The Basics of Propensity Scoring and Marginal Structural Models

Cynthia S. Crowson,
Louis A. Schenck, Abigail B. Green,
Elizabeth J. Atkinson, Terry M. Therneau

Technical Report #84
August 1, 2013

Department of Health Sciences Research
Mayo Clinic
Rochester, Minnesota

Contents

1 Introduction

One of the common questions in medical research is: does a variable x influence a particular outcome y? (Does smoking cause lung cancer? Will this treatment improve disease outcomes? Will this medication increase the risk of cardiovascular disease?) This simple question is often very difficult to answer due to the presence of other confounding factors that are related to the factor of interest x and also affect the outcome. For example, factors that indicate a patient is sicker may predict both that a patient may be more likely to receive a particular treatment (x), and that a patient is more likely to have a poor outcome (y). In this case, failing to account for confounders could produce a biased estimate of the treatment effect. Hence, adjusting for confounders is an important issue in medical research. Randomized controlled trials are one of the best methods for controlling for confounders, as they allow for perfect balance on selected important factors and random balance on all others (both known and unknown, both measured and unmeasured), since the treatment assignment is independent of the other factors. However, randomized controlled trials are not always feasible or practical. Nor are all questions amenable to a trial, e.g., does high blood pressure increase the risk of stroke, since patients cannot be assigned to one value of the predictor.

In observational studies for instance patients are not randomly assigned to treatments, and factors that influence outcome may also have influenced the treatment assignment, which is the definition of confounding. In observational studies the analysis must adjust for the confounding factors to properly estimate the influence of the factor of interest on the outcome. This is true whether the chosen predictor x is a simple yes/no variable such as treatment or a more complex physiologic measurement, though much of the literature on marginal structural models is motivated by the treatment examples. There are two major approaches to adjusting for confounders: the conditional approach and the marginal approach.

A key aspect to both approaches is the concept of a *target population*, essentially a distribution $d(A)$ over the set of confounders A. We would like to estimate the effect of the variable of interest x on the outcome y in a population that had this structure. The most common target population is the distribution of the confounders A in the study sample as a whole, followed by using some external reference population. The conditional approach first creates predictions, followed by a weighted average of predicted values over $d(A)$ for each value of x. The marginal approach first defines case weights so that each substratum of x values, when weighted, has an equivalent distribution $d(A)$ of confounders, and then forms predictions from the weighted sample.

In the conditional approach, the idea is to first form predictions $E(y|x, A)$ for each possible combination of the variable of interest x and possible values of the set of confounders (or adjustors) A. One then averages predicted values over the distribution $d(A)$ for any fixed value of x. Two common approaches for forming the predictions are stratification and modeling. In the former the data is divided into subsets based on A, often referred to as strata, and the relationship $y|x$ is assessed within each stratum. Stratification is challenging, however, for more than 1 or 2 adjustors due to the need to create subsets that are both homogeneous (to avoid confounding within the stratum) and yet large enough to separately examine the $y|x$ relationship within each of them. The more common approach has been to jointly model all the variables using both x and A as covariates. The primary limitation of this approach is that the model must be completely correct, including in particular interrelationships between x and any variables in A. Are the relationships additive, multiplicative, or otherwise in affect? Is the effect linear, non-linear but smooth, or categorical? Are there interactions? Despite the fact that this is the most commonly used method, it is also the most likely to fail. It is too easy to blithely say "the model was adjusted for age and sex" when these were simply added as linear covariates, with no verification that the age effect is actually linear or that it is similar for the two sexes. More worrisome is

the inevitable overfitting that occurs when the set of potential confounders A is large, particularly if selection processes such as stepwise regression are used. More damming is that it can be shown that for time dependent treatment effects modeling may not give the correct result no matter how large the sample size nor how sophisticated the modeler. For example if A influences both treatment and an intermediate outcome y as in figure 1, and then both y and A influence subsequent treatment cycles and outcomes. Bias can be large in this case, with both under or over estimation of the actual treatment effect possible.

Figure 1: Directed acyclic graph depicting a confounding factor (A) that effects both the treatment (T) and the outcome (Y).[18]

The marginal approach is based on the fact that if a sample is balanced with respect to potential confounders, then the estimated effects for treatment will be unbiased, even if confounders are not modeled correctly (or modeled at all). The idea then is to balance the study population for the confounding factors, within each level of x, and produce overall estimates of $E(y|x)$ using the weighted sample. There are two main methods for creating a balanced sample: matched selection and re-weighting. Examples of matched selection include randomized controlled trials and matched case-control studies. Re-weighting has historical roots in survey sampling, where samples may be drawn disproportionately from various subpopulations and are later re-weighted to represent the entire populations. The key idea of re-weighting is to create case weights such that the re-weighted data is balanced on the factor of interest (e.g., treatment assignment) as it would have been in a randomized controlled trial. As with the conditional approach, weights can be estimated either by categorizing into homogeneous subsets of the confounders A or by fitting overall models to esimate probabilities. Limitations of marginal methods are that you can only balance on known factors, the number of balancing variables is limited, and there is a possibility that some patients may have large weights (i.e., a few individuals may represent a large part of the weighted sample). An advantage is that it is possible to balance on factors that influence both the treatment assignment and the outcome, whereas conditional adjustment for such factors may adjust away the treatment effect. In practice an analysis may choose to match on some variables and directly model others.

1.1 An example of the two methods

As an initial example of the two main approaches, we will use data from a study of free light change (FLC) immunoglobulin levels and survival [5]. In 1990 Dr. Robert Kyle undertook a population based study, and collected serum samples on 19,261 of the 24,539 residents of Olmsted County, Minnesota, aged 50 years or more [10]. In 2010 Dr. A. Dispenzieri assayed a subfraction of the immunoglobulin, the free light chain (FLC), on 15,748 samples which had sufficient remaining material to perform the test. A random sample of 1/2 the cases is included in the R survival package as the "flcdata" data set.

In the original analysis of Dispenzieri the subjects were divided into 10 groups based on their total free light chain. For simpler exposition we will divide them into 3 groups consisting of 1: those below the 70th percentile of the original data, 2: those between the 70th and 90th percentile, and 3: those at the 90th or above. The 3 dashed lines in both panels of Figure 2 show the overall survival (Kaplan-Meir) for these cohorts. High FLC is associated with worse survival, particularly for the highest cohort. Average free light chain amounts rise with age, however, in part because it is eliminated through the kidneys and renal function declines with age. Table 1 shows the FLC by age distribution. In the highest decile of FLC (group 3) over half the subjects are age 70 or older compared to only 23% in those below the

70th percentile. The survival distributions of the 3 FLC groups are clearly confounded by age, and to fairly compare the survival of the 3 FLC groups, we need to adjust for age.

The conditional approach to adjustment starts by estimating the survival for every age/sex/FLC combination, and then averaging the estimates. One way to obtain the estimates is by using a Cox model. To allow for non-proportional effects of FLC it was entered as a strata in the model, with age and sex as linear covariates. The assumption of a completely linear age effect is always questionable, but model checking showed that the fit was surprisingly good for this age range and population. Predicted survival curves were then produced from the model for three scenarios: a fictional population with the age/sex distribution of the overall population but with everyone in FLC strata 1, a second with everyone in FLC group 2 and a third with everyone in FLC strata 3. These three curves are shown with solid lines in the panel on the left side of Figure 2.

The marginal approach will first reweight the patients so that all three FLC groups have a similar age and sex distribution. Then ordinary Kaplan-Meier curves are computed for each FLC group using the reweighted population. The solid lines in the right panel of Figure 2 show the results of this process. The new weights can be based on either logistic regression models or tabulating the population by age/sex/FLC groups. (We will use the latter since it provides example data for a following discussion about different weighting ideas.) When dividing into subsets one want to use small enough groups so that each is relatively homogeneous with respect to age and sex, but large enough that there is sufficient sample in each to have stable counts. We decided to use 8 age groups (50-54, 55-59, . . . , 75-79, 80-89, 90+), giving 48 age by sex by FLC subsets in all. All observations in a given group are given the same sampling weight, with the weights chosen so that the weighted age/sex distribution within each FLC stratum is identical to the age/sex distribution for the sample as a whole. That is, the same population target as was used in the conditional approach. The mechanics of setting the weights is discussed more fully in section 3.

In both methods correction for age and sex has accounted for a bit over 1/2 the original distance between the survival curves. The absolute predictions from the two methods are not the same, which is to be expected. A primary reason is because of different modeling targets. The marginal approach is based on the relationship of age and sex to FLC, essentially a model with FLC group as the outcome. The conditional approach is based on a model with survival as the outcome. Also, one of the approaches used continuous age and the other a categorical version. In this case the relationships between age/sex and FLC group and that between age/sex and survival are fairly simple, both approaches are successful, and the results are similar.

2 Propensity scoring

A common approach to dealing with multiple confounding factors that affect both the treatment assignment and the outcome of interest is propensity scoring [19]. A propensity score is the probability (or propensity) for having the factor of interest (e.g., receiving a particular treatment) given the factors present at baseline. A true propensity score is a balancing score; the set of all subjects with the same probability of treatment will have identical distributions of baseline factors among those who were treated and those who were not treated (i.e., all factors will be perfectly balanced at all levels of the propensity score) [1]. Typically the true propensity score is unknown, but it can be estimated using study data and common modeling techniques. For example, logistic regression is commonly used with a binary outcome, and the resulting multivariable logistic regression model can be used to obtain predicted probabilities (i.e., propensity score values) for both the probability of being treated, p, and the probability of not being treated $1 - p$. When there are more than 2 groups, as in the FLC example, ei-

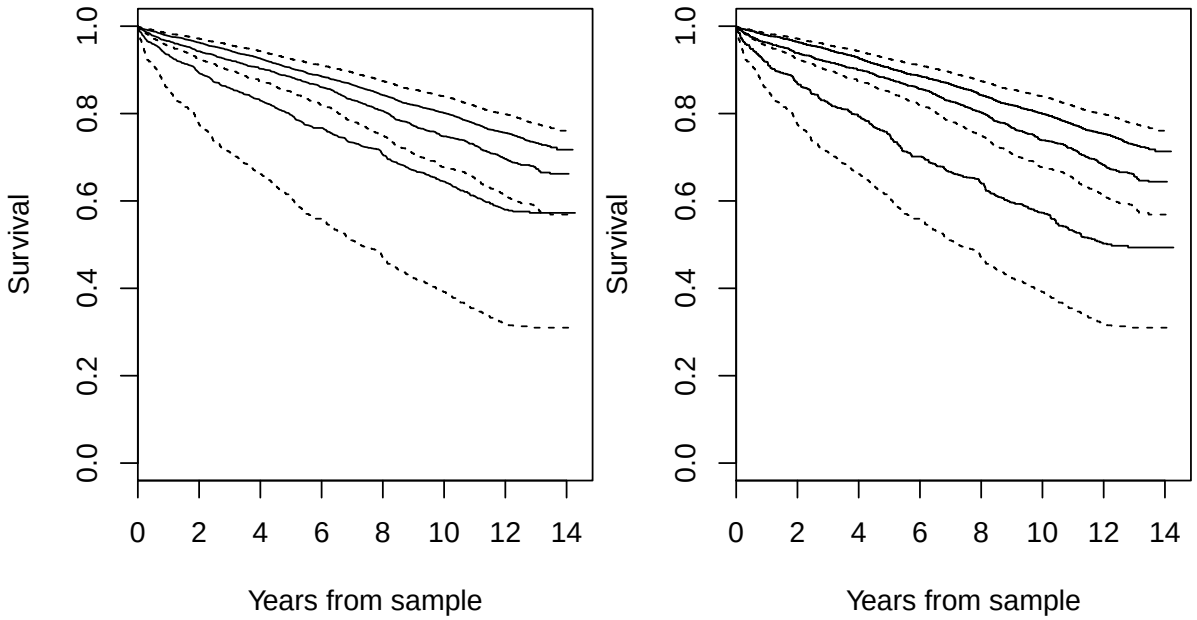

Figure 2: Survival of 15,748 residents of Olmsted County, broken into three cohorts based on FLC value. The dashed lines in each panel show the original survival curves for each group. The solid lines in the left panel were obtained by modeling to adjust for age and sex, and the solid lines in the right panel were obtained by using weighting methods.

ther a model with multinomial outcome or multiple logistic regressions can be used.

2.1 Variable selection

Deciding which factors should be included in the propensity score can be difficult. This issue has been widely debated in the literature. Four possible sets of factors to include are: all measured baseline factors, all baseline factors associated with the treatment assignment, all factors related to the outcome (i.e., potential confounders), and all factors that affect both the treatment assignment and the outcome (i.e., the true confounders) [1]. In practice it can be difficult to classify the measured factors into these 4 categories of potential predictors. For example, it is often difficult to know which factors affect only the treatment assignment, but not the outcome. Some have suggested that including all the available factors, even if the model is overfit, is acceptable when building a propensity score, as prediction overrides parsimony in this case. However, others have shown this approach is flawed, as overfitting the propensity score model can lead to a wider variance in the estimated scores, in particular an overabundance of scores near 0 or 1, which in turn results in a larger variance of the estimated treatment effect. Similarly, inclusion of the factors that only affect the treatment assignment but do not affect the outcome also results in more extreme scores.

The most informative observations in a weighted sample are those with scores close to 1/2, leading to overlap between the set of propensity scores for different values of the primary variable if interest x. The issue of overlap is examined in Figure 3, which shows three scenarios for hypothetical studies comparing the probability of no treatment and treatment. In the left panel, there is a small range of probabilities (35-50%) where patients can receive either treatment. In the middle panel, there is no overlap in probabilities between those who do and do not receive treatment. This represents complete confounding, which can occur when there are treatment guidelines dictating which patients will and

will not be treated. In this case, it is impossible to statistically adjust for the confounding factors and determine the effect of the treatment itself. In the right panel, there is good overlap between the untreated and treated groups. This is the ideal scenario for being able to separate the impact of confounders from that of the treatment.

Returning to the issue of variable selection, the recommended approach is to include only the potential confounders and the true confounders in the propensity score [1]. This will result in an imbalance between the treated and untreated group in the factors that influence only the treatment assignment but not the outcome. However, since these factors do not affect the outcome, there is no advantage to balancing them between the treatment groups. In practice, most factors are likely to affect both the treatment assignment and the outcome, so it may be safe to include all available baseline factors of interest. The exceptions that will require further thought are policy-related factors and time-related factors. For example, treatment patterns may change over time as new treatments are introduced, but the outcome may not change over time. In that case, including a time factor in the propensity score would unnecessarily result in less overlap. It's also important to note that only baseline factors can be included in the propensity score, as factors measured after the treatment starts may be influenced by the treatment.

2.2 Balance

The goal of propensity scoring is to balance the treated and untreated groups on the confounding factors that affect both the treatment assignment and the outcome. Thus it is important to verify that treated and untreated patients with similar propensity score values are balanced on the factors included in the propensity score. Demonstrating that the propensity score achieves balance is more important than showing that the propensity score model has good discrimination (e.g., the c-statistic or area under the receiver operating characteristic curve).

Balance means that the distribution of the factors is the same for treated and untreated patients with the same propensity score. There are a number of ways to assess and demonstrate balance using either matching, stratification or inverse probability treatment weighting (which will be discussed in the next section). In the matching approach, each treated patient is matched to an untreated patient with the same propensity score. In the stratification approach, the patients are divided into groups based on quantiles of propensity scores (e.g., often quintiles of propensity scores are used). Then the differences in covariates between matched pairs of patients or within each strata are examined. If important imbalances are found, the propensity score should be modified by including additional factors, interactions between factors of interest, or non-linear effects for continuous factors. Thus developing a propensity score is an iterative process.

Of note, tests of statistical significance are not the best approach to examining balance, instead the magnitude of differences between the treated and untreated patients with similar propensity scores should be examined. P-values from significance tests are influenced by sample size and in the case of matching, sometimes trivial differences will yield high statistical significance. The recommended approach is to examine standardized differences (defined as the difference in treated and untreated means for each factor divided by the pooled standard deviation). The threshold used to demonstrate balance is not well defined, but it has been suggested that standardized differences <0.1 are sufficient [1].

2.3 Using the propensity score

Once the propensity score has been developed and balance has been shown, several different approaches have been used to examine the question of interest (i.e., does the factor of interest influence the outcome after accounting for confounders?). In fact, both conditional approaches (i.e., stratification, model adjustment) and marginal approaches (i.e., matching and re-weighting) have been used in conjunction with propensity scores as methods to account for confounders.

As mentioned previously, each adjustment method has limitations, which are still applicable

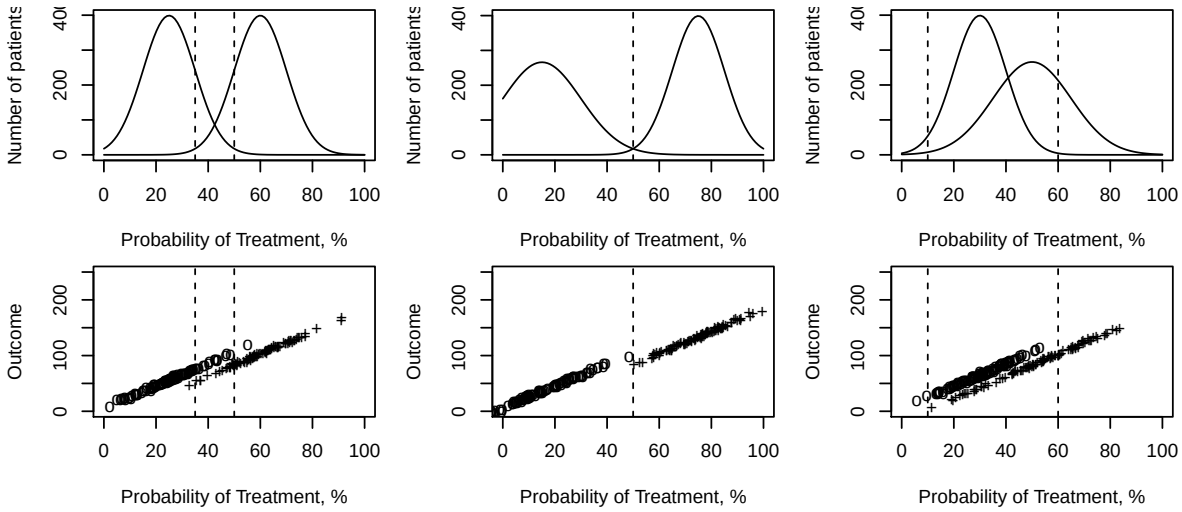

Figure 3: Three scenarios for hypothetical studies comparing the probability of treatment for patients who were untreated (left peak) with those who were treated (right peak) with a small range of probabilities where patients can receive either treatment (left column), no overlap in probabilities between those who do and do not receive treatment (middle column) and good overlap between the untreated and treated groups (right column).

when adjusting for propensity scores. Many reports comparing various methods have been published [1]. Stratification can result in estimates of average treatment effect with greater bias than some of the other methods [14]. Using the propensity score as an adjustor may not adequately separate the effect of the confounders from the treatment effect. Matching on the propensity score often omits a significant portion of the cohort for whom no matches are possible, in the case where people with certain values always or never receive a particular treatment. These issues are discussed further by Austin [1] and Kurth [11].

3 Inverse probability weighting

Inverse probability weighting (IPW) is a method where data is weighted to balance the representation of subgroups within the full data set. In IPW, each observation is weighted by the reciprocal (i.e., the inverse) of the predicted probability of being in the group that was observed for each patient.

	50–59	60–69	70–79	80+	Total
Group 1	2592 (47)	1693 (30)	972 (17)	317 (6)	5574
Group 2	444 (29)	448 (29)	425 (28)	216 (14)	1533
Group 3	121 (16)	188 (25)	226 (29)	232 (30)	767
Total	3157	2329	1623	765	7874

Table 1: Comparison of the age distributions for each of the three groups, along with the row percentages.

This method is commonly used in survey sampling [9]. The important issues to consider when assigning weights are whether balance is achieved, what is the population of interest, and how big are the weights.

We will return to the FLC example to demonstrate the importance of these issues (see Table 1). Restating the three aims for this data set, weights should be chosen so that

1. The weighted age/sex distribution is identical for each FLC group

2. The overall weighted age/sex distribution matches that of the original population, subject to the prior constraint.

3. Individual weights are "as close as possible" to 1, subject to the prior 2 constraints.

Point 1 is important for unbiasedness, point 2 for ensuring that comparisons are relevant, and point 3 for minimizing the variance of any contrasts and reducing potentially undue influence for a small number of observations. For simplicity we will illustrate weights using only age, dividing it into the three coarse groupings shown in table 1.

First we try assigning IPW based on the overall probablilty $w_{ij} = 1/P(\text{age} = i, \text{FLC} = j)$. For the 50–59 age group and the first FLC stratum the probability is 2592/7874, the total count in that cell divided by n, and the weight is 7874/2592= 3.

The results are shown in table 2, with weights shown in the upper half of the table and the new, reweighted table of counts in the lower portion. This achieves balance (trivially) as the reweighted are the same size for each age/FLC group. However, the weighted sample no longer reflects the population of interest, as now each age group is equally weighted, whereas the actual Olmsted County population has far more 50 year olds than 80 year olds. This approach gives the correct answer to a question nobody asked: "what would be the effect of FLC in a population where all ages were evenly distributed", that is, for a world that doesn't exist. In addition, the individual weights are both large and highly variable.

	Age Group			
	50–59	60–69	70–79	80+
	Weights			
FLC 1	3.0	4.7	8.1	24.8
FLC 2	17.7	17.6	18.5	36.5
FLC 3	65.1	41.9	34.8	33.9
	Reweighted Count			
FLC 1	7874	7874	7874	7874
FLC 2	7874	7874	7874	7874
FLC 3	7874	7874	7874	7874

Table 2: Inverse probability weights based on the total sample size. The upper panel shows the weights and the lower panel shows the table of weighted counts. Counts for the three FLC groups are balanced with respect to age, but the overall age distribution no longer matches the population and individual weights are far from 1.

In our second attempt at creating weights, we weight the groups conditional on the sample size of each age group, i.e., $w_{ij} = 1/P(\text{FLC=j} \mid \text{age =i})$; The probability value for a 50–59 year old in the first FLC group is now 2592/3157. Results are shown in Table 3. This method still achieves balance within the FLC groups, and it also maintains the relative proportions of patients in each age group — note that the reweighted counts are the column totals of the original table 1. This is the method used for the curves in the right panel of figure 2, but based on 16 age/sex strata. The weighted sample reflects the age composition of a population of interest. However, the weights are still quite large and variable. There is a natural connection between these weights and logistic regression models: define $y = 1$ if a subject is in FLC group 1 and 0 otherwise; then a logistic regression of y on age and sex is an estimate of $P(\text{FLC group =1} \mid \text{age and sex})$, the exact value needed to define weights for all the subjects in FLC group 1.

	Age Group			
	50–59	60–69	70–79	80+
	Weights			
FLC 1	1.2	1.4	1.7	2.4
FLC 2	7.1	5.2	3.8	3.5
FLC 3	26.1	12.4	7.2	3.3
	Reweighted Count			
FLC 1	3157	2329	1623	765
FLC 2	3157	2329	1623	765
FLC 3	3157	2329	1623	765

Table 3: Inverse probability weights normalized separately for each age group. The upper panel shows the weights and the lower panel shows the total count of reweighted observations. Each FLC group is balanced on age and the overall age distribution of the sample is retained.

The third set of weights retains both balancing and the population distribution while at the same time creating an average weight near 1, by balancing on *both* margins. These are defined as wt = P(FLC=i) P(age=j) / P(FLC=i and age=j). The formula is familiar: it is the reciprocal of the observed:expected ratio from a Chi-square test (i.e., E/O). The resultant weights for our simple example are shown in Table 4. These weights are often referred to as "stabilized" in the MSM literature.

In practice, there will be multiple confounding factors, not just age, so modeling will be needed to determine the IPW. As mentioned in the previous section, propensity scores are designed to balance the treated and untreated groups on the factors that confound the effect of treatment on an outcome, and one way to use a propensity score to account for confounding in the model of the outcome is to use IPW. In IPW, each observation is weighted by the reciprocal (i.e., the inverse) of the predicted probability of receiving the treatment that was observed for each patient, which can be estimated using propensity scoring. Note that all the predicted probabilities obtained from the propensity model, p, will be the probabilities of receiving treatment. The reciprocal of these values, $1/p$ will be the inverse probability weights for patients who were treated. If there are multiple groups a separate model can be fit with each group taking its turn as 'y' =1 with the others as 0 to get the probabilities for observations in that group, or a single generalized logistic regression fit which allows for a multinomial outcome. (If there are only two groups, the common case, only one logistic regression is needed since the fit predicts both the probability p of being in group 1 and that of "not group 1" $= 1 - p =$ group 2.)

	Age Group			
	50–59	60–69	70–79	80+
	Weights			
FLC 1	0.9	1.1	1.3	1.9
FLC 2	1.5	1.1	0.8	0.8
FLC 3	2.8	1.3	0.8	0.4
	Reweighted Count			
FLC 1	2455	1811.1	1262.1	594.9
FLC 2	675.2	498.1	347.1	163.6
FLC 3	337.8	249.2	173.7	81.9

Table 4: Stabilized inverse probability weights, which achieve balance between the FLC groups, maintain the proportion of patients in each age group to reflect the population of interest, and also maintain the original sample size in the weighted sample to avoid inflated variance issues. The upper panel shows the weights and the lower panel shows the sum of the weights.

With the appropriate weights, the weighted study population will be balanced across the treatment groups on the confounding factors. This balance is what allows for unbiased estimates of treatment effects in randomized controlled trials, thus IPW can be thought of as simulating randomization in observational studies.

One problem with IPW is that the weights can have a large standard deviation and a large range in practice. IPW often produces an effective sample size of the weighted data that is inflated compared

to the original sample size, and this leads to a tendency to reject the null hypotheses too frequently [24]. Stabilized weights (SWs) can be used to reduce the Type 1 error by preserving the original sample size in the weighted data sets. The purpose of stabilizing the weights is to reduce extreme weights (e.g., treated subjects with low probability of receiving treatment or untreated patients with high probability of receiving treatment). The stabilization is achieved by the inclusion of a numerator in the IPW weights. When defining weights based on propensity scoring (i.e., using only baseline covariates to determine the probability of treatment), the numerator of the stabilized weights is the overall probability of being treated for those who were treated and of not being treated for those who were not treated [24]. So the formula for SW is then P/p for the treated and $(1-P)/(1-p)$ for the untreated, where P is the overall probability of being treated and p is the risk factor specific probability of treatment obtained from the propensity score for each patient.

It is also important to verify that the mean of the SW is close to 1. If the mean is not close to 1, this can indicate a violation of some of the model assumptions (which will be discussed in a later section), or a misspecification of the weight models [20].

Because extreme weights can occur in IPW and such weights have an untoward influence on the results, weight truncation is commonly used with IPW and SW. Various rules for truncation have been applied. One common approach is to reset the weights of observations with weights below the 1st percentile of all weights to the value of the 1st percentile and to reset the weights above the 99th percentile of all weights to the value of the 99th percentile. Of course, other cutoffs, such as the 5th and 95th percentiles can also be used. Another approach is to reduce the weights that are >20% of the original sample size to some smaller value, and then adjust all the weights to ensure they sum to the original sample size. There is a bias-variance trade off associated with weight truncation, as it will result in reduced variability and increased bias.

4 Marginal structural models

Up to this point, we have primarily focused on adjustment for baseline confounding factors. Marginal structural models (MSMs) are a class of models that were developed to account for time-varying confounders when examining the effect of a time-dependent exposure (e.g., treatment) on a long-term outcome in the presence of censoring. Most of the common methods of adjustment can be difficult or impossible in a problem this complex. For example, in patients with human immunodeficiency virus (HIV), CD4 counts are monitored regularly and are used to guide treatment decisions, but are also the key measure of the current stage of a patient's disease. Not adjusting for CD4 leads to invalid treatment comparisons; an aggressive and more toxic treatment, for instance, may be preferentially applied only to the sickest patients. However, simply adjusting for CD4 cannot disentangle cause and effect in a patient history containing multiple treatment decisions and CD4 levels. MSMs were developed to tackle this difficult problem using IPW methods to balance the treatment groups at each point in time during the follow-up. This approach is "marginal" because the patient population is first re-weighted to balance on potential confounders before estimating the treatment effect. These models involve development of time-varying propensity scores, as well as methods to account for imbalances due to censoring patterns.

An early variant of these models was developed by Dr. Marian Pugh in her 1993 dissertation [16], focused on the problem of adjusting for missing data. The general approach was first published by Drs. James Robins and Miguel Hernán from Harvard in 1999 [7][17]. A Harvard website provides a SAS macro, %msm, for computing these models [13]. The SAS macro uses an inefficient method to compute the Cox models for the time-varying propensity scores and for the outcome of interest, because many standard Cox model software pro-

grams do not allow for subject-specific time-varying weights [7]. To circumvent this software limitation, they use the methods of Laird and Olivier [12] and Whitehead [21] to fit Cox models using pooled logistic regression models (see our example demonstrating equivalence of these methods in Appendix A). This method requires a data set with multiple observations per subject corresponding to units of time (e.g., months). Logistic regression models are fit using this data set with time as a class variable to allow a separate intercept for each time, which mimics the baseline hazard in a Cox model. The macro call for the %msm macro is quite complex because the macro fits several models that all work together to make the resulting MSM estimations. These models include:

1. Two separate models for the numerators and denominators of the stabilized case weights, which are logistic regression models of the probability of receiving the treatment (or of having the factor of interest that you want to balance on). As previously mentioned, the denominator of the SW is often obtained using propensity scoring and the numerator is often just the overall probability of treatment. In MSMs there are time-varying propensity scores which are fit using both baseline and time-varying factors. The numerator is typically obtained from a model including only the baseline factors. This is similar to stratified randomization, which is often used to prevent imbalance between the treatment groups in key factors. Because the numerator and denominator models share common factors, they should be correlated, which should result in a weight that is less variable than using only the denominators. This approach is particularly useful when large weights are a problem.

   So for these models, the baseline or non-varying factors influencing the treatment assignment (e.g. past treatment history) are included in the numerator model and both the baseline and the time-varying factors influencing the treatment decision are included in the denominator model. The resulting predicted probabilities obtained from the numerator and denominator models are used to construct the SW for each subject at each time point during follow-up.

   Once the patent initiates the treatment, the rest of his/her observations are excluded from these models, and his/her weights do not change (i.e., the probability of initiating treatment does not change once the patient is actually on the treatment). The model assumes the patients are on (or are exposed to) the treatment from the time of initiation of the treatment until the last follow-up.

2. Censoring models are also available in the %msm macro. There are two logistic regression models for the numerator and denominator of SW for censoring. The macro allows up to 4 different sets of censoring models to model different types of censoring. For these models, the binary outcome is censoring of a particular type. The 2 most common types of censoring are administrative censoring and lost-to-follow-up. Administrative censoring occurs when subjects have complete follow-up to the last available viewing date (e.g., today or the last possible study visit or the day the chart was abstracted). This type of censoring is very common in Rochester Epidemiology Project studies. Lost-to-follow-up censoring occurs when a subject is lost before the end of the study. Many factors may influence this type of censoring, as perhaps the subject has not returned due to an unrecorded death, or perhaps they are now feeling well and decided the study was no longer worth participating in. If patients who are censored differ from those who are still being followed, the distribution of risk factors of interest will change over time, which will introduce imbalance with respect to the risk factors. The models of censoring help to adjust for administrative or lost-to-follow-up censoring by assigning high weights to patients with a high probability of censoring who were not actually censored, so they will represent the patients

who were censored. This is similar to the "redistribute to the right" weighting that occurs in the Kaplan-Meier method of estimating the probability of events occurring over time.

3. The final model is fit using a pooled logistic regression model of the outcome of interest (e.g., death) incorporating time-varying case weights using the SW computed from the previous models. As with the other models, this model is computed using the full data set including an observation for each time period for each subject. The SW used at each time point is the product of the SW from the treatment models and the SW from the censoring models.

   In addition to weighting the model by the case weights, this model can also include additional variables that may influence the outcome of interest, which were not included in the models of the treatment assignment. The question of including as adjustors the same variables that were included in the models used to establish the weights is one that has received much discussion (e.g., in the context of adjusting for the matching factors in a case-control study). If the case weights truly achieve balance, then there is no need to include them in the model of the outcome of interest. However, the price of SWs is that the weighted population may not be fully adjusted for confounding due to the baseline covariates used in the numerator of the weights, so the final model must include these covariates as adjustors [4].

4.1 MSM assumptions

There are several key assumptions inherent in MSMs: exchangeability, consistency, positivity and that the models used to estimate the weights are correctly specified [4]. The exchangeability assumption has also been referred to as the assumption of no unmeasured confounding. This is an important assumption, but unfortunately it cannot be verified empirically.

Consistency is another assumption that is difficult to verify. In this context, consistency means that the observed outcome for each patient is the causal outcome that results from each patient's set of observed risk factors. Note that this definition differs from the usual statistical definition of consistency that the bias of an estimator approaches zero as the sample size increases.

Positivity, also known as the experimental treatment assumption, requires that there are both treated and untreated patients at every level of the confounders. If there are levels of confounders where patients could not possibly be treated, such as the time period before a particular treatment existed, then this creates structural zero probabilities of treatment. Contraindications for treatment can also violate the positivity assumptions. The obvious way to deal with these violations is to exclude periods of zero treatment probability from the data set. However, if the problem occurs with a time varying confounder, exclusions can be difficult.

A related issue is random zeroes, which are zero probabilities resulting by chance usually due to small sample sizes in some covariates levels can also be problematic. Parametric models or combining small subgroups can be used to correct this problem. Weighted estimates are more sensitive to random zeroes than standard regression models. And while additional categories of confounders are thought to provide better adjustment for confounding, the resulting increase in random zeroes can increase the bias and variance of the estimate effect. Sensitivity analysis can be used to examine this bias-variance trade off. It may be advantageous to exclude weak confounders from the models to reduce the possibility of random zeroes.

5 Example using %msm macro

Our example data set used the cohort of 813 Olmsted County, Minnesota residents with incident rheumatoid arthritis (RA) in 1980-2007 identified by Dr. Sherine Gabriel as part of her NIH grant studying heart disease in RA [15]. Patients with RA have an increased risk for mortality and for heart disease. Medications used to treat RA may have beneficial or adverse effects on mortality and heart disease in patients with RA. However, the long-term outcomes of mortality and heart disease have been difficult to study in randomized clinical trials, which usually do not last long enough to observe a sufficient number of these long-term events. In addition, observational studies of the effect of medications on outcomes are confounded due to channeling bias, whereby treatment decisions are based on disease activity and severity. Therefore, MSMs might help to separate the treatment effects from the effects of disease activity and severity confounders on the outcomes of mortality and heart disease.

Methotrexate (MTX) was approved for use in treatment of RA in 1988, and is still the first line treatment for RA despite the introduction of several biologic therapies since 1999. In this cohort of 813 patients with RA, 798 have follow-up after 1988 (mean age: 57 years). RA predominately affects women in about a 3:1 female:male ratio (this cohort is 69% female). The mean follow-up was 7.9 years (range: 0 to 28.6 years) during which 466 patients were exposed to MTX and 219 patients died. The majority of censoring was for administrative reasons (i.e., they were followed through today and we need to wait for more time to pass to get more follow-up). Very few patients were lost to follow-up. Figure 4 shows the number of patients exposed to MTX and not exposed to MTX who were under observation according to months since RA diagnosis. This figure demonstrates that the proportion of patients exposed to MTX changes over time, as it was low at the start of follow-up and it was nearly 50% in the later months of follow-up.

5.1 Step 1 - Choosing time scale and defining time, event and treatment variables

The first step is to prepare the data set needed for analysis. As with any time related analysis, the first consideration is what time scale to choose for the analysis. The most common time scales are time from entry into the study or time from diagnosis. Note that the %msm macro documentation states that all subjects must start at time 0, but we could not find any methodological reason for this assertion, except that it is important to have a sufficient number of patients under observation at time 0. In these models, intercepts are estimated for each time period (compared to time 0 as the reference time period) by modeling time as a class variable.

Since our study population is an incidence cohort of patients with RA, the qualification criteria for a patient to enter this study was RA diagnosis. Thus time 0 was when RA was first diagnosed. Since the study cohort begins in 1980 and MTX was not approved for RA until 1988, patients diagnosed prior to 1988 were entered into the model in 1988. Note that for this example once the patient has been exposed to MTX they will be considered in the treatment group, even if they stop taking the drug.

Our example uses exposure to MTX (0/1 unexposed/exposed) for treatment and death (i.e., 'died') as the outcome variable. Time t (0, 1, 2, 3, ...) is the number of time periods since the patient entered the study. Other examples typically use months as the time periods for running MSMs. In our data set, this led to some computational issues. We found that each of the models fit during the MSM modeling process needs to have at least one event and at least one censor during each time period. This was also true for the reference time period (time 0), which all other time periods were compared to. In our example there were some long periods between events, so the time intervals we defined were irregular. Issues like model convergence or extreme model coefficient values often resulted from a

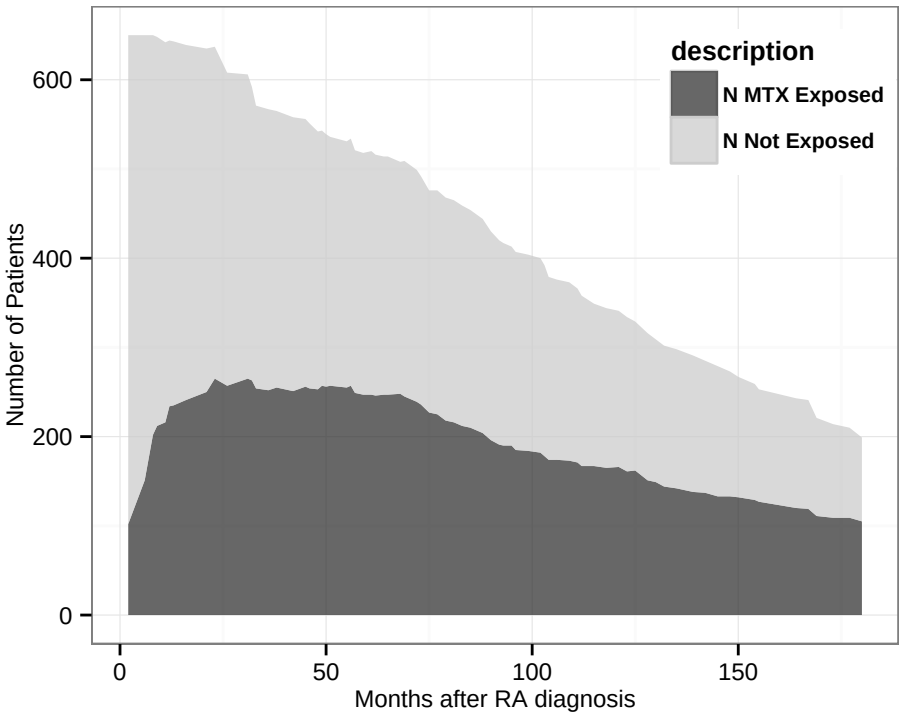

Figure 4: This figure shows the number of patients who were exposed and not exposed to methotrexate (MTX) according to months after RA diagnosis in a stacked area plot. At 50 months after RA diagnosis, approximately 250 patients were exposed to MTX and approximately 300 patients were not exposed to MTX for a total of approximately 550 patients under observation. This figure shows that the total number of patients under observation is changing over follow-up time and that the proportion of patients who were exposed to MTX changes over time, as the proportion of MTX exposed patients is low at the start of follow-up and is around 50% in the later months of follow-up.

lack of events in a time period. We found that warnings regarding "Quasi-complete separation of data points" and "maximum likelihood estimate may not exist" could be ignored since we were not interested in the accuracy of the estimated intercepts for each time period, and estimates for the coefficients of interest were largely unaffected by these warnings. We found the coefficients for each of the time periods with no events were large negative numbers on the logit scale (i.e., near zero when exponentiated) indicating no information was added to the overall model for the time periods with no events.

Because of problems with model convergence due to months without events, we used irregular periods of time that were sometimes several months long. To determine the periods of time that would work, we examined distributions of the event and censoring times. Then starting with the first month of follow-up, we grouped time periods together until we had a minimum of 1 event and 1 censor in each time period. Note that there are some other options to improve computational issues without creating irregular time periods, such as the use of cubic splines to model the baseline hazard. This would make assumptions about the functional form of the baseline hazard function, so the results would no longer match the Cox model results exactly, but it might be easier and perhaps less arbitrary than defining

irregular time periods. This is an issue for further investigation and it will not be discussed further in this report.

In addition, the patients were followed until death, migration from Olmsted County or December 31, 2008. However, this resulted in small numbers of patients under observation at later time points, so to avoid computational issues related to small sample sizes, we chose to truncate follow-up at 180 months after RA diagnosis.

At this point we created a data set with multiple observations per patient (i.e. one observation for each time period), and we can define the event variable and the treatment variable. In addition the macro requires several other time and event related variables. The time and event related variables we used in the %msm macro were:

id patient id

censored indicator for alive at the last followup date (0=died, 1=alive)

died indicator that the patient has died (1=died, 0=alive)

t time period (0, 1, 2, ...)

exposed_mtx indicator for exposure to MTX (0=untreated, 1=treated)

eligible indicator for treatment eligibility (1=eligible (prior to and in the time period where MTX exposure starts), 0=ineligible (beginning in the first time period after MTX exposure starts). The macro sets the probability of treatment (e.g., pA_d and pA_n) to 1 after exposure begins.

The programming for the data setup for the %msm macro was complicated due to the large number of time-varying factors included in the model, and the irregular time periods that were used to facilitate computations. Thus we have not provided code for constructing the full data set, but we have provided an example of the data set structure that was used with the macro. Table 5 shows a small sample of what the data set looks like so far.

5.2 Step 2 - Choosing and defining variables for treatment, censoring and final models

The next step is to choose the variables of interest for each of the models involved in the MSM process. You may want to review the section on variable selection for propensity scoring and also consult your investigator to help determine which variables may be related to both treatment assignment and outcome, as well as what factors may influence censoring in your study.

In our example, we first tried to include all patients diagnosed with RA since 1980 in our models beginning at time 0, which was RA diagnosis. However, MTX was introduced in 1988, so patients diagnosed prior to 1988 could not receive MTX in the time periods prior to 1988. At first we added an indicator variable for whether MTX had been introduced or not. This led to problems with positivity, as the probability of receiving MTX prior to its introduction was zero. This violation of an assumption of the MSM models led to extreme weights. Thus we excluded time periods prior to the introduction of MTX in 1988 for the patients who were diagnosed with RA prior to 1988.

We also started with a longer list of variables of interest, which we shortened to keep this example simple. So here is the list of baseline and time-dependent variables we will use in this example:

- baseline factors

  age Age in years at RA diagnosis

  male sex indicator (0=female, 1=male)

  rfpos indicator for rheumatoid factor positivity (0=negative, 1=positive)

  smokecb indicator for current smoker

  smokefb indicator for former smoker

  yrra calendar year of RA diagnosis

- time-varying factors

id	t	eligible	died	age	male	yrra	exposed_mtx
1	0	1	0	26	0	1988	0
1	1	1	0	26	0	1988	0
2	0	1	0	45	0	2001	0
2	1	1	0	45	0	2001	0
2	2	1	0	45	0	2001	0
2	3	1	0	45	0	2001	0
2	4	1	0	45	0	2001	0
2	5	1	0	45	0	2001	0
2	6	1	0	45	0	2001	1
2	7	1	0	45	0	2001	1
3	0	1	0	84	0	1989	0
3	1	1	0	84	0	1989	0
3	2	1	0	84	0	1989	0
3	3	1	1	84	0	1989	0
4	0	1	0	54	1	1993	1
4	1	0	0	54	1	1993	1
4	2	0	0	54	1	1993	1
4	3	0	0	54	1	1993	1
4	4	0	0	54	1	1993	1
4	5	0	0	54	1	1993	1
4	6	0	0	54	1	1993	1

Table 5: A sample of the data set structure for the rheumatoid arthritis example (see Step 1).

cld_ indicator for chronic lung disease

dlip_ indicator for dyslipidemia

dmcd_ indicator for diabetes mellitus

hxalc_ indicator for alcohol abuse

erosdest_ indicator for erosions/destructive changes

ljswell_ indicator for large joint swelling

nodul_ indicator for rheumatoid nodules

esrab indicator variable for abnormal erythrocyte sedimentation rate

exra_sev_ indicator for extra-articular manifestations of RA

hcq_ indicator for exposure to Hydroxychloroquine

othdmard_ indicator for exposure to other disease modifying anti-rheumatic medications

ster_ indicator for exposure to glucocorticosteroids

These variables need to be added to the data set described in the previous section. Note that baseline variables should have the same value for all observation for each patient. The time-dependent variables in our example data set are all absent/present variables, which are defined as 0 prior to the development of each characteristic, and they change to 1 when the characteristic develops and remain 1 for the remaining observation times. Note that continuous time-varying variables (e.g., systolic blood pressure) can also be defined, and these variables would typically change value whenever they are re-measured during follow-up. Note also that time-dependent variables should be defined at the start of each time period. Typically a new time interval would be started when the value of a time-dependent variable needs to be changed, but that may not be possible in this instance due to the computational issues mentioned in the previous section.

5.3 Step 3 - Calling the %msm macro

Once the data set is ready and the variables to use in each model have been chosen, then it is time to call the macro. Here is the macro call corresponding to our example data set:

```
%msm(
      data=ra_mtx_msm,
      id = id,
      time = t,

      /* Structural Model Settings */
      outcMSM = died,
      AMSM = exposed_mtx,
      covMSMbh = t ,
      classMSMbh= t,
      covMSMbv= age male yrra rfpos smokecb smokefb,

      /* Settings for treatement and possible censoring weight models*/
      A = exposed_mtx,
      covCd =  t age male yrra rfpos smokecb smokefb
               esrab erosdest_ exra_sev_ hcq_
               othdmard_ ster_ cld_ nodul_ ljswell_
               dmcd_ hxalc_ dlipcd_ ,
      classCd = t,
      covCn =  t age male yrra rfpos smokecb smokefb,
      classCn = t,

      covAd =  t age male yrra rfpos smokecb smokefb esrab erosdest_
               exra_sev_ hcq_ othdmard_ ster_ cld_ nodul_ ljswell_
               dmcd_ hxalc_ dlipcd_,
      classAd = t,
```

```
    covAn= t  age male yrra rfpos smokecb smokefb,
    classAn= t ,

    eligible= eligible,
    cens = censored,

    /* Data analysis settings */
    truncate_weights=1,
    msm_var=1,
    use_genmod=0,

    /* Output Settings */
    libsurv=out_exp,
    save_results = 1,
    save_weights = 1,
    debug=0,

    /* Survival curves and cumulative incidence settings */
    survival_curves = 1,
    use_natural_course=1,
    bootstrap=0,
    treatment_usen = exposed_mtx,
    treatment_used = exposed_mtx,
    time_start=0,
    time_end=120
);
```

Note that the lists of numerator variables for the treatment and censoring models include only the baseline variables. The numerators of the weights are used to stabilize the weights, and only baseline variables are included in these models by convention. See a previous section for more discussion of these issues.

Now it is time to run the macro. While the macro performs all steps of the model process at once (if it runs correctly) it is important to look at each step one-by-one, so we will examine each step individually in the next few sections.

5.4 Step 4 - Examining the treatment models

The model for the denominator of the treatment weights is arguably the most important model in the determination of the weights. This model should be examined to be sure the directions of the coefficients make sense in the medical context of the study (i.e., factors that increase the probability of receiving treatment have positive coefficients and factors that decrease the probability of receiving treatment have negative coefficients). The coefficients for the baseline variables in the model for the numerator of the weights may or may not agree closely with the coefficients of these same variables in the denominator model, depending on how the time-varying variables may have impacted the model. Table 6 summarizes the results of the treatment weight models for our example data set.

5.5 Step 5 - Examining the censoring models

The next step is to examine the censoring models. As most of the censoring in our example data set was administrative, the majority of factors in the model are unrelated to censoring, with the exception of factors indicative of time (e.g., yrra, age). Table 7 summarizes the results of the censoring weight models.

Model for Denominator of Stabilized Weights, MTX Exposure					
Parameter	DF	Estimate	Std Err	Wald Chi-Sq	Pr >ChiSq
Intercept	1	-133.9	20.0872	44.4173	<.0001
age	1	-0.0268	0.00395	46.0334	<.0001
male	1	-0.0573	0.1207	0.2256	0.6348
yrra	1	0.0658	0.0100	43.0049	<.0001
rfpos	1	0.9003	0.1314	46.9777	<.0001
smokecb	1	0.1273	0.1487	0.7326	0.3920
smokefb	1	0.1433	0.1247	1.3212	0.2504
erosdest_	1	0.8361	0.1149	52.9355	<.0001
exra_sev_	1	-0.0729	0.2388	0.0932	0.7602
hcq_	1	0.0297	0.1143	0.0676	0.7948
othdmard_	1	0.4644	0.1453	10.2169	0.0014
ster_	1	0.9093	0.1194	57.9615	<.0001
cld_	1	0.1345	0.1342	1.0043	0.3163
nodul_	1	0.2334	0.1314	3.1562	0.0756
ljswell_	1	0.3500	0.1210	8.3587	0.0038
dmcd_	1	0.2178	0.1680	1.6807	0.1948
hxalc_	1	-0.4271	0.2209	3.7364	0.0532
dlipcd_	1	0.2405	0.1170	4.2263	0.0398
esrab	1	0.5432	0.1164	21.7762	<.0001

Time variables not shown.

Model for Numerator of Stabilized Weights MTX Exposure					
Parameter	DF	Estimate	Std Err	Wald ChiSq	Pr >ChiSq
Intercept	1	-123.8	18.3873	45.2986	<.0001
age	1	-0.0138	0.00354	15.1999	<.0001
male	1	0.0406	0.1119	0.1319	0.7165
yrra	1	0.0610	0.00919	44.1148	<.0001
rfpos	1	1.0258	0.1255	66.8070	<.0001
smokecb	1	0.0554	0.1384	0.1600	0.6892
smokefb	1	0.1123	0.1204	0.8708	0.3507

Time variables not shown.

Table 6: Treatment Models (see Step 4).

Model for Denominator of Stabilized Weights - Censoring					
Parameter	DF	Estimate	Std Err	Wald ChiSq	Pr >ChiSq
Intercept	1	1781.9	81.5345	477.6382	<.0001
exposed_mtx	1	-0.0120	0.1223	0.0096	0.9219
age	1	0.00714	0.00405	3.1016	0.0782
male	1	0.0122	0.1163	0.0109	0.9167
erosdest_	1	0.1464	0.1143	1.6404	0.2003
exra_sev_	1	0.1323	0.1943	0.4635	0.4960
yrra	1	-0.8860	0.0406	475.2877	<.0001
hcq_	1	0.0459	0.1140	0.1619	0.6874
othdmard_	1	-0.0696	0.1333	0.2728	0.6015
ster_	1	0.0340	0.1367	0.0620	0.8034
rfpos	1	-0.00719	0.1194	0.0036	0.9520
cld_	1	-0.0811	0.1357	0.3569	0.5503
nodul_	1	-0.1512	0.1243	1.4782	0.2241
ljswell_	1	-0.0306	0.1268	0.0583	0.8092
dmcd_	1	-0.1266	0.1478	0.7344	0.3914
hxalc_	1	0.1394	0.2101	0.4405	0.5069
dlipcd_	1	-0.0342	0.1201	0.0810	0.7759
smokefb	1	0.1300	0.1243	1.0944	0.2955
smokecb	1	0.1144	0.1545	0.5481	0.4591
esrab	1	0.2394	0.1437	2.7758	0.0957

Time variables not shown.

Model for Numerator of Stabilized Weights - Censoring					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr >ChiSq
Intercept	1	1761.7	80.5099	478.8203	<.0001
exposed_mtx	1	-0.00056	0.1109	0.0000	0.9959
age	1	0.00865	0.00359	5.7903	0.0161
male	1	-0.0198	0.1128	0.0308	0.8607
yrra	1	-0.8759	0.0401	476.4679	<.0001
rfpos	1	0.00876	0.1153	0.0058	0.9395
smokecb	1	0.1420	0.1440	0.9734	0.3238
smokefb	1	0.1256	0.1200	1.0953	0.2953

Time variables not shown.

Table 7: Censoring models (see Step 5).

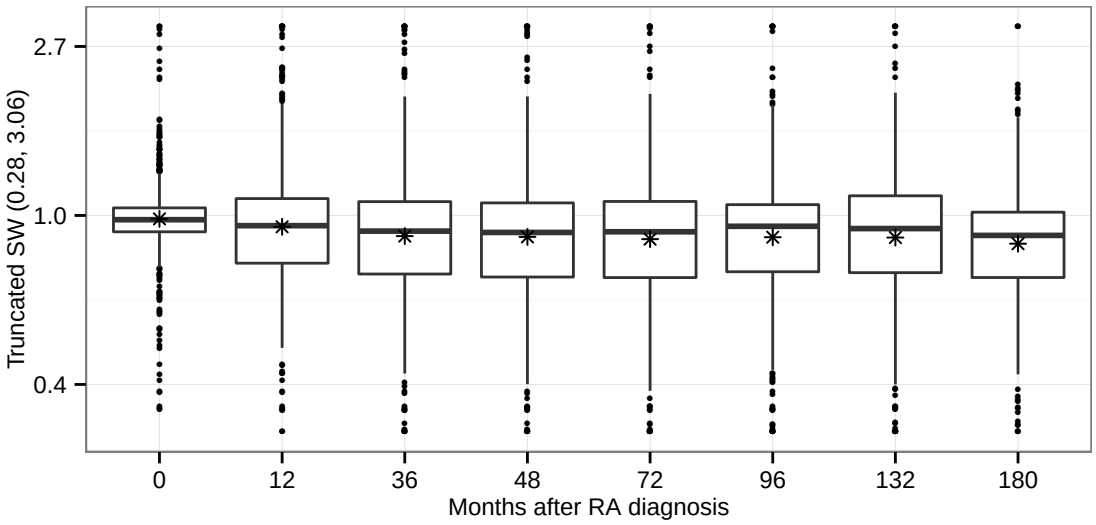

Figure 5: The distribution of weights at approximately 12 month intervals (see Step 6).

5.6 Step 6 - Examining the weights

The stabilized weights for treatment are computed by obtaining the predicted probabilities from the numerator and denominator models for each observation in the data set and then dividing these. This work is done by the macro. Note that once a patient is exposed to the treatment, both the numerator and the denominator are set to 1 for the remaining time periods, resulting in weights of 1.

The stabilized weights for censoring are computed by obtaining the predicted probabilities from the numerator and denominator models for each observation in the data set and then dividing these. This work is also done by the macro. Then the censoring weights are multiplied with the treatment weights to obtain the weights used in fitting the final model.

The treatment weights adjust for imbalances in the characteristics of the treated and untreated patients, and the censoring weights adjust for imbalances in the study population that develop over time.

When developing weighting models, it is important to visually inspect the individual weights being created. In some instances a poor performing covariate will cause many extreme weights to be created. Note that the macro also has an option for truncating extreme weights, but if many patients receive extreme weights it is important to understand why and to check model setup issues.

It is also important to verify that the mean of the stabilized IPW is close to 1. If the mean is not close to 1, this can indicate a violation of some of the model assumptions (e.g., positivity), or a misspecification of the weight models [20]. It is also important to verify that the sum of the weights is close to the actual sample size.

Figure 5 shows the distribution of the truncated stabilized weights, sw, over time. The mean (*) and the median (middle horizontal bar) and quartiles (horizontal box edges) are shown. The weights in the box plot were truncated below the 1st percentile and above the 99th percentile. The smallest actual weight using %msm was 0.15 and largest was 27.7. The weights were truncated to reduce variance as described in section 3. Notice the means and medians of the weights are reasonably close to 1.

5.7 Step 7 - Examining balance

The next step is to examine whether the inverse probability weights have resulted in balanced data

for the important factors confounding the treatment effect and the outcome. While data balance is often examined when using propensity scoring, it is often ignored in the MSM literature, likely due to the complexity of the time-varying weighting. In our example data we examined plots of the median covariate values in the treated and untreated groups over time both in the original and in the weighted data sets. If balance is improved with weighting, the central tendency of the covariate values should be closer together in the weighted data set. If balance is achieved, the treated and untreated lines should coincide.

Figure 6 below show the unweighted and weighted average age for patients exposed to MTX and not exposed to MTX. While the weighted median ages do not seem to be much closer together than the unweighted median ages, it is important to note that the age distributions of treated and untreated patients were largely overlapping and did not differ greatly to begin with. Despite this, we were a bit surprised that the age distributions did not appear to get closer together in the weighted sample. This could potentially indicate that we should spend more time trying to accurately model the age effect by considering non-linear age effects or interactions between age and other factors, such as sex.

Figure 7 show the unweighted and weighted proportions of patients with various characteristics in the treated and untreated groups. Notice that the weighted proportions tend to be closer together than the unweighted proportions.

5.8 Step 8 - The final model

The next step is fitting the final model to examine the estimated effect of the treatment on the outcome after applying weights to adjust for confounding factors. Note that the macro refers to the final model as the "structural" model, but we find that terminology a bit confusing, so we will refer to it as the "final" model. Table 8 summarizes the outcome model.

The adjusted odds ratio is 0.83 from the MSM, and indicates MTX has no statistically significant effect on mortality in patients with RA. The unadjusted odds ratio from Appendix A, 0.676 (95% CI: 0.515, 0.888), indicates patients exposed to MTX experienced improved survival. However, this apparent protective effect of MTX on mortality may have been due to confounding, since we no longer see this effect in the adjusted analysis. Also of note, the confidence intervals from the MSM analysis are quite wide (i.e., 95% CI: 0.56, 1.23).

Due to concerns regarding the possible over-inflation of the variance in the weighted final model, bootstrap variance estimation is recommended once the model is finalized. In fact, bootstrapping is a default setting in the macro call, but it is time-consuming and produces a great deal of output (e.g., >4000 pages in our example). Table 9 summarizes the outcome model obtained using bootstrapping. Note that the confidence intervals obtained using the bootstrap method are slightly narrower, .058–1.19.

The macro also provides estimates of the absolute treatment effect over time, in the form of survival plots. Figure 8 shows 3 survival curves generated by the %msm macro: with treatment, without treatment and natural course. The treated and untreated curves are obtained using the re-weighted study population, so these 2 curves can be compared to each other and their difference provides a visual representation of the absolute magnitude of the treatment effect. The natural course survival curve is obtained by fitting the final model without the treatment weight or the treatment indicator. The censoring weights are still applied to correct for any changes over time in the study population. So this curve represents what is currently happening in a study population with the characteristics and treatment patterns of the original study population, if all subjects were followed for entire study period. Since the natural course curve depicts the survival experience of the original study population, instead of the re-weighted study population, it is not comparable to the untreated and treated curves. Also of note, in our example the natural course curve is below both the untreated and the treated curves. Intuitively, the natural course should be between the treated and untreated curves,

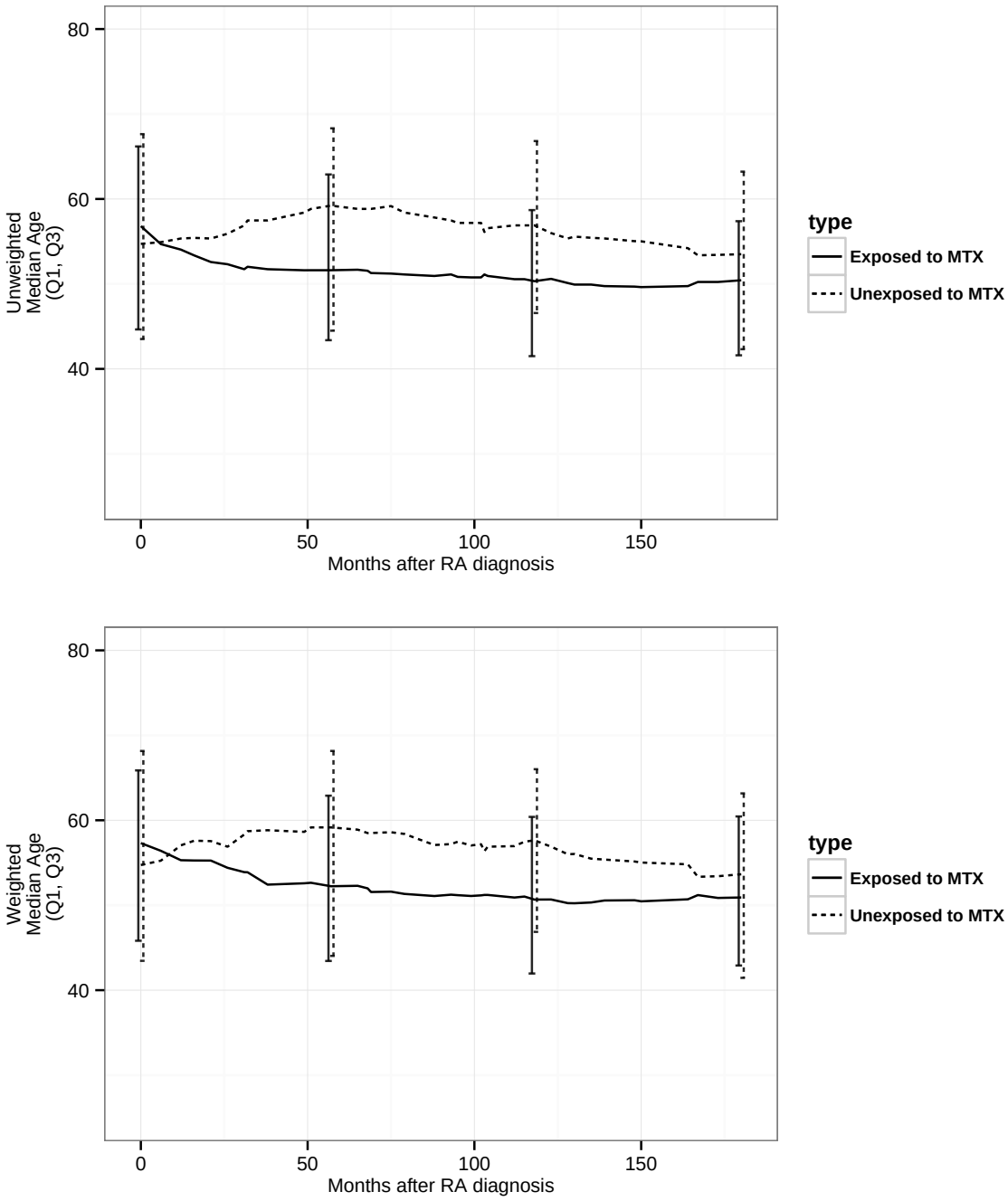

Figure 6: The distribution of median age over time in treated and untreated groups in the original data set (upper panel) and the weighted data set (lower panel) (see Step 7).

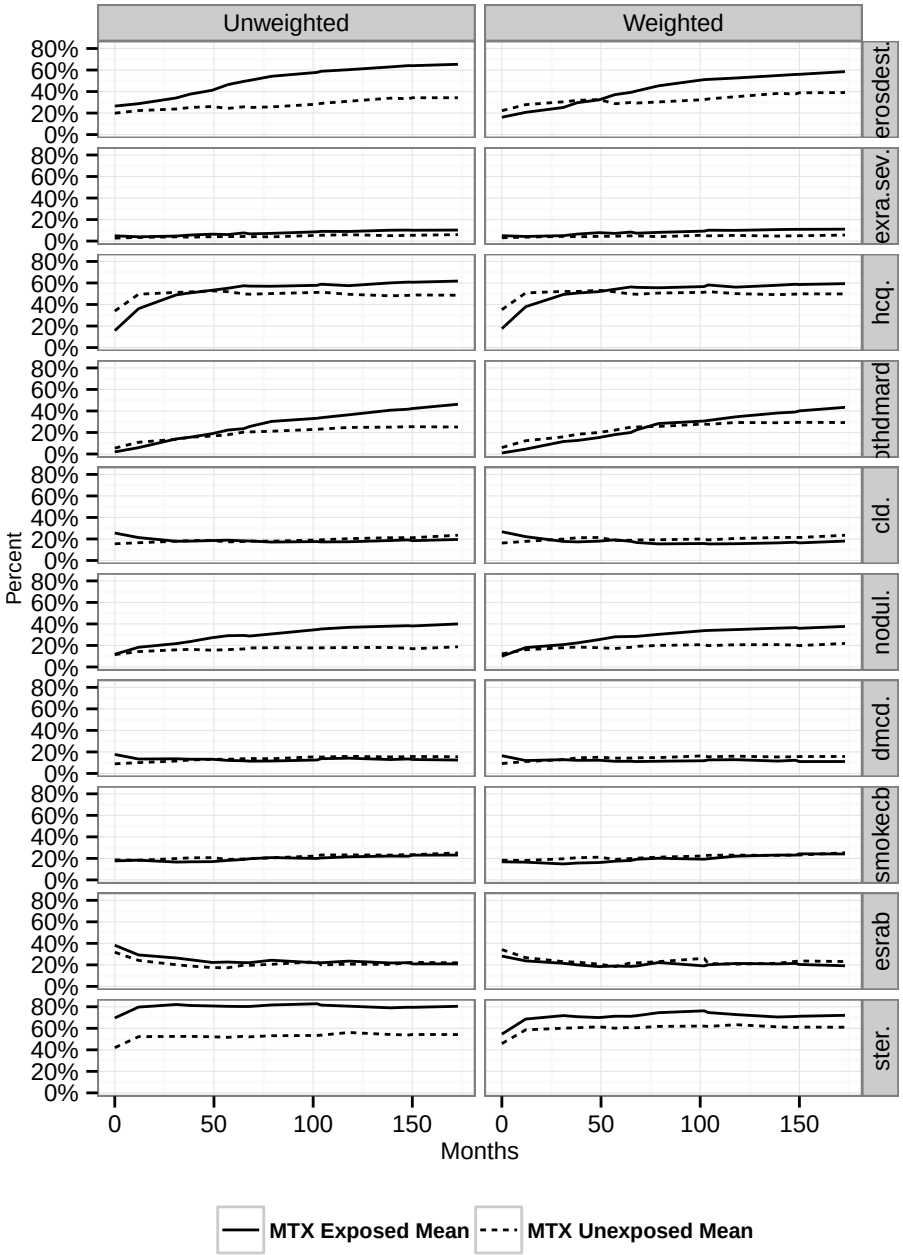

Figure 7: Unweighted and weighted proportions of various characteristics over time for the treated and untreated groups. (see Step 7).

Name	Estimate	Std	LB	UB	Z	P Value
Intercept	35.9350	31.906	-26.60	98.47	1.1263	0.2601
exposed_mtx	-0.1838	0.201	-0.58	0.21	-0.9126	0.3615
age	0.1254	0.009	0.11	0.14	13.5427	<.0001
male	-0.0718	0.176	-0.42	0.27	-0.4070	0.6840
yrra	-0.0257	0.016	-0.06	0.01	-1.6044	0.1086
rfpos	0.5997	0.188	0.23	0.97	3.1973	0.0014
smokecb	1.3954	0.236	0.93	1.86	5.9154	<.0001
smokefb	0.7453	0.216	0.32	1.17	3.4560	0.0005

Time variables not shown.

Table 8: Final outcome model (see Step 8).

Variable	Est	Bias	Std Err	95% lower	95% upper	Pr > |Z|
Intercept	35.93	-17.01	51.45	-64.92	136.7	0.48
exposed_mtx	-0.18	-0.02	0.18	-0.54	0.1	0.31
age	0.12	-0.00	0.01	0.10	0.1	<.0001
male	-0.07	-0.03	0.16	-0.38	0.2	0.65
yrra	-0.02	0.01	0.02	-0.07	0.0	0.31
rfpos	0.59	-0.01	0.25	0.10	1.0	0.01
smokecb	1.39	-0.08	0.24	0.90	1.8	<.0001
smokefb	0.74	-0.02	0.30	0.15	1.3	0.01

Time variables not shown.

Table 9: Final model with bootstrap errors. (Note that excessive digits were trimmed from the output to fit it on the page.) (See Step 8).

and this would be the case the same study population was used to draw all 3 curves. However, since the treated and untreated curves are obtained using the re-weighted study population and the natural course is obtained using the original study population, these curves are not comparable. The utility of the natural course curve is questionable and it does not appear in any published results using this methodology to our knowledge. We describe it here because it is part of the macro output.

Figure 9 shows the survival curve estimates using the bootstrap. Note the confidence intervals for the survival estimates are quite wide in this example. For example, the confidence intervals span more than 20% in each group by 100 months after RA diagnosis.

6 Counterfactuals

Much of the literature on MSMs focuses on counterfactuals and causal effects. Using counterfactuals to explain MSMs can be confusing. The context for counterfactuals is that the ideal, hypothetical study would be to observe all the study subjects under both conditions (e.g., both treated and untreated, or both exposed and unexposed), and then the true causal effect would be the difference between the outcomes when the same set of patients were exposed and when they were unexposed. The average of the individual causal effects across the entire population is often referred to as the Aver-

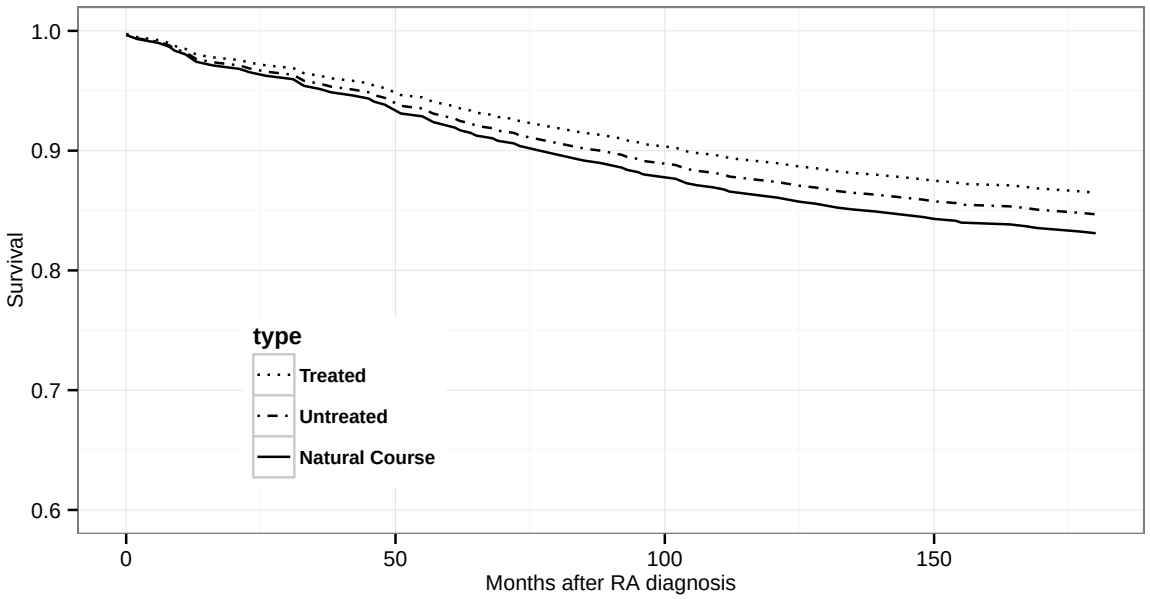

Figure 8: The survival curve estimates using the survival with and without treatment obtained using the re-weighted population and the natural course obtained using the original study population. Comparison of the treated and untreated curves provides a visual representation of the absolute treatment effect. The comparability of the natural course to the other curves is questionable.

age Causal Effect (ACE). In reality, each patient can only be exposed or unexposed; it is rarely possible to examine the same patients under both conditions. Even in a cross-over situation, there may be differences due to which order the treatments / exposures are experienced. The counterfactual refers to the unobserved state for each patient, so if the patient was treated, then the counterfactual is the outcome that would have occurred if the patient was not treated. Since we can rarely observe the same patients under both conditions, the only way to determine causal effects is to compare two groups that are identical in every way except that one group was treated and the other was not, which occurs when patients are randomly assigned to treatments and is referred to as exchangeability. MSMs attempt to fully adjust for confounders to simulate randomization and achieve balance and exchangeability in order to estimate causal effects.

7 Practical considerations

MSMs are time-consuming, as they require carefully fitting several different models for exposure, censoring and then the outcome of interest. Model modifications may be needed to achieve balance and to achieve a reasonable distribution of weights. In addition, sensitivity analyses are required to examine the effect of model assumptions. It may be necessary to fit several different variations of each of the models included in the MSM to see how the result is affected by modifying these models. Published papers reporting MSMs typically include results from several sensitivity analyses to strengthen

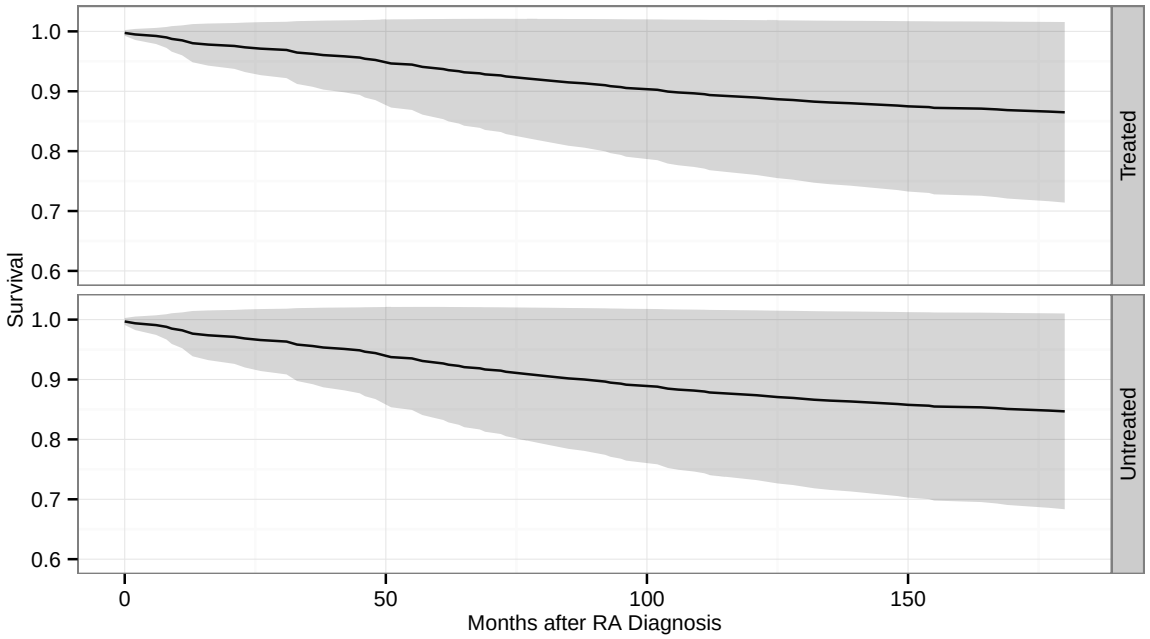

Figure 9: The survival curve with 95% confidence intervals estimated using the bootstrap (see Step 8).

the findings enough to convince skeptical readers by demonstrating stability of the estimated effect under various choices of model specifications.

Finally, the complexities of MSMs result in variances that are typically 20% higher than the variance in the original unadjusted model of treatment effect. Thus, this methodology requires adequate sample size to allow for a conclusive result in the face of this larger variance.[20]

8 Additional information

For more information on marginal structural models, examine the following references.

- The Basics of MSM [17]
- A medical example using MSM [7]
- Adjusted survival curves [3]
- More details/depth [8]
- Checking balance [6] [2]
- Pooled logistic regression models [12] [21]
- Propensity scores [19] [1]
- G-computation, an alternative to inverse probability weighting [22]

R users may want to investigate the *ipw* package [23] [22]. Work is underway to implement the MSM methods using the R *survival* library. A paper how using Stata to fit a Marginal Structural Model is available on the Harvard website (`http://www.hsph.harvard.edu/causal/`).

9 Acknowledgements

This work was supported by the Mayo Clinic Center for the Science of Health Care Delivery, by the National Institute of Arthritis and Musculoskeletal and Skin Diseases of the National Institutes of Health under Award Number R01AR46849, and by the Rochester Epidemiology Project, which is supported by the National Institute on Aging of the National Institutes of Health under Award Number R01AG034676. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

Appendices

A Pooled logistic vs. Cox model

As previously mentioned, the %msm macro uses pooled logistic regression models instead of Cox models to estimate the weights and also to model the long-term outcome. Pooled logistic regression models are equivalent to Cox models, and the theory demonstrating this fact was published by Laird and Olivier [12] and Whitehead [21]. The logistic regression models are fit using a data set with multiple observations per subject corresponding to units of time (e.g., months), and time is fit as a class variable to allow a separate intercept for each time. The separate intercepts for each unit of time mimic the baseline hazard in a Cox model.

Here is a simple example demonstrating how to create a data set with an observation for each month for each patient. The time interval for each observation ranges from t to tstop=t+1, and the exposed_mtx variable is a time-dependent covariate with the value of 0 prior to exposure to methotrexate, which changes to 1 at the start of exposure to methotrexate and remains 1 until the end of follow-up.

```
data ra;
    set <dataset>;
    tmlfu=lfudt-indexdt;
    tmmtx=mtx1dt-indexdt;
    keep id tmlfu dead tmmtx t tstop exposed_mtx;

    maxmonths=floor(tmlfu/30);
    exposed_mtx=0;
    do t=0 to maxmonths;
        if maxmonths=0 then do;
            if tmmtx^=. then exposed_mtx=1;
            end;
        else if . < tmmtx < tmlfu*(t+1)/maxmonths then exposed_mtx=1;
        tstop=t+1;
        output;
    end;
run;
```

When we compare the results of the pooled logistic model and the Cox model run on this data set, they are very similar, as expected.

```
**Cox model using monthly dataset created above**;
proc phreg;
```

```
      model (t,tstop)*dead(0)=exposed_mtx/rl ties=efron;

run;
```

Cox model

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr >ChiSq
exposed_mtx	1	-0.38848	0.13831	7.8898	0.0050

Parameter	Hazard Ratio	95% Hazard Ratio Confidence Limits
exposed_mtx	0.678	0.517, 0.889

```
**pooled logistic regression model using time in months as a class variable**;
proc logistic des;
        class t/ref=first;
        model dead=exposed_mtx t/rl;
run;
```

Logistic

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr >ChiSq
Intercept	1	-10.1760	10.2381	0.9879	0.3203
exposed_mtx	1	-0.3916	0.1390	7.9370	0.0048
t1	1	-4.5914	59.0956	0.0060	0.9381
t2	1	3.5533	10.2865	0.1193	0.7298
t3	1	-4.5744	59.2558	0.0060	0.9385
t4	1	-4.5705	59.2959	0.0059	0.9386
t5	1	3.5748	10.2865	0.1208	0.7282
t6	1	3.5851	10.2865	0.1215	0.7274
t7	1	-4.5583	59.4850	0.0059	0.9389
t8	1	4.6986	10.2542	0.2100	0.6468
t9	1	3.6062	10.2865	0.1229	0.7259
t10	1	3.6135	10.2865	0.1234	0.7254
Remaining Time variables not shown.					
t339	1	-3.5949	593.1	0.0000	0.9952
t340	1	-3.5851	681.5	0.0000	0.9958
t341	1	-3.5851	681.5	0.0000	0.9958
t342	1	9.8745	10.3109	0.9171	0.3382
t343	1	-3.5751	830.5	0.0000	0.9966
t344	1	-3.5751	830.5	0.0000	0.9966
t345	1	-3.5751	830.5	0.0000	0.9966
t346	1	-3.5751	830.5	0.0000	0.9966
t347	1	-3.5648	1168.4	0.0000	0.9976
t348	1	-3.5648	1168.4	0.0000	0.9976

Parameter	Odds Ratio	95% Odds Ratio Confidence Limits
exposed_mtx	0.676	0.515, 0.888

Note that the odds ratio from the pooled logistic regression model is similar to the hazard ratio from the Cox model. The intercept in a Cox model is the baseline hazard function, which does not appear on the standard output from a Cox model. The baseline hazard function in the Cox model is the same as the intercept and time variables in the logistic regression model. We can easily demonstrate this in R:

```
#run Cox model in R
fit1<-coxph(Surv(t, tstop, dead)~ exposed.mtx, data=s1)

#obtain survival curve for average covariate values
surv1<-survfit(fit1)

#transform survival function to cumulative hazard function
cumhaz1<- -log(surv1$surv)

#transform cumulative hazard function to hazard function
hazard1<-diff(c(0,cumhaz1))

#run pooled logistic regression model
lmfit1<-glm(dead~exposed.mtx + factor(t),
          family-binomial(logit), data=s1)

#obtain estimated intercept at mean covariate values
#coefficient 1 is the intercept term
#coefficient 2 is the exposed.mtx term and the mean
#          of this 0/1 variable is 0.4358
#coefficients 3 to 350 are the time coefficients
est2<-exp(c(0,lmfit$coef[1] + lmfit$coef[2]*0.4358 +
                            lmfit1$coef[3:350])

#The values of hazard1 and est2 are identical.
```

Figure 10 shows the baseline hazard function for this example. Note that the estimated intercepts from the pooled logistic regression model are identical.

A disadvantage of the pooled logistic regression model is that the estimation of the time coefficients can lead to computational issues. Time intervals where the data set does not have any events will yield a baseline hazard value near 0. As you can see in the figure, this happens quite often in our data set, and the Cox model has no trouble dealing with it. However, in the pooled logistic regression model, this results in problems with model convergence or extreme coefficient values. We found that warnings regarding "Quasi-complete separation of data points" and "maximum likelihood estimate may not exist" could be ignored, since we were not interested in the accuracy of the estimated intercepts for each time period, and estimates for the coefficients of interest were largely unaffected by these warnings. However, the %msm macro will stop executing when these errors occur, so if these errors arise during fitting of the

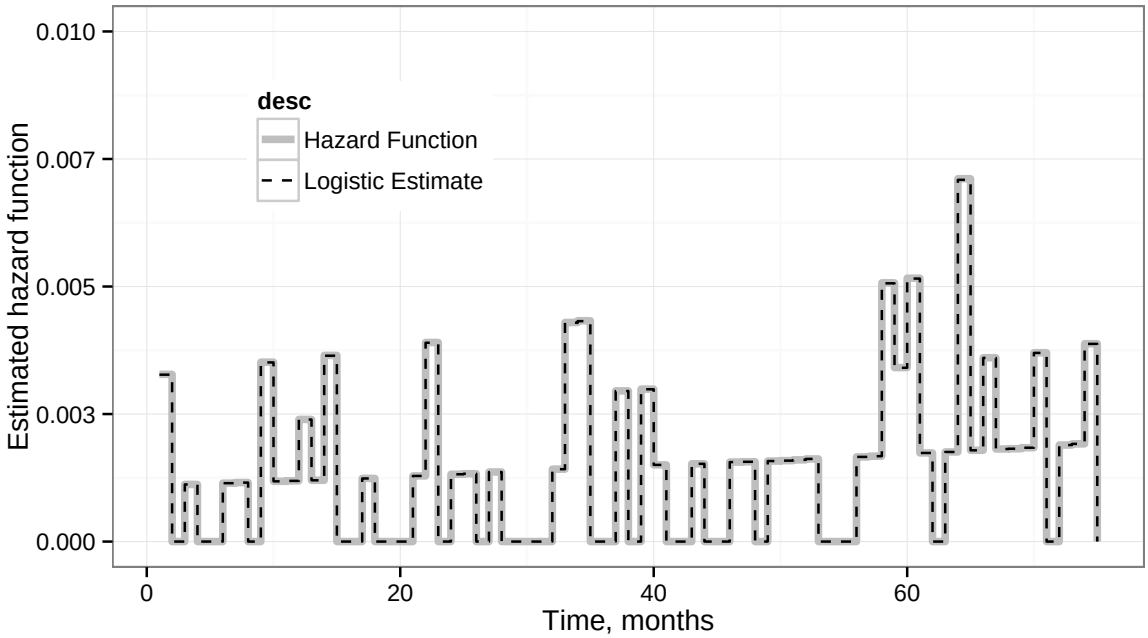

Figure 10: Baseline hazard function plotted with the logistic estimate.

treatment and/or censoring models, then the problem must be fixed in order to obtain the fit of the final model.

B %MSM macro documentation

The documentation below was taken from the %msm documentation and has been partially reproduced for convenience. See msm.sas code comments and msmdoc.pdf for additional details (`http://www.hsph.harvard.edu/causal/files/2012/10/msm.zip`).

data=	Data set name.
id=	Subject unique identifier (variable name).
time=	Time of follow-up (variable name) first observation for each subject must be time=0.(In descriptions below denote time index with m.)
	Structural Model Settings
outcMSM=	Time-varying outcome variable 1: event, 0: no event; missing if cens=1 or cens2=1 or cens3=1.
AMSM=	Time-varying exposure variable(s).
AMSM_class=	Categorical variables in AMSM (other than two levels). Default reference level will be first level.
covMSMbv=	Baseline variables to include in weighted models.

covMSMbh=	Baseline hazard variables (e.g., splines for time) to include in weighted models.
contMSMbh = 1	All variables listed in baseline hazard variables are continuous.
covMSM_extra =	Extra variables to include in the weighted model for both competing risk and outcMSM. When calculating the survival curves the user will need to modify the extra_variables_def submacro. This will be used in the creation of the variables for the weighted models and also in the curves data set.
extra_variables_used =	Extra variables to keep in the data sets that will be used to create the variables in covMSM_extra. Do not need to include the time variables, or any non-categorical variables listed in covMSMbv. You WILL NEED to include categorical variables that are listed in the class variables since these variables are converted to the equivalent binary variables and the original variables are not kept.
time_knots =	If non-missing, these are used for creating covMSMbh time categorical/ spline variables for final model and curves submacro.
classMSMbv =	Categorical variables in covMSMbv.
classMSMbh =	Categorical variables in covMSMbh.
inter_MSM=	Interactions with variables listed in AMSM.
inter_time=	Interactions with variables listed in covMSMbh.

Settings for treatment and possible censoring weight models.

A=	Time-varying outcome variable 1: treated, 0: untreated if missing, weights set to 1.
covAd=	Baseline and time-varying variables in treatment model for denominator of weights.
classAd=	List of categorical variables in covAd.
covAn=	Baseline variables in treatment model for numerator of weights.(Anything listed here also needs to be listed in covMSMbv.)
classAn=	List of categorical variables in covAn.
eligible=	Eligibility variable for treatment 1: yes, 0: no If 0 then pA_d and pA_n will be set to 1.
Cens=	Time-varying censoring indicator. 1: yes, 0: no
covCd=	Baseline and time-varying variables in model for denominator of weights
classCd=	Categorical variables in covCd
covCn=	Baseline variables in model for denominator of weights. (Anything listed here also needs to be listed in covMSMbv.)
classCn=	Categorical variables in covCn
eligible_cens =	Eligibility variable for Cens 1: yes, 0: no If 0 then pC_d and pC_n will be set to 1

Settings for second and third censoring models (Same as for cens).

Cens2=, covC2d=, classC2d=, covC2n=, classC2n=, eligible_cens2 =, Cens3=, covC3d= , classC3d=, covC3n=, classC3n=, eligible_cens3=,	
use_stabilized_wts = 1	Include the numerator model in the calculation of the weights
user_defined_weights = 0	Use user defined weights in analysis, skip calculate_weights macro
	Data analysis settings.
class_ref = first	Method for coding reference level for binary dummy variables for class variables.This level will be represented when all binary variables are equal to 0. first = default coding is to use lowest level for all categorical variables. last = use largest level as reference level in coding the binary variables.
use_genmod = 0	When equal to 1, use PROC GENMOD for final weighted model for outcMSM when not running bootstraps or calculating the analytic variance.Can be used for obtaining an estimate of the robust variance. Otherwise use PROC LOGISTIC.
msm_var = 1	Calculate analytic estimate of variance
truncate_weights = 0	0 default: use all 1 truncates weights below and above 1 and 99 percentile 2 truncates weights below and above 2 and 98 percentile, etc
user_defined_limits =	User defined lower and upper limits for the weights. This must be a list with two numbers separated with a space: user_defined_limits = lower_limit upper_limit. this will only be used when trun- cate_weights = 1 and user_defined_limits is not missing. If only one entry is listed or the list is impty then the method will used the percentile values given in the truncate_weights option.
	Bootstrap settings.
bootstrap= 0	Use bootstrap samples to estimate the variance of the parameters of the weighted models.
nboot= 200	Number of bootstrap samples to use.
bootstart = 0	Starting sample in current run of bootstrap samples (0 = original data).
bootend = 200	Last sample to use in current run of bootstrap samples (should be <= nboot.)
bootlib = work	Libname for saving bootstrap results.
bootname = boot	Name of data set for holding bootstrap results
bseed= 12345	Random number seed for boostrap Computational settings

just_models = 0	Estimate the treatment and censoring weights only. Can be used for testing possible models.
override_warnings= 0	Override the warnings from the weight models. Normally, when the weight models do not converge any further analyses are not performed. Use this option when running bootstraps to continue when there are warnings.
logistic_options =	Possible options to include in the PROC LOGISTIC model state- ments. (e.g. possible increasing of maxiter).
	Output settings
no_listing = 0	Suppress all listing output useful when calling macro from a loop and do not desire the output.
print_boot= 0	Print out the individual model results for each bootstrap sample.
save_results= 1	Save bootstrap results into sas data sets.
save_weights= 0	Save the weights to a permanent data set. (The macro will keep the weights in a temporary data set called _weights.)
libsurv= work	User defined libname for saving results, work directory will be default
results_name= _results	User defined name of file for saving results.
weights_name = _weights	User defined name of file to save calcuated weights in.
debug= 0	1 = keep intermediary data sets, 0 = delete intermediary data sets
	Survival curves and cumulative incidence settings.
survival_curves = 0	Calculate the survival probabilities for two situations where A = 0 and A = 1 for all time points. (This can be generalized to other types of treatment histories as described in the examples.)
AMSM_type = 0	Default type of function of treatment will be AMSM = A 1 = A_m and $\frac{1}{m}\sum_{j=0}^{j=m-1} A_j$ 2 = Categorical variable, which is a function of time treated. For curves submacro user must list the times where the function changes. 3 = User defined function of treatment. For curves macro this will need to be coded by user.
AMSM_knots=	When amsm_type = 2, this lists the times where amsm changes levels.

use_natural_course = 0	Include a natural course analysis that removes the treatment weights and AMSM from weighted models (only used when survival_curves = 1). When use_natural_course=1 and survival_curves = 1 all treatment variables that are to be included in the censoring weight models need to be included separately to make sure that they are not included in the weight models that are used in the natural course weighted models. Any treatement variables should not be included in covCd and covCn (and any secondary censoring models.) but included in the following two macro variables.
treatment_usen =	Treatment variables to include in the numerator models for each censoring variable.
treatment_used =	Treatment variables to include in the deniminator models for each censoring variable.
competing_risk =	An additional outcome for which a marginal structural model is fit. This is done assuming the same structural model and estimated weights as used for outcMSM. When using survival_curves = 1, a risk difference is also calculated at each time point. The user can select how this variable, riskdi, is defined. The possible variables are km_nontreat, km_treat, km_nc for survival probabilities, for never treated, always treated and under the natural course. There are corresponding variables for the cumulative risk: ci_nontreat, ci_treat, and ci_nc. The variables for _nc are only included when use_natural_course = 1. When there is a competing risk, these variables are calculated using methods suggested in Gooley.
riskdiff1 = ci_treat	One variable from the list { km_nontreat, km_treat, km_nc, ci_nontreat, ci_treat, ci_nc }.
riskdiff0 = ci_nontreat	select from ci_nontreat ci_nc for calcluating ci_treated - riskdiff0
print_ci = 0	
time_start = 0	Time point to start calculating the survival probabilities.
time_end = 60	Final time point to use in the survival probabilities.
time_step = 1	Time step used in calculating survival probabilities. Will assume that if this is not 1 then the probabilities are constant between the various time point values.

References

[1] P. C. Austin. *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies*, Multivariate Behavioral Research 2011;46: 399-424

[2] S. Belitser, E. Martens, W. Pestman, R. Groenwold, A de Boer, O. Klungel. *Measuring balance and model selection in propensity score methods*, Pharmacoepidemiology and drug safety 2011;20:1115-1129

[3] S. Cole, M. Hernán. *Adjusted survival curves with inverse probability weights*, Computer Methods and Programs in Biomedicine 2004;75:45-49

[4] S. Cole, M. Hernán. *Constructing Inverse Probability Weights for Marginal Structural Models*, American Journal of Epidemiology 2008;168:656-664

[5] A. Dispenzieri, J. A. Katzmann, R. A. Kyle, D. R. Larson, T. M. Therneau, C. L. Colby, R. J. Clark, G. P. Mead, S. Kumar, L. J. Melton III, S. V. Rajkumar, *Use of Nonclonal Serum Immunoglobulin Free Light Chains to Predict Overall Survival in the General Population*, Mayo Clin Proc. 2012;87(6):517-523

[6] R. Groenwold, F. de Vries, A. de Boer, W. Pestman, F. Rutten, A. Hoes, O. Klungel1. *Balance measures for propensity score methods: a clinical example on beta-agonist use and the risk of myocardial infarction*, Pharmacoepidemiology and drug safety 2011;20:1130-1137

[7] M. Hernán, B. Brumback, J Robins. *Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men*, Epidemiology, 2000;11:561-568

[8] M. Hernán, M. McAdams, N. McGrath, E. Lanoy, D Costagliola. *Observation plans in longitudinal studies with time-varying treatments*, Statistical Methods in Medical Research, 2009;18:2752

[9] L. Kish. *Weighting for Unequal P_i*, Journal of Official Statistics, 1992;8:183-200

[10] R. A. Kyle, T. M. Therneau, S. V. Rajkumar, D. R. Larson, M. F. Plevak, J. R. Offord, A. Dispenzieri, J. A. Katzmann, L. J. Melton III, *Prevalence of Monoclonal Gammopathy of Undetermined Significance*, New England Journal of Medicine 2006;354:1362-9.

[11] T. Kurth, A.M. Walker, R.J. Glynn, K.A. Chan, J.M. Gaziano, K. Burger, J.M. Robins. *Results of Multivariable Logistic Regression, Propensity Matching, Propensity Adjustment, and Propensity-based Weighting under Conditions of Nonuniform Effect*, American Journal of Epidemiology, 2006;163:262-270

[12] N. Laird, D. Olivier. *Covariance analysis of censored survival data using log-linear analysis techniques.* Journal of the American Statistical Association 1981; 76: 231-240.

[13] R. Logan, E Tchetgen, M. Hernán. *%MSM: SAS software for survival analysis using marginal structural models* April 5, 2011.

[14] J. K. Lunceford, M. Davidian. *Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study*, Statistics in Medicine, 2004;23:2937-2960.

[15] : E. Myasoedova, C.C. Crowson, H.M. Kremers, T.M. Therneau, S.E. Gabriel. *Is the incidence of rheumatoid arthritis rising?: results from Olmsted County, Minnesota, 1955-2007. Arthritis Rheum.* 2010 Jun;62(6):1576-82.

[16] M. Pugh. *Inference in the Cox Proportional Hazards Model with Missing Covariate Data*, dissertation, Harvard School of Public Health, 1993.

[17] J. Robins, M. Hernán, Babette Brumback. *Marginal Structural Models and Causal Inference in Epidemiology*, Epidemiology 2000;11:550-560

[18] J.M. Robins. *Data, Design,, and Background Knowledge in Etiologic Inference.* Epidemiology 2001; 11; 313-320.

[19] P. Rosenbaum, D. Rubin. *The Central Role of the Propensity Score in Observational Studies for Causal Effects*, Biometrika, 1983;70:41-55.

[20] D. Suarez, R. Borrás, X. Basagaña. *Difference Between Marginal Structural Models and Conventional Models in Their Exposure Effect Estimates: A Systematic Review.* Epidemiology 2011; 22: 586-588.

[21] J. Whitehead. *Fitting Cox's regression model to survival data using GLIM.* Applied Statistics 29, 268-275.

[22] W. M. van der Wal, M. Prins, B. Lumbreras, R. B. Geskus. *A simple G-computation algorithm, to quantify the causal effect of a secondary illness on the progression of a chronic disease.* Statistics in Medicine 2009; 28: 2325-2337.

[23] W.M. van der Wal, R.B. Geskus. *ipw: An R Package for Inverse Probability Weighting.* Journal of Statistical Software, 2011; 43:1-23. `http://www.jstatsoft.org/v43/i13/`.

[24] S. Xu, C. Ross, M. A. Raebel, S. Shetterly, C. Blanchette, D. Smith. *Use of Stabilized Inverse Propensity Scores as Weights to Directly Estimate Relative Risk and Its Confidence Intervals.* Value in Health 2010; 13: 273-277.