Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1356-1294

Evaluating health management programmes over time: application of propensity score-based weighting to longitudinal data

Ariel Linden DrPH MS[1] and John L. Adams PhD[2]

[1]President, Linden Consulting Group, Hillsboro, OR, USA
[2]Senior Statistician, RAND Corporation, Santa Monica, CA, USA

Keywords
disease management, health management, inverse probability of treatment weights, longitudinal data, propensity score

Correspondence
Correspondence
Ariel Linden
Linden Consulting Group
Hillsboro OR 97124
USA
E-mail: alinden@lindenconsulting.org

Accepted for publication: 25 November 2009

doi:10.1111/j.1365-2753.2009.01361.x

Abstract

Health management programmes are generally evaluated as point treatment studies in which only a baseline and outcome measurement are used in the analysis, even when multiple observations for each individual are available. By summarizing observations into two distinct measurements the evaluator loses any ability to discern patterns of change in the outcome variable over time in relation to the intervention. There are several statistical models available to evaluate longitudinal data that are typically regression-like in form and designed to adjust for clustering at the individual level. Most evaluators of longitudinal studies tend to adjust for the effect of time-dependent confounding by including these covariates as independent variables in the model. However, this standard adjustment approach is likely to provide biased estimates. In this paper we describe the application of the propensity score-based weighting technique to longitudinal data to estimate the effect of treatment on an outcome. This method reweights each treatment pattern to represent the entire population at each time point and provides an unbiased treatment effect. We illustrate the technique using data from a disease management programme and demonstrate its superiority over standard analytical adjustments in correcting for time-dependent confounding for each time period under study.

Introduction

Health management programmes are generally evaluated as point treatment studies in which only a baseline and outcome measurement are used in the analysis. For some types of data, such as surveys, it may be that only two measurements are available because of the prohibitive cost and resources needed to collect observations more frequently. In other instances, while more periodic data may actually be available (such as medical claims data or biometric feeds from remote telemonitoring devices), the evaluator may nevertheless choose to aggregate the data into pre- and post-intervention observations, possibly to simplify the analysis or make the results easier to understand for non-researchers.

While it is perfectly valid to summarize multiple observations into two distinct measurements for the purpose of evaluation, the evaluator loses any ability to discern patterns of change in the outcome variable over time in relation to the intervention. For researchers and programme administrators alike, establishing the temporal relationship between treatment and outcome may be as important as determining whether there is a programme effect at all.

There are several statistical models available to evaluate longitudinal data and they generally share some basic commonalities: these models are typically regression-like in form [e.g. random effects, fixed effects, mixed effects, general estimating equations (GEE)], and they are designed to adjust for clustering at the individual level (e.g. outcomes measured repeatedly within individuals will generally be highly correlated and may produce inaccurate estimates using standard regression models).

Longitudinal data from health management interventions are particularly susceptible to the influence of time-dependent confounders and therefore, any of the aforementioned models must adjust for this bias accordingly. A time-dependent confounder is a variable that obscures the true relationship between treatment and outcome if a past level of that variable independently predicts both the exposure to treatment and the outcome. As an example, perceived health status has been shown to predict mortality [1], but it is likely to predict participation in a health management programme if a low score is used as criteria for enrolment. As illustrated in Fig. 1, health status confounds the effect of the programme on mortality because health status predicts both programme participation and mortality. Additionally, an effective

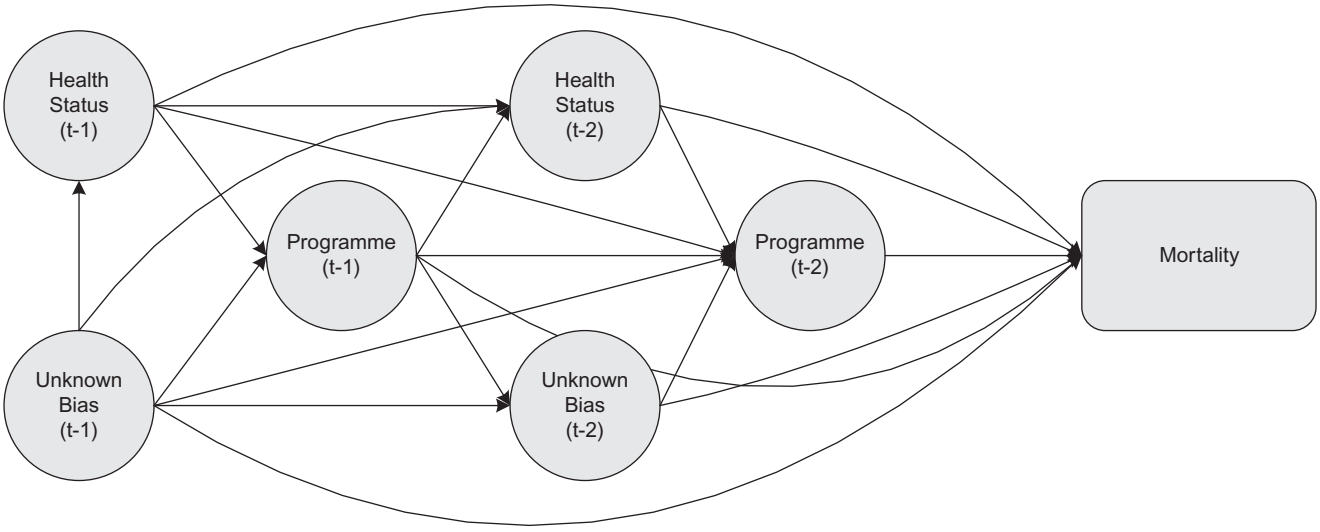

Figure 1 Illustration of the time-dependent relationship between health status, unmeasured sources of bias, programme participation and the outcome – mortality. Adapted from Robins *et al.* [2]

intervention will impact future measurements of health status that in turn may impact whether that person will continue receiving the intervention or get discharged from the programme. In this situation, health status is confounded by prior treatment.

Most evaluators of longitudinal studies tend to adjust for the effect of confounding by including these covariates as independent variables in the model (as they would in point treatment studies using standard regression models). However, this standard adjustment approach has been shown to provide biased estimates [2].

A recently developed modelling approach for longitudinal observational data generally referred to as *marginal structural models* (MSM) adjusts for time-dependent confounding as well as other biases present in non-randomized studies using a time-updated propensity score-based weighting procedure [2,3]. The weight is based on the conditional probability of an individual receiving his/her own treatment at each time point, and is called the 'inverse probability of treatment weight' (IPTW) [2,3]. This weighting mechanism weights the treated participants to the population from where they were drawn at each time point, thereby allowing unbiased population estimates to be calculated at each period or across all time points in aggregate.

This paper presents a non-technical introduction to the IPTW procedure for evaluating longitudinal data. It builds on our accompanying paper in the current issue [4], which described the weighting procedure for use in point treatment studies. For the purpose of illustration, data from a chronic disease management programme will be used with medical costs as the primary outcome.

Inverse probability of treatment weighting in longitudinal data

As described in Linden and Adams [4], Robins [3] and Robins, Hernán and Brumback [2] have applied the weighting concept developed in the survey sciences to adjust for imbalances in sampling pools [5] to the study of treatment effects in observational health care studies, using the estimated propensity score to represent the conditional probability of treatment [6,7]. Treated subjects are given a weight of 1/(propensity score) and non-treated subjects are given a weight of 1/(1 – propensity score) [2].

The IPTW mechanism can be thought of as creating a pseudopopulation comprised of 'copies' of the original subjects who account not only for themselves but for subjects with similar characteristics who received the alternate exposure [8]. More specifically, a programme participant with a low estimated propensity score will contribute many more copies of him/herself to the pseudopopulation than a participant with a high estimated propensity score (e.g. a programme participant with a propensity score of 0.01 will contribute 100 copies while a participant with a propensity score of 1.0 will contribute only one copy). The interpretation of the weighting formula for non-participants is analogous [4].

In point treatment studies, the IPTW allows us to view the pseudopopulation as one in which all individuals are considered conditionally exchangeable by ensuring that balance is achieved between treated and non-treated groups on pre-intervention characteristics [9]. In contrast, the IPTW in longitudinal studies is updated at each time point, ensuring that balance is achieved not only at baseline, but at each observation up to the most recent period. As such, the IPTW allows the researcher to control for effects of time-dependent confounders as well as many other potential biases that may have occurred after the initial baseline measurement was taken (see Fig. 1).

The first step in calculating the IPTW in longitudinal data is to estimate the propensity score for each person at each period. More specifically, pooled logistic regression is used to estimate the probability of assignment to the treatment group conditional on covariates from the current and past periods. Table 1 provides an example of such a model in which there are three observation periods. The first measurement period represents the baseline, and therefore, the covariates include only those variables measured prior to the start of the intervention. The second measurement is taken at some point after the intervention commences and thus the model includes covariates from the prior measurement (period 1)

Table 1 A hypothetical propensity score model for a programme with three measurement periods

Measurement period	Covariates
1	(Period 1)
2	(Periods 1 and 2) + outcome (period 1)
3	(Periods 1, 2 and 3) + outcome (periods 1 and 2)

The treatment variable from the prior period could also be added into the current measurement period model, but that may be more important for programmes in which treatment is not constant.

and from the current period (period 2). The outcome variable measured at the prior period can be added to the model as well; however, care must be taken to ensure that the outcome from the current period is not included in the propensity score model, as it will be used in the outcome model at a later stage of analysis. Additionally, prior treatment status can also be added as a covariate to capture the persistency of treatment. The third measurement (and any subsequent period model) is analogous to period 2. Boosted logistic regression [10] is worth considering as an alternative to the standard logistic model in estimating the propensity score. Regression boosting, commonly referred to as multiple additive regression trees, is a general, automated, data-adaptive modelling algorithm that can estimate the non-linear relationship between the outcome variable (in this case, treatment assignment) and a large number of covariates including multiple level interaction terms resulting in greater accuracy over standard linear models [11].

Once the propensity score is estimated for each person period, the IPTW weights can then be generated for treated and non-treated individuals. In studies where the propensity score distribution has large variability (possibly because of some covariates being highly correlated with treatment), it is possible that some treatment patterns will have extremely large weights. In this situation, Robins *et al.* [2] and Hernán *et al.* [12] recommend replacing the IPTW with *stabilized weights* to reduce this variability and ensure the estimated treatment effect remains unbiased. Generally, a review of summary statistics after generating the IPTW weights will help determine if stabilized weights are necessary. A user-written program for Stata called *propwt* (available from A. Linden) generates both stabilized and unstabilized IPTW weights (in addition to generating several other potentially useful weights that can be used in either point treatment or longitudinal studies).

Upon completion of this step, each individual will have a weight for each period in which they have data, accounting not only for themselves but for subjects with similar characteristics who received the alternate exposure to treatment up to that period. In effect, this process decouples the relationship between a person's probability of receiving the intervention and their time-updated covariate mix.

Model estimation using inverse probability of treatment weight in longitudinal analyses

Unbiased treatment effects can be estimated by fitting the appropriate statistical model using the IPTW as the specified weight (e.g. in the Stata software package one would specify the IPTW as either an analytical weight or sampling weight). Like any other outcome variable, the choice of model depends on the distribution of the outcome variable. This can be logistic regression for dichotomous variables, ordinary least squares for continuous variables, Poisson for rates or rare events, and pooled logistic regression for survival or censored cases [13]. Some researchers prefer the use of generalized linear modelling for its flexible distributional assumptions [14]. Regardless of which of these traditional regression models are used, standard errors must be adjusted to correct for within subject correlation by either clustering at the individual level or using robust standard errors [15]. Conversely, evaluators can choose from among more complex models specifically designed to account for within subject correlation in longitudinal data, such as GEE, random effects, fixed effects or mixed effects models (readers are referred to Rabe-Hesketh and Skrondal [16] and Fitzmaurice *et al.* [17] for a comprehensive discussion on these models). Standard errors must also account for the variability in weights from the IPTW.

As the IPTW implicitly incorporates covariates into the statistical model by way of the propensity score, it is generally sufficient to include only the treatment variable into the structural component of the regression (e.g. the right side). However, covariates may be added if reviewing their contribution to the model is important. As of yet 'doubly robust estimators' [18–20] that require the use of IPTW and covariate within the same regression model, do not appear to have been applied to longitudinal models. In point treatment studies using IPTW, an estimator is doubly robust if it remains consistent when either model (propensity score or outcomes regression) is correctly specified. Therefore, an evaluator is given two chances, instead of only one, to make a valid inference. Future efforts should focus on extending this technique to longitudinal data.

Example of the inverse probability of treatment weight applied to a longitudinal analysis of a health management programme

For the purpose of illustrating how the MSM/IPTW concept can be applied to longitudinal analyses we use data from a health management programme that invites individuals with chronic conditions to enrol in a nursing intervention intended to improve clinical indices of care while reducing medical costs. The data consist of 24 monthly observations for 155 programme participants and 7713 non-participants (for a total of 188 832 observations).

The first 12 months of data representing the pre-programme baseline period are presented in Table 2. As would be expected in a non-randomized programme, participants were significantly older, sicker and more costly than non-participants. In addition to this obvious selection bias, one can easily envision several potential sources of confounding when reviewing the data in a longitudinal context. For example, individuals with a chronic illness who are not getting their prescriptions filled may be more likely to be targeted for programme enrolment. To get the prescription, these patients would have to see their doctor, which in turn may reduce the likelihood of the patient presenting to the emergency department for an acute exacerbation of the condition, which may further result in a hospital admission and ultimately incurring high costs.

Table 2 Baseline (12 months) characteristics of programme participants and non-participants

Variable*	Participants	Non-participants	*P*-value†
N	155	7713	
Age	56.45 (9.2)	46.59 (11.0)	<0.0001
Female (%)	45.81	53.26	0.06
Congestive heart failure (%)	9.68	0.65	<0.0001
Coronary heart disease (%)	30.32	3.16	<0.0001
Chronic obstructive pulmonary disease (%)	14.19	4.36	<0.0001
Diabetes (%)	67.10	9.87	<0.0001
Hospital admissions	0.21 (0.5)	0.04 (0.3)	<0.0001
Emergency department visits	0.43 (1.1)	0.12 (0.4)	<0.0001
Doctor office visits	9.65 (6.3)	3.81 (4.4)	<0.0001
Prescriptions	46.45 (28.4)	11.62 (16.4)	<0.0001
Total costs	$13 522 (17 585)	$3107 (8857)	<0.0001

*Unless otherwise noted, variables presented are means and standard errors.

†*P*-values for means were derived using *t*-tests for independent samples and *P*-values for percentages were derived using *t*-test for proportions.

In this scenario, the patient's level of compliance with the medication regimen, or their ability to schedule a doctor visit, or their preference for using the emergency department as their primary source of care could all confound the relationship between programme enrolment and costs.

In generating the propensity scores for each person period, we followed the logic described in Table 1. The propensity score for the baseline was estimated using all the variables listed in Table 2 as covariates and aggregated into one annual block (rather than 12 monthly increments). Similarly, the propensity score for the first programme month was estimated using the baseline values and those covariates of the first month (excluding current period costs). All subsequent person period propensity scores were estimated in a similar fashion. At the end of this process step, a total of 13 propensity scores were estimated per individual (one baseline and 12 programme periods).

Raw and stabilized IPTW weights were then generated for each person period using the *propwt* user-written program described earlier. A review of the summary statistics indicated that stabilizing the weights reduced the variability around the mean to a large degree (the mean and standard deviation for the raw IPTW was 1.42 and 18.05, while the mean and standard deviation for the stabilized IPTW was 0.99 and 1.96, respectively). This suggested that the stabilized weights would be preferred over raw weights when estimating the outcome model in the next stage of analysis.

We estimated the outcome model using generalized linear modelling in Stata (version 10.1) using the IPTW as the weight and clustering on the individual to provide robust standard errors ('clustered robust' option in Stata). To simplify the exposition, costs were regressed on the treatment covariate only and the stabilized weight was employed as the analytical weight ('aweight' option in Stata). Figure 2 illustrates the difference in monthly costs for participants over non-participants using a naïve estimate (costs regressed on the treatment variable without the weights to adjust for selection bias or time-dependent confounding). As shown, the monthly medical costs of programme participants ranges are significantly higher than non-participants in every month of the programme, with point estimates (representing the mean difference between the groups) ranging from $604 to $1274 over the 12-month period. The naïve estimate for the aggregate 12-month period was $1204 higher costs for the treatment group per month (95% CI = $927.9, $1119.9).

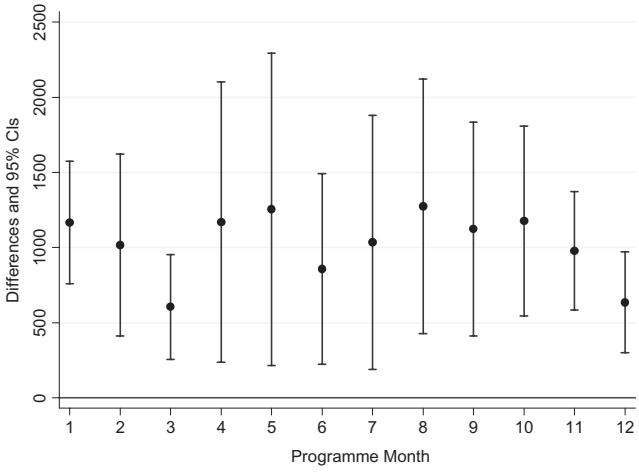

Figure 2 Naïve estimate of programme effect on costs. Values and 95% CIs represent the treated group's monthly costs relative to the non-treated comparison group (red line at zero).

Figure 3 illustrates the difference in monthly costs for participants over non-participants using the IPTW weights in the model estimation. As shown, once adjustments are made to control for selection bias and time-dependent confounding, the participants' group no longer have statistically higher medical costs than non-participants (monthly point estimates for the difference between groups range from –$761 to $847). The IPTW adjusted estimate for the aggregate 12-month period was $98 higher costs for the treatment group per month (95% CI = –$198.6, $394.2). These estimates were consistent when the process was replicated in the SAS statistical software package (SAS Institute, Inc., Cary, NC, USA) using 'Proc Genmod' with an independence correlation matrix and clustering on the individual (in this context, the 'Proc Genmod' statement produces GEE estimates).

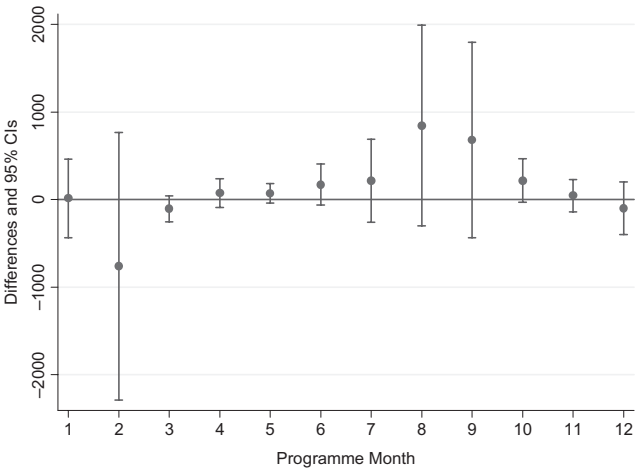

Figure 3 Weighted estimates of programme effect on costs. Values and 95% CIs represent the treated group's monthly costs relative to the non-treated comparison group (red line at zero).

For comparison purposes, we re-estimated the model using standard regression techniques. Here, costs were regressed on the treatment variable and all other covariates for the current and past periods (utilizing the same variables as those used to produce the IPTW for the weighted model). The standard model estimate was –$168.7 per month (95% CI = –$276.3, –$61.1) for the treatment group, suggesting that the participant group exhibited a small but statistically significant monthly drop in costs relative to the non-participant group.

In summary, upon controlling for selection bias and time-dependent confounding, medical costs for participants in a chronic disease management programme were no higher than non-participants, whereas standard regression techniques provided biased estimates indicating a slight monthly decrease in medical costs for programme participants.

Limitations of the inverse probability of treatment weight technique in longitudinal data

As with any evaluation of observational data, the foremost limitation is that we presume that all biases and confounding have been adjusted for in the model, an assumption that cannot truly be tested outside of a randomized study. One limitation specific to inverse probability weighting is that propensity scores for programme participants must be different from zero. In effect, no treatment effect can be estimated for people who have no probability of receiving the treatment [4]. A related problem will arise if everyone in the population will either receive or not receive the intervention at a specific time point. From a mathematical perspective, logistic regression cannot produce estimates if all values for the outcome are identical, thus the propensity score cannot be estimated.

Conclusion

In this paper we have described the application of the marginal structural modelling weighting technique to longitudinal data to estimate the effect of treatment on an outcome. This method reweights each treatment pattern to represent the entire population at each time point and provides an unbiased treatment effect. As illustrated in the example provided, the IPTW technique is superior to standard analytical adjustments because of its ability to correct for time-dependent confounding at each time period under study. This provides us with confidence that the groups remain essentially equivalent (assuming there is no residual confounding), thereby allowing us to make causal inferences about treatment effects. Given the robustness of this analytical technique, the IPTW should be considered as an alternative procedure for use with longitudinal observational data to evaluate health management programme effectiveness.

References

1. Idler, E. L. & Benyamini, Y. (1997) Self-rated health and mortality: a review of twenty-seven community studies. *Journal of Health and Social Behavior*, 38 (1), 21–37.
2. Robins, J. M., Hernán, M. A. & Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550–560.
3. Robins, J. M. (1998) Marginal structural models. In 1997 Proceedings of the Section on Bayesian Statistical Science, pp. 1–10. Alexandria, VA: American Statistical Association.
4. Linden, A. & Adams, J. (2010) Using propensity score-based weighting in the evaluation of health management program effectiveness. *Journal of Evaluation in Clinical Practice*, 16, 175–179.
5. Horvitz, D. G. & Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.
6. Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
7. Linden, A., Adams, J. & Roberts, N. (2005) Using propensity scores to construct comparable control groups for disease management program evaluation. *Disease Management & Health Outcomes*, 13 (2), 107–127.
8. Hernán, M. A., Hernández-Diaz, S. & Robins, J. M. (2004) A structural approach to selection bias. *Epidemiology*, 15, 615–625.
9. Rosenbaum, P. R. (1987) Model-based direct adjustment. *Journal of the American Statistical Association*, 82, 387–394.
10. Ridgeway, G. (1999) The state of boosting. *Computing Science and Statistics*, 31, 172–181.
11. McCaffrey, D., Ridgeway, G. & Morral, A. (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9 (4), 403–425.
12. Hernán, M. A., Brumback, B. & Robins, J. M. (2002) Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statistics in Medicine*, 21, 1689–1709.
13. D'Agostino, R. B., Lee, M.-L. & Belanger, A. J. (1990) Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statistics in Medicine*, 9, 1501–1515.
14. McCullagh, P. & Nelder, A. (1989) Generalized Linear Models, 2nd edn. London: Chapman and Hall.
15. White, H. A. (1980) A heteroscedasticity-consistent covariance matrix estimator and a direct test of heteroscedasticity. *Econometrica*, 48, 817–838.
16. Rabe-Hesketh, S. & Skrondal, A. (2008) Multilevel and Longitudinal Modeling Using Stata, 2nd edn. College Station, TX: Stata Press.

17. Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. (2004) Applied Longitudinal Analysis. Hoboken, NJ: John Wiley & Sons Inc.
18. Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 1096–1120 (with Rejoinder, 1135–1146).
19. Robins, J. M. (2000) Robust estimation in sequentially ignorable missing data and causal inference models. Proceedings of the American Statistical Association Section on Bayesian Statistical Science, 6–10.
20. Bang, H. & Robins, J. M. (2005) Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962–973.

