Title
Editorial: Observational studies, time-dependent confounding, and marginal structural models

Permalink
https://escholarship.org/uc/item/3n42235s

Journal
Arthritis and Rheumatology, 67(3)

ISSN
2326-5191

Author
McCulloch, CE

Publication Date
2015

DOI
10.1002/art.38931

Peer reviewed
Observational Studies, Time-Dependent Confounding, and Marginal Structural Models

Charles E. McCulloch

Observational studies have an undeniable role in medical research and often have significant advantages over randomized trials (1), such as broader generalizability. But, we regularly want to use observational studies to draw causal conclusions about the ability of interventions to prevent and treat disease, which can be problematic. In companion articles published in this issue of Arthritis & Rheumatology, Yang, Lapane, and coworkers (2,3) consider the causal impact of glucosamine/chondroitin and nonsteroidal antiinflammatory drugs (NSAIDs) on knee osteoarthritis (OA), based on data from the Osteoarthritis Initiative (OAI), a longitudinal observational study. They use a statistical technique called “marginal structural models” with the goal of strengthening their ability to infer causation. This is becoming a more common approach to addressing issues of confounding in observational studies. When should this technique be used, and what are its advantages and disadvantages?

It is well understood that a key threat to drawing causal conclusions from observational studies is confounding. For example, a history of knee surgery could easily be associated with both NSAID use and OA, and failure to account for differences in rates of knee surgery in comparing users and nonusers of NSAIDs could lead to biased estimates of the causal effect of NSAIDs. Fortunately, multiple regression analysis is in the statistical toolkit of virtually any medical researcher and can, in a wide variety of situations, adjust for or control for the effects of history of knee surgery to reduce or eliminate the effects of confounding.

But what about a variable such as inflammation? Inflammation could also easily be associated with both NSAID use and OA. The conventional wisdom is that we should not adjust or control for inflammation because it may be one of the paths by which NSAIDs could work to reduce OA. If we adjusted or controlled for it by multiple regression, we might be removing or reducing the true causal effect of NSAIDs on OA that occurs through reduced inflammation. So, variables such as inflammation (called mediator or intermediate variables) should not be adjusted or controlled for (4). So far, so good. We should adjust for confounders but not intermediate variables.

Unfortunately, it is possible for a variable to be both a confounder and an intermediate variable in a longitudinal study. For example, inflammation during the first year of a study (between baseline and year 1) could cause higher NSAID use at the year 1 visit. So, it is a confounder during year 1, and we should adjust for it. NSAID use might then lead to reductions in inflammation between year 1 and 2 and be associated with lower rates of OA. So, it is an intermediate variable during year 2, and we should not adjust for it. Variables that are simultaneously confounders and intermediate variables are known as “time-dependent confounders.”

The technique of marginal structural models sets out to solve this dilemma by accounting for the association of inflammation and NSAID use not by adjusting for it in a regression model but instead by using weighting methods. The basic idea is to weight the data to break the association between the time-dependent confounder and the outcome. Continuing the hypothetical inflammation example, suppose that there are 50 participants with no inflammation during year 1, 10% (or n = 5) of whom reported NSAID use at year 1, and 100 participants with inflammation during year 1, 20% (or n = 20) of whom report NSAID use at year 1. The weighting used in marginal structural models is called “inverse probability of treatment” weighting. So, in the no inflammation group, those who took NSAIDs are weighted by 10 (from 1/0.1 = 10) and those who did not take NSAIDs are weighted by 1.11 (from 1/0.9 = 1.11). Thus, the 5 participants in the no inflammation group who took NSAIDs are essentially inflated (by the...
weighting) to a group of 50 (or, $5 \times 10 = 50$) who took NSAIDs, and the 45 participants in the no inflammation group who did not take NSAIDs are inflated to a group of 50 (or, $45 \times 1.11 \approx 50$) who did not take NSAIDs.

Similar calculations within the inflammation group (using weights of $1/0.2$ and $1/0.8$) lead to inflated groups of 100 (see Figure 1). So, a weighted analysis behaves as if it has equal numbers of subjects taking and not taking NSAIDs within the inflammation and within the no inflammation groups. This weighting breaks the association between inflammation and NSAID use and, hence, eliminates confounding due to inflammation. But, it keeps constant the proportions with and without inflammation, so that effect estimates reflect the inflammation/no inflammation division. Importantly, because we have not controlled for the effects of inflammation using a regression analysis, we will not remove the intermediate effect of NSAIDs on OA that occurs through reduction of inflammation. Marginal structural models solve the time-dependent confounding dilemma by using a weighting system that removes the effects of confounding without adjusting away the intermediate effects.

Marginal structural models are one of the few ways to deal with time-dependent confounders, and they therefore deserve strong consideration when faced with that situation. They come with significant drawbacks, however. These include decreased precision, the requirement to build and check subsidiary statistical models that are not primary to the scientific question, and difficulties of use with studies that do not have regularly scheduled visits. I will briefly elaborate on each of these points.

Inherent in the use of marginal statistical models is weighting by the inverse probability of treatment. In the hypothetical example above, the no-inflammation NSAID group was inflated by a factor of 10. This serves to magnify any errors associated with those individuals and to reduce the precision of estimated effects. For example, in Table 3 of the Yang et al study (2), the width of the confidence intervals (and hence the standard errors) of the estimated treatment effects are more than 40% larger on average for the marginal structural model than for a more standard longitudinal regression analysis (generalized estimating equations). Translating this into a required sample size (which is proportional to the square of a standard error) means that to achieve equivalent precision, we would need to double the sample size to accommodate the marginal structural model analysis.

Typically, the probability of treatment depends on many potential confounding factors. So, usual practice is to build a logistic regression model for the probability of treatment as a function of those confounders to get an estimated probability of treatment to use in the weighting. The effort required to build and check this subsidiary model is of the same magnitude as that required for building and checking the model for the outcome of interest. In studies where the outcome of interest is a time to event outcome (e.g., time to total knee replacement), it is also common to build a model for the probability of censoring (5), which adds to the modeling burden. Finally, with many confounding factors, a not-insignificant proportion of the observations may have a very small probability of treatment (or no treatment), leading to very large inverse probability weights and unstable and highly imprecise estimates. It is therefore common to “truncate” the weights to reduce this sensitivity (2,3,6), often subjectively.

The use of marginal structural models is facilitated when the observations are made on a regularly spaced grid, such as in the OAI, which had (essentially) yearly visits. In less controlled data collection situations, such as trying to use an electronic health record system to conduct an observational study, visits will be much more haphazard. In such a case, there is considerable

![Figure 1](image.png)
extra effort required to put the data into amenable formats as well as subjective decisions to be made about time intervals for the analysis.

Marginal structural models, like their regression counterparts, depend on some key assumptions. Not surprisingly, but crucially, they require that all confounders be measured so they can be included in the modeling. This assumption will rarely be met in practice. Also not surprisingly, they depend on the correctness of the modeling assumptions (e.g., linearity of numeric predictors or inclusion of all needed interactions). This is exacerbated by the additional modeling burden described above.

In summary, it is important to consider the potential for time-dependent confounding in longitudinal studies because standard methods of analysis may then give biased estimates. Marginal structural models are a possible solution to the time-dependent confounding dilemma. However, their use should not be embarked upon lightly. They require substantial additional modeling efforts, and they trade a potential reduction in bias for what can be a significant decrease in precision.

**AUTHOR CONTRIBUTIONS**

Dr. McCulloch drafted the article, revised it critically for important intellectual content, and approved the final version to be published.

**REFERENCES**