and just the exposed observations (solid line), respectively. These lines appear to fit the data well. However, one may fallaciously infer that the treatment is effective for the values of L prior to where the lines cross and harmful afterwards. But clearly, this conclusion was made based on extrapolations of the regression lines to regions of L where each treatment was never observed. In reality, these data were generated with a constant positive treatment effect over the entire range of L. The R code for the data generation and plots are available in Box 1.

Figure 1 about here

Box 1: R code for the positivity violation data generation and plots

```
set.seed(934)

#Set up the axis
plot(xlab="L",ylab="Y",xlim=c(0,5),ylim=c(0,5),x=0,y=0,xaxt="n",
yaxt="n",type="l")
axis(side=1,labels=F)
axis(side=2,labels=F)

#Generate 20 data points (l, a, y)
ssize<-20
l<-runif(min=0,max=5,n=ssize)
a<-rep(0,ssize)
a[l<2.5]<-0
a[l>=2.5]<-1
y<-(1+l-0.2*l^2)+a+rnorm(n=ssize,sd=0.5)

#Plot the points
points(x=l[a==0],y=y[a==0],pch="0")
points(x=l[a==1],y=y[a==1],pch="1")

#Fit and plot the regression lines
mod0<-lm(y~l,subset=(a==0))
abline(mod0,lty=2,lwd=2)
mod1<-lm(y~l,subset=(a==1))
abline(mod1,lty=1,lwd=2)
```


Aside from the statistical difficulty in estimating a treatment effect when positivity is violated, one may question the existence of an average "effect" when there are members of the population who could never have received one or the other exposure. Note that the positivity violation resulting from a true contraindication, where no patient with certain covariate values should receive a given treatment, is distinct from a *practical* positivity violation (or "data sparsity"), where no or few patients with certain covariate values are *observed* to receive a given treatment. The latter may be addressed using statistical methods [27,28] while the former indicates an ill-defined causal question. [29]

4. CONTRAINDICATION OCURRING AFTER INITIAL TREATMENT EXPOSURE

The solution to the problem of baseline violations of the positivity condition is conceptually simple; one should remove subjects from the analysis who are known to be ineligible for both treatments and evaluate the treatment effects on the remaining set. But what happens if subjects *become* ineligible for a treatment during follow-up? The estimation of expected population outcomes under sustained exposure to

medication in settings where exposure may vary in time also requires *longitudinal* positivity, where every subject must be eligible for either medication given the values of their confounding variables at every time point at which a treatment change is potentially made. [30] In our example, we are interested in the effect of sustained exposure to warfarin versus sustained exposure to a DOAC. Thus, longitudinal positivity will be violated if a patient develops a contraindication to either medication at any time point. In this section, we demonstrate how ignoring or improperly adjusting for longitudinal positivity violations can result in misleading conclusions. In the subsequent section we describe treatment strategies and a modified version of IPW which produce valid conclusions.

4.1 Data Generation

To illustrate the problem, we generate data according to three distinct scenarios illustrated through Directed Acyclic Graphs (DAGs). DAGs are used to represent the causal relationships between variables [31]. The arrows (or directed edges) between variables in the graph indicate that one variable (the parent) affects another (the child) in the direction of the arrow. A path between two variables is an unbroken route that proceeds along or against the direction of the arrows. A path is considered open unless 1) conditioning on a variable blocks the path (denoted by a square around the variable) or 2) the path goes through an unadjusted *collider*: a variable that is affected by two parent variables. Adjusting for a collider opens the previously blocked path. If there is an open path between the outcome and exposure other than the path of interest, estimation of the causal effect will be biased.[31]

The first two scenarios we consider correspond to the top DAG in Figure 2 and the third corresponds to the bottom DAG. Specifically, the variable A_0 represents the exposure to a DOAC, "1", or warfarin, "0", in the first year of follow-up, and A_1 represents exposure in the second year. The variable RF denotes the renal failure status (yes/no) at the beginning of the second year; patients who develop renal failure necessarily receive warfarin at the next time point (A_1=0). The outcome Y is major bleeding event (yes/no) by the end of follow-up. The baseline variable W represents a single confounder of the initial exposure and outcome while U represents an unmeasured baseline risk factor of major bleeding. For instance, U may be the use of over the counter non-steroidal anti-inflammatory drugs. This simplified structure assumes that the outcome may not occur prior to the second year and is observed for all subjects. In these DAGs, no arrows go from either exposure to Y, indicating that there is no true treatment effect. In the top DAG, RF causes Y and exposure has no effect on RF. In the bottom DAG, we consider the illustrative scenario where RF no longer causes Y, exposure causes RF, and the

unmeasured risk factor U also affects RF. RF is therefore a collider [32] in the bottom DAG as it is caused by A_0 and U.

Figure 2 about here

The data generation R code for the three scenarios is presented in Box 2. Importantly, the outcome models for the first and third scenarios only include main terms while the outcome model in the second scenario includes an interaction between W and U. Summaries of the generated datasets of n=25,000 subjects each are in Web Table 1 in the Web Appendix. From this table, we see that after the initial exposure, roughly 10% developed renal failure though the percentages are imbalanced between those with early exposure to warfarin and to DOACs, respectively. Almost all subjects without renal failure who were taking a DOAC stayed on DOACs while 17% previously on warfarin switched over to DOACs. No subjects who developed renal failure subsequently took DOACs. Without adjustment, a slightly higher percentage of subjects had major bleeding in the initial warfarin group compared to the initial DOAC group.

Box 2: R code for the data generation of the DOAC-warfarin example

```
set.seed(454)

#Baseline covariates
W<-rnorm(n=25000)
U<-rnorm(n=25000)

#Early exposure: 1 is DOAC, 0 is warfarin
A0<-rbinom(size=1,n=25000,p=plogis(0.5-0.5*W))

#Renal failure (scenarios 1 and 2)
RF<-rbinom(size=1,n=25000,p=plogis(-2.8+W))
#Renal failure (scenario 3)
RF<-rbinom(size=1,n=25000,p=plogis(-4+W+1.5*U+A0))

#Late exposure
A1<-rbinom(size=1,n=25000,plogis(-2+5*A0-2*W))
A1[RF==1]<-0 #Everyone who gets renal failure goes on Warfarin

#Outcome (scenario 1)
Y<-rbinom(size=1,n=25000,p=plogis(-2+0.5*RF+0.5*W+2*U))
#Outcome (scenario 2)
Y<-rbinom(size=1,n=25000,p=plogis(-2+0.5*RF+0.5*W+2.5*W*U))
#Outcome (scenario 3)
Y<-rbinom(size=1,n=25000,p=plogis(-2+0.5*W+2*U))

#Save observed data as dataframe
DAT=as.data.frame(cbind(W,A0,RF,A1,Y))
```

4.2 Regression analyses

We precede with a series of possible regression analyses that may or may not include renal failure in the model. Since Y is binary, a naïve approach might use logistic regression analysis. The R code for all analyses is presented in Box 3.

4.2.1 Intent-to-treat analyses

We begin with intent-to-treat (ITT) analyses, which consider only the initial exposure to medication (A_0). Given that RF is a mediator of the relationship between A_0 and Y (i.e. an intermediate variable along the path from A_0 to Y), we do not need to adjust for it in order to estimate the causal effect of A_0. However, we must adjust for the confounder W.

In Table 1 we present the odds ratios and confidence intervals from logistic regression models, unadjusted and adjusted for W, respectively. From this table, we see that with adjustment for the confounder W, we do not conclude in any scenario that there is any difference in effect between early exposure to DOACs and warfarin on major bleeding. Without adjustment, it appears that DOACs are safer. Given that in this simulated example we know that there is no relative effect of DOACs versus warfarin, we therefore see that the adjusted ITT analysis gives the correct result.

Table 1 about here

4.2.2 Adherence adjusted analyses

We may also be interested in the effect of the exposure over time. For instance, we may be interested in the effect of both early (A_0) and more recent (A_1) exposure to DOACs on the outcome, where a corresponding analysis is often called “adherence adjusted”. How then should we proceed?

Simple regression analysis – which we will see leads to incorrect conclusions -- would involve including covariates A_0 and A_1 in the regression. Conceptually, we want to adjust for all confounders of the A_0 – Y and A_1 – Y relationships. In the upper DAG in Figure 2 (scenarios 1 and 2), RF is a confounder of the A_1 – Y relationship, but it is not a mediator of A_0 – Y. So adjustment for RF may give the correct conclusion as long as the model correctly extrapolates outcomes under A_1=1 for patients with RF. In the lower DAG (scenario 3), in addition to being a confounder of A_1, RF is a collider of A_0 and U, so adjusting for RF will create an open path between U and A_0 leading to a misleading (non-causal) association between A_0 and Y. In fact, even including A_1 as a covariate will create an open path between U and A_0 because A_1 is a descendant of RF. [32] Thus in both cases,

both the adjusted and unadjusted analyses may produce misleading conclusions about the existence of an effect.

In Table 2 we give the results of four naïve modeling strategies, all of which are logistic regressions with A_0 and A_1 as covariates. The details of these models are:

- M1: Ignore RF. Include W as a covariate.
- M2: Subset on the subjects who do not develop renal failure (RF=0). Include W as a covariate.
- M3: Subset on subjects who do not develop renal failure (i.e. we artificially censor them) and add in inverse probability of censoring weights.[4] These weights are estimated using a censoring model that adjusts for the baseline covariate and early exposure.
- M4: Include RF and W as covariates.

Note that for M3, the automatic standard error estimate in the regression function (lm or glm) output, based on the Fisher information matrix, does not take into account the uncertainty of the weights. We provide the code for a nonparametric bootstrap in the Web Appendix A2. [33]

For M1, the results indicate that in all scenarios exposure to DOACs is harmful in the early time period and protective in the later time period. The effect of A_1 is biased because we have not adjusted for RF, a confounder of the A_1 – Y association. In scenarios 1 and 2, the effect of A_0 is biased because adjusting for A_1 creates an open path A_0 ->RF -> Y. In scenario 3, adjusting for A_1 opens the path A_0 ->U -> Y

For M2 and M3 where we subset on RF=0, we obtain null (correct) results for both exposures in the first scenario. This is because we have a nearly correct model of the outcome in the subpopulation without renal failure which occurs independently of exposure conditional on measured factors. In the second scenario, the results suggest that later exposure to DOACs is protective, because the outcome model is no longer correctly specified with respect to W (due to the interaction between W and the unobserved U). In the third scenario, the results suggest that only early exposure is slightly protective. This occurs because the selection on the collider RF induced a correlation between A_1 and Y. [32] When adjusting for RF in the regression model (M4), we have similar conclusions as the previous two models.

Table 2 about here

Box 3: R code for the ITT and adherence adjusted regression analyses

```
#ITT analysis without adjustment
ITT1<-glm(Y~A0,family=binomial())
exp(coefficients(ITT1)) #OR estimates
exp(confint(ITT1)) #95% confidence intervals on the OR scale

#ITT analysis with adjustment
ITT2<-glm(Y~A0+W,family=binomial())
exp(coefficients(ITT2)) ; exp(confint(ITT2))

#Adherence adjusted model with adjustment (M1)
mod1<-glm(Y~A0+A1+W,family=binomial())
exp(coefficients(mod1)) ; exp(confint(mod1))

#Adherence adjusted model with subsetting on RF (M2)
mod2<-glm(Y~A0+A1+W,family=binomial(),subset=(RF==0))
exp(coefficients(mod2)) ; exp(confint(mod2))

#Adherence adjusted model with subsetting on RF
#with censoring weights (M3)
censor_RF<-function(DAT){ #Create function for use in bootstrap
pC<-1-predict(glm(RF~A0+W,family=binomial(),data=DAT),type="response")
cmod<-
glm(Y~A0+A1+W,family=quasibinomial(),subset=(RF==0),weights=(1/pC),data
=DAT)
return(coefficients(cmod)) }
exp(censor_RF(DAT))
bsres<-bootstrap.vect(censor_RF,DAT)
#code for bootstrap.vect available in the Web Appendix A2
exp(cbind(bsres$CIlow,bsres$CIhigh))

#Adherence adjusted model with adjustment for RF (M4)
mod4<-glm(Y~A0+A1+W+RF,family=binomial())
exp(coefficients(mod4)) ; exp(confint(mod4))
```

From the above results, it is clear that ignoring the development of post-exposure contraindications, or treating them as censoring or as a covariate, can lead to false conclusions, though this depends on the underlying structure of the data generation. The bias was largest when RF was a collider of past treatment and unmeasured risk factors of the outcome. Even when RF is not a collider, bias can arise from outcome model misspecification over regions of the covariate space where only one exposure group is observed. Because we are extrapolating, we lack the needed data (patients with A_1=1 and RF=1) to judge the appropriateness of the model. The censoring weights were not able to correct for the bias as developing RF is both a confounder of A_1 and Y and the determinant of censoring. Since the weights cannot adjust for RF (which would

result in zero probabilities of not being censored and thus weights of infinite size), this confounding remains unadjusted for.

From the DAG and the above analyses, we see that post-exposure contraindication is a time-dependent confounder, as it affects further treatment and the outcome. It is well-known in epidemiology that it is necessary to adjust for time-dependent confounding in order to estimate the effect of treatments that may change in time. [10] However, renal failure also creates a positivity violation for the exposure A_1 because anyone who has the contraindication has a zero probability of receiving $A_1=1$. Therefore, we cannot adjust for the time-dependent confounder using inverse probability of treatment weights in the standard way. (Note that this would involve estimating the probability of $A_1=1$ for all subjects, including those with renal failure, for whom we would obtain a probability of zero. [1]) We demonstrate one way to appropriately analyze this type of data in the next section.

5. TREATMENT STRATEGIES WITH IPW

Using the counterfactual approach to causal inference, one may define mean population-level effects as the contrast between summaries of the outcomes that would have been observed had the entire population of interest taken treatment 1 versus the entire population taken treatment 2. For example, the difference in the means of the counterfactual outcomes under either treatment is commonly known as the average treatment effect. However, in this context, it is unrealistic to impose sustained treatment with DOACs regardless of the intermediate patient outcomes due to the above-mentioned potential for developing contraindications. Therefore, the counterfactual contrast between the fixed treatment strategies of "always treat with a DOAC" versus "always treat with warfarin" cannot be imposed on the population of interest. Specifically, the positivity condition for the always-treat-with-DOACs strategy does not hold for patients who develop renal failure.

5.1 Contrasts of interest

However, we may alternatively contrast the realistic treatment strategies R_D: "always treat with a DOAC unless the patient develops renal failure, in which case treat with warfarin" versus R_W: "always treat with warfarin". These are called "treatment strategies" because they correspond with how a doctor may make treatment decisions in a clinical setting based on a patient's characteristics or outcomes after previous treatment. These strategies may be made more complex to correspond with what is

done in practice so that competing treatment strategies may be compared through analysis. [8]

In this setting and with these definitions, the condition of positivity means that patients must have a non-null probability of 1) initiating warfarin and then remaining on warfarin and 2) initiating DOACs and then either staying on DOACs or switching if renal failure develops. It does not mean that every patient in the analysis must have followed one of these two strategies, just that they all had the potential to follow either one. For estimation with IPW, we also require that we have observed a sizeable number of patients following each of these strategies (as the analysis will be powered by these sample sizes).

Define Y(R) to be the potential outcome that a patient *would have had* if they had (perhaps counterfactually) followed the treatment regimen R. The marginal parameter of interest may then be $P[Y(R_D)=1]-P[Y(R_W)=1]$, defined as the difference in probabilities of major bleeding under the counterfactual situations where all patients had followed strategy R_D versus R_W. We could alternatively estimate the marginal risk ratio $P[Y(R_D)=1]/P[Y(R_W)=1]$ or the marginal odds ratio $(P[Y(R_D)=1]/\{1-P[Y(R_D)=1]\})/(P[Y(R_W)=1]/\{1-P[Y(R_W)=1]\})$ as desired.

5.2 Applying IPW

A modified version of IPW can be applied to estimate any of these contrasts. Standard implementation approaches can be found elsewhere[1,10] but notably involve estimating the conditional probabilities of exposure to both warfarin and DOACs at each time point, the latter of which we know to be zero for certain subjects. The goal of standard IPW is to reweight the sample so that treated and untreated subjects at each time point (with the same treatment histories) are comparable in terms of their covariate histories, as would occur in a sequentially randomized trial with randomization probabilities only conditional on past treatment. The reweighted population is often referred to as a pseudo-population.[25]

The goal of the modified procedure is to reweight the subjects in order to achieve balance in terms of covariate histories between the *treatment strategy* groups rather than between those treated vs not. The first step is to use all patient observations to estimate the probability of following each strategy through the two time points, conditional on confounders. For those subjects who followed one or the other strategy, we construct weights using the inverse of their probability of following the strategy that was followed. With those same subjects, we then fit a weighted regression model adjusting for a variable indicating which strategy was followed. The type of regression

will determine the type of contrast estimated (i.e. the risk difference, risk ratio, or odds ratio, as defined above).

To estimate the probability of following strategy R_W, one must first fit a model to estimate $P(A_0 = 0 \mid W)$, the conditional probability of initially receiving warfarin. A second model may estimate $P(A_1 = 0 | A_0 = 0, W, RF = 0)$, the conditional probability of subsequently receiving warfarin if warfarin was initially received amongst subjects who did not develop renal failure. We can use this probability to compute

$$P(A_1 = 0 \mid A_0 = 0, W, RF) = \begin{cases} P(A_1 = 0 \mid A_0 = 0, W, RF = 0) & \text{if } RF = 0 \\ 1 & \text{if } RF = 1, \end{cases}$$

each subject's conditional probability of receiving warfarin at the second time point had warfarin been received at the first. This probability is set to 1 for subjects who develop renal failure. The probability of following R_W is therefore the product of $P(A_0 = 0 \mid W)$ and $P(A_1 = 0 \mid A_0 = 0, W, RF)$. We compute this probability, called p_W, for each patient.

For the probability of following strategy R_D, we first estimate the conditional probability of receiving DOACs at the first time point, $P(A_0 = 1 \mid W)$. For the second time point, we compute $P(A_1 = 1 \mid A_0 = 1, W, RF = 0)$. The probability of following the strategy R_D throughout the two time points is then

$$p_D = P(A_0 = 1 \mid W) * \begin{Bmatrix} P(A_1 = 1 \mid A_0 = 1, W, RF = 0) & \text{if } RF = 0 \\ 1 & \text{if } RF = 1 \end{Bmatrix}$$

which we can compute for each patient. Code for estimating these probabilities with logistic regressions is given in Box 4.

The final step is to fit a weighted regression. The weights are $1/p_W$ for patients who followed strategy R_W and $1/p_D$ for patients who followed strategy R_D. All other patients receive a weight of 0 (i.e. are excluded from the final regression step). A linear regression conditional on an indicator of the treatment strategy followed will produce coefficient estimates corresponding to estimates of the marginal risk difference. A log-linear regression will produce estimates of the (log) marginal risk ratio, and a logistic regression will produce (log) odds ratios. In Box 5, we give the code to estimate each of these contrasts of interest, using the bootstrap function to construct confidence intervals provided in the Web Appendix A2.

Box 4: Estimating the probabilities needed for the IPW weights

```
#Model for probability of treatment at tp 0
A0mod<-glm(A0~W,family=binomial())
#Model for probability of treatment at tp 1 conditional on A0
A1mod<-glm(A1~A0+W,family=binomial(),subset=(RF==0))

#Estimated probability of A0 = 0 (warfarin)
P0<-1-predict(A0mod,type="response")
#Estimated probability of following RW for tps 0 & 1
PW<-((1-predict(A1mod,
newdata=as.data.frame(cbind(W,A0=0)),type="response"))*(RF==0)+1*
(RF==1))*P0

#Estimated probability of A0 = 1 (DOAC)
P1<-1-P0
#Estimated probability of following RD for tps 0 & 1
PD<-
((RF==0)*predict(A1mod,newdata=as.data.frame(cbind(W,A0=1)),type=
"response")+ (RF==1))*P1
```

Following this procedure for each of the scenarios, the estimates and bootstrapped 95% confidence intervals of the marginal odds ratio were 1.05 (0.97, 1.13), 0.93 (0.85, 1.03), and 1.04 (0.96, 1.11), respectively. The estimates of the marginal risk difference and ratio were similarly null.

Conclusions about the performance of statistical estimators can be misleading when drawn from the analysis of a single finite data sample. To validate these conclusions, we present the results of a Monte Carlo simulation study in the Web Appendix A3 with the same data-generating functions and models as above. We also investigate a fourth scenario where A_0 and A_1 have small positive effects on the outcome.

One will often obtain superior estimation (lower in finite sample bias and estimation variance) by constructing stabilized weights. [25] We give the related code in the Web Appendix A4.

Box 5: Fitting weighted regressions to obtain IPW estimates

```
#subset of patients following RW at first AND second time points
RW01<-(A0==0&A1==0)
#subset of patients following RD at first AND second time points
RD01<-(A0==1&A1==1) | (A0==1&RF==1&A1==0)

#Weights
w<-(1/PW*RW01 + 1/PD*RD01)

#Those patients who followed either strategy
sub<-(RD01==1|RW01==1)

#RISK DIFFERENCE: weighted linear regression
IPTWmod1<-lm(Y~RD01,weights=w,subset=sub)

#RISK RATIO: weighted log-linear model
IPTWmod2<-
 glm(Y~RD01,weights=w,subset=sub,family=quasibinomial(link=log))

#ODDS RATIO: weighted logistic regression
IPTWmod3<-
 glm(Y~RD01,weights=w,subset=sub,family=quasibinomial())

#code to obtain the standard error and 95% CI using
#nonparametric bootstrap is available in the Web Appendix A2.
```

6. DISCUSSION

The treatment-strategy IPW method we described here can be used in a more general longitudinal context by estimating the necessary probabilities at each time point to obtain the final patient-specific weights. [2] However, because a number of patients receive a zero weight, the standard errors can become large. Other more efficient methods that can estimate the marginal parameters described in the previous section include the parametric g-formula [34] and longitudinal targeted minimum loss-based estimation, [35] which can be implemented with machine learning methods. For instance, one may use the nested outcome expectation approach[36] where nuisance modeling of the outcome should be conditional on observed medication exposure. The predictions from these models should then be made based on the treatment-strategy of interest conditional on the observed history.

We chose to estimate what is sometimes known as a "saturated" marginal structural model, [30] i.e. the mean counterfactual outcome under a specific intervention. In our case, this intervention corresponds to each of the treatment strategies. One may also choose to model the outcome with respect to baseline covariates using a single

weighted regression with the same inverse probability weights, i.e. estimate the parameters of a marginal structural model. Alternative approaches to estimating effects in the presence of time-varying eligibility to treatment exist [2,37] though a comparison is beyond the scope of this paper.

We demonstrated the potential for bias in the setting of a prospective cohort with measured baseline confounder W. However, the bias due to colliders and longitudinal positivity violations varies based on the underlying mechanisms and the modeling approach used. The conclusions regarding the bias caused by improperly dealing with time-varying eligibility are also relevant for the analysis of RCTs in which contra-indications may develop over time. Similar to observational studies, we may remove all patients with contraindications to either treatment option at baseline but it is not advisable to censor patients post-randomization due to emerging contraindications (models M2 and M3). [3] Nearly identical ITT or treatment strategy approaches can be used in RCTs as well. [3] We therefore encourage investigators to take note of time-varying eligibility in their studies and plan their analytical protocols with these lessons in mind.

REFERENCES

1. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol.* 2008;168(6):656-664.
2. Gong Q, Schaubel DE. Estimating the average treatment effect on survival based on observational data and using partly conditional modeling. *Biometrics.* 2017;73(1):134-144.
3. Hernan MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. *N Engl J Med.* 2017;377(14):1391-1398.
4. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics.* 2000;56(3):779-788.
5. Mertens BJ, Datta S, Brand R, Peul W. Causal effect estimation strategies in a longitudinal study with complex time-varying confounders: A tutorial. *Stat Methods Med Res.* 2017;26(1):337-355.
6. Fewell Z, Hernán MA, Wolfe F, Tilling K, Choi H, Sterne JA. Controlling for time-dependent confounding using marginal structural models. *The Stata Journal.* 2004;4(4):402-420.
7. Murray EJ, Hernan MA. Adherence adjustment in the Coronary Drug Project: A call for better per-protocol effect estimates in randomized trials. *Clin Trials.* 2016;13(4):372-378.
8. Hernan 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.
9. Chakraborty B, Murphy SA. Dynamic Treatment Regimes. *Annu Rev Stat Appl.* 2014;1:447-464.
10. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 2000;11(5):561-570.
11. Excellence NIfHaC. *Atrial fibrillation: management.* National Clinical Guidelines Center2014.
12. Connolly SJ, Pogue J, Eikelboom J, et al. Benefit of oral anticoagulant over antiplatelet therapy in atrial fibrillation depends on the quality of international normalized ratio control achieved by centers and countries as measured by time in therapeutic range. *Circulation.* 2008;118(20):2029-2037.
13. Law S, Ghag D, Grafstein E, Stenstrom R, Harris D. A pharmacoeconomic study of traditional anticoagulation versus direct oral anticoagulation for the treatment of venous thromboembolism in the emergency department. *CJEM.* 2016;18(5):340-348.
14. Patel MR, Mahaffey KW, Garg J, et al. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med.* 2011;365(10):883-891.
15. Connolly SJ, Ezekowitz MD, Yusuf S, et al. Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med.* 2009;361(12):1139-1151.
16. Granger CB, Alexander JH, McMurray JJ, et al. Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med.* 2011;365(11):981-992.
17. Lopez-Lopez JA, Sterne JAC, Thom HHZ, et al. Oral anticoagulants for prevention of stroke in atrial fibrillation: systematic review, network meta-analysis, and cost effectiveness analysis. *BMJ.* 2017;359:j5058.

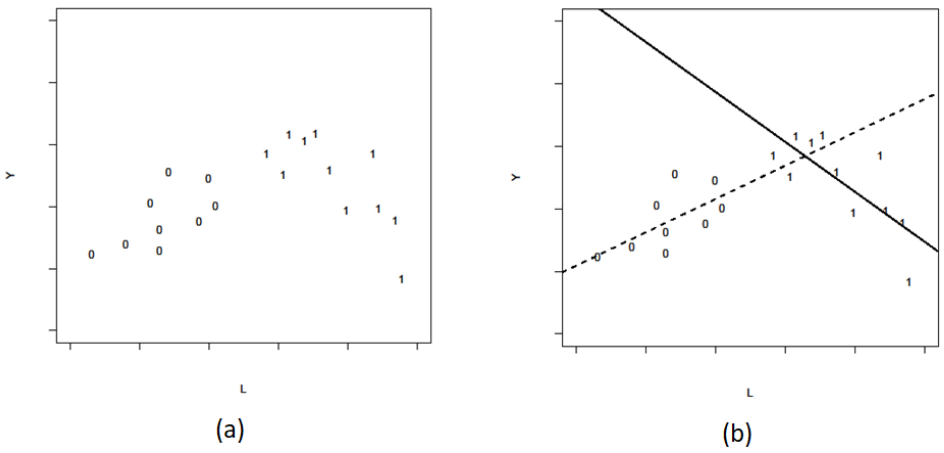
18. Vinogradova Y, Coupland C, Hill T, Hippisley-Cox J. Risks and benefits of direct oral anticoagulants versus warfarin in a real world setting: cohort study in primary care. *BMJ.* 2018;362:k2505.
19. Barrett A, Moore M, Ferrins P, Thornton P, Murphy P, Quinn J. From a direct oral anticoagulant to warfarin: reasons why patients switch. *Ir J Med Sci.* 2018;187(3):719-721.
20. Ikeda T, Yasaka M, Kida M, Imura M. A survey of reasons for continuing warfarin therapy in the era of direct oral anticoagulants in Japanese patients with atrial fibrillation: the SELECT study. *Patient Prefer Adherence.* 2018;12:135-143.
21. Pisters R, Lane DA, Nieuwlaat R, de Vos CB, Crijns HJ, Lip GY. A novel user-friendly score (HAS-BLED) to assess 1-year risk of major bleeding in patients with atrial fibrillation: the Euro Heart Survey. *Chest.* 2010;138(5):1093-1100.
22. Di Lullo L, House A, Gorini A, Santoboni A, Russo D, Ronco C. Chronic kidney disease and cardiovascular complications. *Heart Fail Rev.* 2015;20(3):259-272.
23. Rubin DB. Randomization Analysis of Experimental-Data - the Fisher Randomization Test - Comment. *J Am Stat Assoc.* 1980;75(371):591-593.
24. Hernan MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obesity.* 2008;32:S8-S14.
25. Hernán MA, Robins JM. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC; 2020.
26. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol Methods.* 2008;13(4):279-313.
27. Ju C, Schwab J, van der Laan MJ. On adaptive propensity score truncation in causal inference. *Stat Methods Med Res.* 2019;28(6):1741-1760.
28. Petersen ML, Porter KE, Gruber S, Wang Y, van der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Stat Methods Med Res.* 2012;21(1):31-54.
29. Westreich D, Cole SR. Invited Commentary: Positivity in Practice. *Am J Epidemiol.* 2010;171(6):674-677.
30. Robins JM. Robust Estimation in Sequentially Ignorable Missing Data and Causal Inference Models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science*1999:6-10.
31. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology.* 1999;10(1):37-48.
32. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. *Epidemiology.* 2004;15(5):615-625.
33. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med.* 2016;35(30):5642-5655.
34. Zhang Y, Young JG, Thamer M, Hernan MA. Comparing the Effectiveness of Dynamic Treatment Strategies Using Electronic Health Records: An Application of the Parametric g-Formula to Anemia Management Strategies. *Health Serv Res.* 2018;53(3):1900-1918.
35. van der Laan MJ, Gruber S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int J Biostat.* 2012;8(1).
36. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics.* 2005;61(4):962-973.
37. Wang L, Zhou XH, Richardson TS. Identification and estimation of causal effects with outcomes truncated by death. *Biometrika.* 2017;104(3):597-612.

ITT unadjusted regression, $Y \sim A_0$			
Covariate	**OR Scenario 1**	**OR Scenario 2**	**OR Scenario 3**
A_0	0.86(0.81,0.92)*	0.85(0.79,0.90)*	0.87(0.82,0.92)*
ITT adjusted regression, $Y \sim A_0 + W$			
Covariate	**OR Scenario 1**	**OR Scenario 2**	**OR Scenario 3**
A_0	1.02(0.96,1.08)	0.96(0.90,1.02)	1.01(0.95,1.08)
W	1.43(1.39,1.48)*	1.32(1.28,1.36)*	1.40(1.36,1.45)*

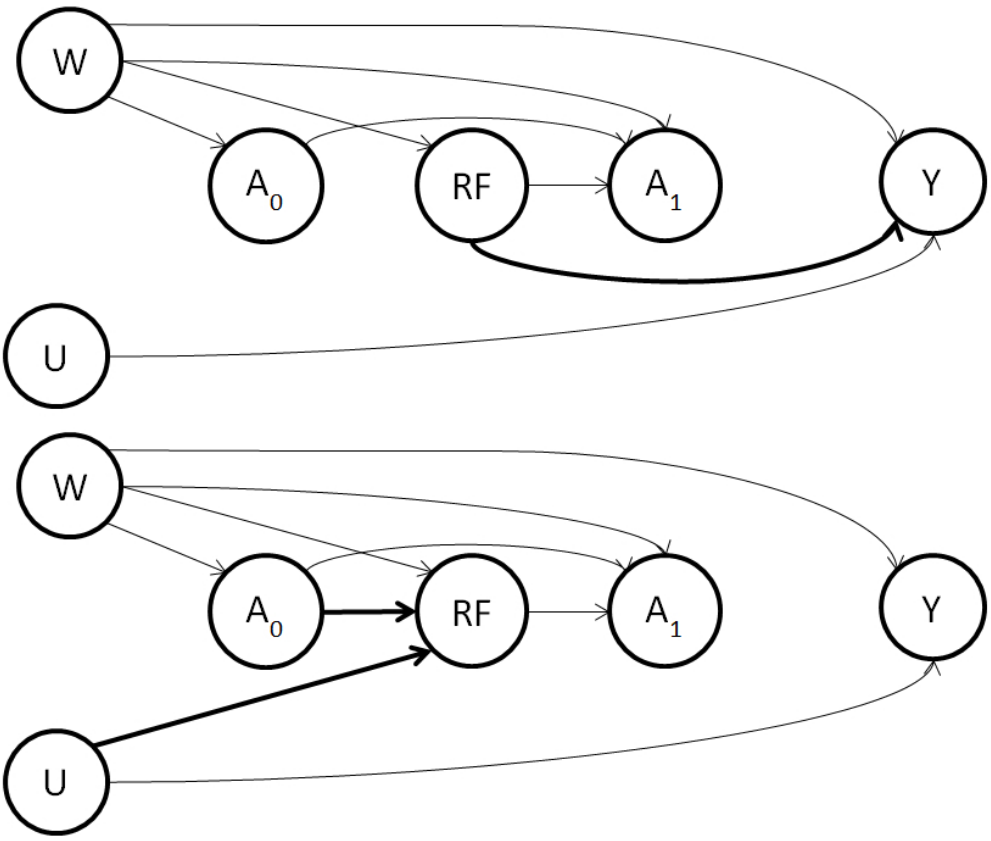
Table 1. ITT Analyses With Odds Ratio (OR) Estimates in the Three Scenarios.
*Asterisks indicate whether the confidence interval excludes the null.

M1. Adherence adjusted regression, $Y \sim A_0 + A_1 + W$			
Covariate	**OR (Scenario 1)**	**OR (Scenario 2)**	**OR (Scenario 3)**
A_0	1.15(1.06,1.25)*	1.13(1.04,1.23)*	1.63(1.51,1.76)*
A_1	0.82(0.75,0.90)*	0.77(0.70,0.84)*	0.41(0.38,0.45)*
W	1.38(1.34,1.43)*	1.26(1.22,1.31)*	1.20(1.16,1.24)*
M2. Adherence adjusted regression subsetting for RF=0, $Y \sim A_0 + A_1 + W$			
Covariate	**OR (Scenario 1)**	**OR (Scenario 2)**	**OR (Scenario 3)**
A_0	1.05(0.95,1.16)	1.06(0.96,1.17)	0.90(0.81,1.00)*
A_1	0.92(0.83,1.03)	0.83(0.75,0.93)*	0.92(0.82,1.03)
W	1.38(1.33,1.44)*	1.21(1.17,1.26)*	1.26(1.21,1.31)*
M3. Adherence adjusted regression subsetting for RF=0 and using inverse probability of censoring weights†, $Y \sim A_0 + A_1 + W$			
Covariate	**OR (Scenario 1)**	**OR (Scenario 2)**	**OR (Scenario 3)**
A_0	1.05(0.95,1.16)	1.05(0.95,1.17)	0.90(0.82,0.99)*
A_1	0.93(0.84,1.04)	0.84(0.74,0.93)*	0.92(0.83,1.03)
W	1.38(1.33,1.44)*	1.26(1.21,1.31)*	1.26(1.21,1.30)*
M4. Adherence adjusted regression, $Y \sim A_0 + A_1 + W + RF$			
Covariate	**OR (Scenario 1)**	**OR (Scenario 2)**	**OR (Scenario 3)**
A_0	1.08(1.00,1.18)	1.05(0.96,0.31)	0.85(0.78,0.93)*
A_1	0.90(0.82,0.99)*	0.86(0.78,0.95)*	0.94(0.85,1.04)
W	1.38(1.33,1.42)*	1.25(1.21,1.30)*	1.23(1.19,1.27)*
RF	1.31(1.17,1.46)*	1.42(1.28,1.59)*	4.27(3.83,4.75)*

Table 2. Regression Analyses Investigating the Effect of Exposure Over Time. †The confidence interval for inverse censoring uses a nonparametric bootstrap with 500 resamples to take into account the uncertainty in the weights. *****Asterisks indicate whether the confidence interval excludes the null.



(a) A simple setting with a single baseline covariate, L, and outcome Y. The observations are denoted as either "0" (warfarin) or "1" (DOAC) corresponding to a binary exposure. In this extreme example of violation of the positivity condition, small values of L determine that a subject will not be exposed, while larger values imply that they will. (b) The same observations with a dotted regression line fit with the non-exposed subjects and a solid regression line fit with the exposed subjects.



DAGs representing the data-generation in the DOAC-warfarin example, in scenarios 1 and 2 (above) and scenario 3 (below). In the bottom DAG, renal failure (RF) is a collider of A0 and U while in the top DAG it is not. The bolded lines indicate the arrows that are different between the DAGs.

217x183mm (96 x 96 DPI)

WEB APPENDIX for "A tutorial on dealing with time-varying eligibility for treatment: Comparing the risk of major bleeding with DOACs versus warfarin"

AUTHORS: Mireille E. Schnitzer[1], Robert W. Platt[2], and Madeleine Durand[3]

AFFILIATIONS

[1]Faculty of Pharmacy and Department of Social and Preventative Medicine, Université de Montréal, Montreal, Canada
[2]Departments of Pediatrics and of Epidemiology, Biostatistics, and Occupational Health, McGill University, and the Research Institute of the McGill University Health Centre, Montreal, Canada
[3]Internal Medicine Service, Department of Medicine, Centre Hospitalier de l'Université de Montréal, and the Research Center of Centre Hospitalier de l'Université de Montréal, Montréal, Canada

A1. SUMMARY MEASURES FOR THE THREE SIMULATED SCENARIOS

	A_0=1 (DOAC)	A_0=0 (warfarin)
Scenario 1		
	n=15314	*n*=9686
Baseline variables		
W, *mean (sd)*	-0.17 (0.98)	0.30 (0.96)
U, *mean (sd)*	0.00 (0.99)	0.00 (1.00)
Post-initial exposure variables		
RF, *p*	0.07	0.10
A_1, *p*	0.84	0.15
A_1 amongst those with RF=0	0.91	0.17
A_1 amongst those with RF=1	0	0
Y, *p*	0.22	0.25
Scenario 2		
	n=15314	*n*=9686
Baseline variables		
W, *mean (sd)*	-0.17 (0.98)	0.30 (0.96)
U, *mean (sd)*	0.00 (0.99)	0.00 (1.00)
Post-initial exposure variables		
RF, *p*	0.07	0.10
A_1, *p*	0.84	0.15
A_1 amongst those with RF=0	0.91	0.17
A_1 amongst those with RF=1	0	0
Y, *p*	0.21	0.24
Scenario 3		
	n=15314	*n*=9686
Baseline variables		
W, *mean (sd)*	-0.17 (0.97)	0.30 (0.96)
U, *mean (sd)*	0.00 (0.99)	0.00 (1.00)
Post-initial exposure variables		
RF, *p*	0.14	0.07
A_1, *p*	0.78	0.16
A_1 amongst those with RF=0	0.91	0.17
A_1 amongst those with RF=1	0	0
Y, *p*	0.22	0.24

Web Table 1. Baseline and Post-initial Exposure Summary Measures for Both Initial Exposure Groups in the Simulated Data Example.

A2. CODE FOR GENERAL NONPARAMETRIC BOOTSTRAP AND APPLIED TO IPTW

When inverse probability weights are used, it is often recommended to estimate the standard error using the bootstrap. Here we provide a simple general nonparametric implementation that can be applied to an arbitrary estimator of a vector parameter.

Box A2.1: Nonparametric bootstrap function for a vector parameter

```
bootstrap.vect<-function( funct, DAT , K=500, n=dim(DAT)[1] ){
result.full<-funct(DAT)
mat.est<-matrix(0,ncol=length(result.full),nrow=K)
for (k in 1:K){
 #resample the original dataset
 resamp<-sample(1:n, n, replace=T)
 DATk<-as.data.frame(DAT[resamp,])
 estimate<-funct(DATk)
 mat.est[k,]<-estimate
 }
VAR=apply(mat.est,2,var)
CIlow=apply(mat.est,2,function(x) quantile(x,0.025))
CIhigh=apply(mat.est,2,function(x) quantile(x,0.975))

#Output: variance, 95% confidence limits,
#and the bootstrap estimates
return(list(VAR=VAR,CIlow=CIlow,CIhigh=CIhigh,
estList=mat.est))

}
```

In order to apply the above bootstrap, we must create a function that inputs the data and outputs the vector estimate. Here is such a function of the IPTW estimator described in this manuscript for both the risk ratio and odds ratio.

Box A2.2: Function for IPTW

```
iptw_RF<-function(DAT){
#define regimen W as: stay on warfarin
#define regimen D as: stay on DOAC unless RF, then warfarin
#patients following reg W at first time point
RW0<-DAT$A0==0
#patients following reg W at first AND second time points
RW01<-(DAT$A0==0&DAT$A1==0)
#patients following reg D at first time point
RD0<-DAT$A0==1
#patients following reg D at first AND second time points
RD01<-(DAT$A0==1&DAT$A1==1) | (DAT$A0==1&DAT$RF==1)

A0mod<-glm(A0~W,family=binomial(),data=DAT)
A1mod<-glm(A1~A0+W,family=binomial(),subset=(RF==0),data=DAT)

P1<-predict(A0mod,type="response");  P0<-1-P1
P00<-((1-predict(A1mod,
 newdata=as.data.frame(cbind(W=DAT$W,A0=0)),type="response"))*
 (DAT$RF==0)+1*(DAT$RF==1))*P0
P11<-((DAT$RF==0)*
 predict(A1mod,newdata=as.data.frame(cbind(W=DAT$W,A0=1)),
 type="response") + (DAT$RF==1))*P1

w<-(1/P00*RW01 + 1/P11*RD01)

sub<-(RD01==1|RW01==1)

#5% weight truncation
#w[w>quantile(w[sub],0.05)]<-quantile(w[sub],0.05)

#Option 1: RISK DIFFERENCE
IPTWmod<-lm(Y~RD01,weights=w,subset=sub,data=DAT)
est_RD<-IPTWmod$coef[2]

#Option 2: RISK RATIO
IPTWmod<-glm(Y~RD01,weights=w,subset=sub,
 family=quasibinomial(link=log),data=DAT)
est_RR<-exp(IPTWmod$coef[2])

#Option 3: ODDS RATIO
IPTWmod<-
glm(Y~RD01,weights=w,subset=sub,family=quasibinomial(),data=DAT)
est_OR<-exp(IPTWmod$coef[2])

return(c(est_RD,est_RR,est_OR)) #output all three estimates
}
```

The next box shows how to use the above functions to obtain bootstrapped estimates of the standard errors of the risk difference, risk ratio, and odds ratio given data O=(W,A0,RF,A1,Y) as defined previously.

Box A2.3: Using the bootstrap and IPTW functions

```
#Both functions must first be run
#Save the data as a dataframe
DAT<-as.data.frame(cbind(W,A0,RF,A1,Y))

#K is the number of bootstrap resamples
bsres<-bootstrap.vect(iptw_RF,DAT,K=500)

#Estimated standard error for each parameter estimate
sqrt(bsres$VAR)

#Lower-limit of the 95% confidence interval for each estimate
bsres$CIlow

#Upper-limit of the 95% confidence interval for each estimate
bsres$CIhigh

#List of bootstrapped estimates (vector of size K)
#Can use to plot histogram and/or identify failed runs
Bsres$estList
```

A3. SIMULATION STUDY RESULTS

We compare the performance of the four adherence analyses (M1-M4) with the performance of the IPW application contrasting the two treatment strategies. Recall that, in our simulated data with the data-generating functions given in Box 2, there is no true difference between the effects of DOACs and warfarin on bleeding (i.e. no arrows leading from either exposure node to the outcome, Y). Thus, a correct result from this analysis would result in null estimates (OR=1).

We independently sample 1,000 datasets of size n=25,000 and run the models M1-M4 and IPW for treatment strategy on each of the 1,000 datasets. The parameters estimated by the first four methods are the exponential of the coefficients in the models M1-M4, i.e. the conditional odds ratios related to A_0 and A_1, respectively. The parameter estimated by the IPW method contrasting strategies R_D and R_W is the marginal odds ratio $(P[Y(R_D)=1]/\{1- P[Y(R_D)=1]\})/(P[Y(R_W)=1]/\{1- P[Y(R_W)=1]\})$. We then save the point estimates from each model run on each dataset and also note whether the resulting confidence interval includes the null. If not, this is considered a false positive, incorrectly concluding that there is an effect of anticoagulant choice on bleeding.

Parameter	**M1. Adjust for W**	**M2. Subset on RF**	**M3. Subset on RF with weights**	**M4. Adjust for RF**	**IPW for strategy**
Scenario 1					
OR of					
A_0	1.07 (31.9)	1.00 (4.2)	1.00 (5.4)	1.00 (4.8)	
A_1	0.90 (61.0)	1.00 (4.6)	1.00 (5.2)	1.00 (4.6)	
R_D vs R_W					1.00 (5.4)
Scenario 2					
OR of					
A_0	1.11 (69.4)	1.08 (29.2)	1.07 (27.7)	1.02 (7.7)	
A_1	0.82 (98.3)	0.88 (63.4)	0.89 (61.8)	0.94 (20.3)	
R_D vs R_W					1.00 (6.6)
Scenario 3					
OR of					
A_0	1.61 (100.0)	0.82 (95.4)	0.82 (97.6)	0.81 (99.7)	
A_1	0.41 (100.0)	1.00 (3.6)	1.01 (5.7)	1.00 (3.6)	
R_D vs R_W					1.00 (5.0)

Web Table 2. Mean of the Point Estimates and % False Positives for Regression Methods M1-M4 and the IPW Treatment Strategy Approach over 1,000 Simulated Draws.

Web Table 2 presents the mean of the odds ratio point estimates and the % of false positives over the 1,000 simulated datasets in each scenario. For scenario 1, there is a small average bias for both exposure effects, leading to 32% and 61% false positives for A_0 and A_1 respectively. For the models M2-M4 there is no bias with roughly 5% false positives in each case as desired. There is similarly no bias for the IPW approach and 5% false positives, indicating that the standard errors were well-estimated by bootstrap. For scenario 2, we see that for all regression models, there is an upwards mean bias in the odds ratio estimates of A_0 and a downwards bias for A_1. The biases are particularly large for M1, which ignored the contraindication entirely, and lowest for M4, which adjusted for RF in the regression model. The high level of false positives in all scenarios suggests that an investigator is likely to conclude that there is indeed an effect. The IPW method had no bias on average and 6.6% type 1 error. In scenario 3, where RF is a collider, models M1-M4 produced the largest average bias overall resulting in very high levels of type 1 errors. In particular, the direction of bias of the effect of A_0 for models M2-M4 was reversed compared to the previous scenario, suggesting a protective effect of early exposure. In M2-M4, there was no average bias in the estimation of the effect of A_1, which makes sense as these models correctly adjusted for the confounder RF of the relationship A_1 – Y. The IPW estimator was unbiased on average with optimal type 1 error.

In order to evaluate the IPW estimator for a true parameter off the null, we repeated the same simulation as scenario 3, except that we changed the outcome generation to

```
Y0<-rbinom(size=1,n=ssize,p=plogis(-2+0.5*W+2*U+0.2*A0+0.4*A1))
```

so that both early and later treatments had positive effects on the probability of an outcome. The results are in Web Table 3.

Parameter	M1. Adjust for W	M2. Subset on RF	M3. Subset on RF with weights	M4. Adjust for RF	IPW for strategy
OR of					
A_0	1.78 (1.00)	0.96 (0.15)	0.94 (0.23)	0.93 (0.35)	
A_1	0.56 (1.00)	1.31 (1.00)	1.32 (1.00)	1.31 (1.00)	
R_D vs R_W					1.38 (1.00)

Web Table 3. Mean of the Point Estimates and % True Positives (power) for Regression Methods M1-M4 and the IPW Treatment Strategy Approach over 1,000 Simulated Draws. In this scenario, both early and later exposures have small positive effects (conditional OR=1.22 and 1.49, resp.). True effect of treatment strategy OR = 1.38.

A4. CODE FOR STABILIZED WEIGHTS

Better finite-sample properties can be obtained by computing stabilized weights that include a numerator probability that does not adjust for W. Weight stabilization has a negligible impact on the current finite sample simulation results.

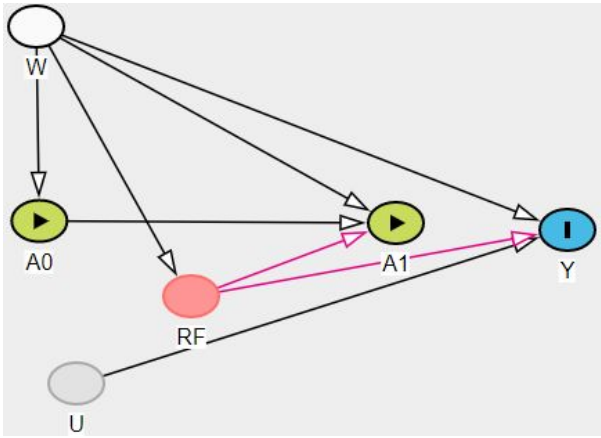
Box A4.1: Stabilized weights for IPW for treatment strategy

```
#denominator (same as before)
A0mod<-glm(A0~W,family=binomial())
A1mod<-glm(A1~A0+W,family=binomial(),subset=(RF==0))

P1<-predict(A0mod,type="response")
P0<-1-P1
P00<-((1-predict(A1mod,
newdata=as.data.frame(cbind(W,A0=0)),type="response"))*(RF==0)+1*
(RF==1))*P0
P11<-
((RF==0)*predict(A1mod,newdata=as.data.frame(cbind(W,A0=1)),type=
"response")+ (RF==1))*P1

#numerator
P1n<-mean(A0)
P0n<-1-P1n
P00n<-((1-mean(A1[RF==0&A0==0]))*(RF==0)+1*(RF==1))*P0n
P11n<-((RF==0)*mean(A1[RF==0&A0==1])+ (RF==1))*P1n

#stabilized weights
ws<-(P00n/P00*RW01 + P11n/P11*RD01)
```

A5. DAGITTY CODE FOR THE DAGS PRESENTED IN FIGURE 2

An anonymous reviewer generously provided DAGitty (http://www.dagitty.net/dags.html) code to reproduce the DAGs presented in Figure 2, which we relay to the reader.

```
dag {
bb="0,0,1,1"
A0
[exposure,pos="0.170,0.488"]
A1
[exposure,pos="0.405,0.490"]
RF [pos="0.270,0.566"]
U [latent,pos="0.194,0.655"]
W [adjusted,pos="0.168,0.288"]
Y [outcome,pos="0.518,0.497"]
A0 -> A1
RF -> A1
RF -> Y
U -> Y
W -> A0
W -> A1
W -> RF
W -> Y
}
```




```
dag {
bb="0,0,1,1"
A0
[exposure,pos="0.170,0.488"]
A1
[exposure,pos="0.405,0.490"]
RF [pos="0.270,0.566"]
U [latent,pos="0.194,0.655"]
W [adjusted,pos="0.168,0.288"]
Y [outcome,pos="0.518,0.497"]
A0 -> A1
A0 -> RF
RF -> A1
U -> RF
U -> Y
W -> A0
W -> A1
W -> RF
W -> Y
}
```