

Oops, I D-I-D It Again! Advanced Difference-in-Differences Models in SAS®

E Margaret Warton and Melissa M Parker, Kaiser Permanente Division of Research

ABSTRACT

The quasi-experimental, longitudinal difference-in-differences (D-I-D) study design is increasingly used in epidemiological and healthcare research. D-I-D models generate a causal estimate of the change in an outcome due to an intervention or exposure, after subtracting the expected background change observed in a reference group. Advantages of this method include preservation of time-ordering, accounting for changes in secular trends and regression to the mean, and, when a fixed cohort is used, controlling for unmeasured confounding. In this companion to our 2016 Western Users of SAS conference paper “How D-I-D You Do That? Basic Difference-in-Differences Models in SAS®”, we introduce advanced difference-in-differences methods and present a practical, step by step approach for implementation in SAS. Topics covered include: power and sample size considerations, options for modeling a binary outcome, balancing case mix differences in exposed and reference groups, and assessing heterogeneity of treatment effects. We will illustrate these methods using data from a study on the impact of introducing a value-based insurance design (VBID) medication plan at Kaiser Permanente Northern California on changes in medication adherence.

INTRODUCTION

Difference-in-Differences (D-I-D) study designs are becoming more commonly used in healthcare and other epidemiological research. This quasi-experimental longitudinal cohort study design analyzes changes in pre- and post-exposure outcomes between exposed and control groups, and so it is often used to analyze observational data resulting from a "natural experiment" when a suitable control group can be identified. Benefits of the study design include preservation of time-ordering, accounting for changes in secular trends, and, when a fixed cohort is used, controlling for both measured and unmeasured confounding. In addition, data can be collected prospectively or retrospectively, and the results can be interpreted causally, so long as time-ordering is maintained. In a previous paper (Warton, Parker, and Karter 2016), we introduced the basics of implementing and interpreting results of Difference-in-Differences models with a continuous outcome in SAS® software. In this subsequent paper, we build on the prior paper, introducing power and sample size calculation methods, analysis of D-I-D studies with binary outcomes, use of propensity scores for case-mix adjustment between exposed and control groups, and implementing and analyzing Difference-in-Difference-in-Differences (DDD) models using SAS.

A BRIEF REVIEW OF D-I-D BASICS

As the name suggests, difference-in-differences studies are concerned with whether the pre/post change in an outcome is different between two exposure groups. The basic equation for the model is

$$\mu_{it} = \beta_0 + \beta_{\text{post}} * \text{Post} + \beta_{\text{exp}} * \text{Exposure} + \beta_{\text{interaction}} * \text{Post} * \text{Exposure} + \epsilon_{it}$$

where μ_{it} is the expected mean value for subject i at time t , **Post** is a binary indicator that the outcome measurement was made in the post period, **Exposure** is a binary indicator that the subject is exposed to the intervention or treatment during the study, and ϵ_{it} is the error term for the outcome measure of subject i at time t . As usual, errors are assumed to be normally distributed with a mean of zero.

In Figure 1, the post minus pre change in the unexposed group (dashed line) is A2-A1, or ΔA , while the post minus pre change in the exposed group (solid line) is B2-B1, or ΔB . Matching with the figure, the difference in the changes between the two groups can be represented mathematically as

$$\begin{aligned} \Delta B - \Delta A &= ((\beta_0 + \beta_{\text{post}} + \beta_{\text{exp}} + \beta_{\text{interaction}}) - [\beta_0 + \beta_{\text{exp}}]) - ((\beta_0 + \beta_{\text{post}}) - \beta_0) \\ &= (\beta_0 + \beta_{\text{post}} + \beta_{\text{exp}} + \beta_{\text{interaction}} - \beta_0 - \beta_{\text{exp}}) - (\beta_0 + \beta_{\text{post}} - \beta_0) \\ &= (\beta_{\text{post}} + \beta_{\text{interaction}}) - (\beta_{\text{post}}) \\ &= \beta_{\text{interaction}} \end{aligned}$$

where β_0 is the intercept in the unexposed group in the pre period, β_{exp} is the difference between the unexposed and the exposed intercepts in the pre period, β_{post} is the change between pre and post periods in the unexposed group, and $\beta_{interaction}$ is the difference in the slopes of the pre/post lines in the exposed compared to the unexposed. In a simple D-I-D study with a continuous outcome, the point estimate for the interaction term in the model answers the difference-in-differences question: How much more (or less) did the outcome in the exposed group change compared to the unexposed group over the same pre-post period?

A key aspect of the D-I-D design is that it includes time-ordered repeated measurements of the outcome of interest. Because outcomes are measured among the same individuals before and after the exposure without crossover, cohort members serve as their own controls, eliminating both measured and unmeasured confounding. In addition, because the outcome measures are time-ordered and the control group accounts for temporal trends, the results from a D-I-D analysis can be interpreted causally.

In a basic D-I-D study with a fixed cohort, the same study subjects are measured in both the pre and the post periods, thus care must be taken to account for the repeated outcome measures inherent in the design. Difference-in-differences data can be analyzed using ANOVA, generalized linear, or mixed models to account for repeated measures. Because the subjects serve as their own controls, if case-mixes are similar between the exposed and unexposed groups, no adjustment for additional confounders is required, and thus the model statement needs to include only the outcome measure, an exposure group indicator variable, a pre vs post period indicator variable, and an interaction term between the exposure and period variables. More complex D-I-D study designs can include adjustment for case-mix differences between exposure groups (generally through use of propensity scores) and comparison of outcome changes between exposure groups within specific strata (difference-in-difference-in-differences, or DDD).

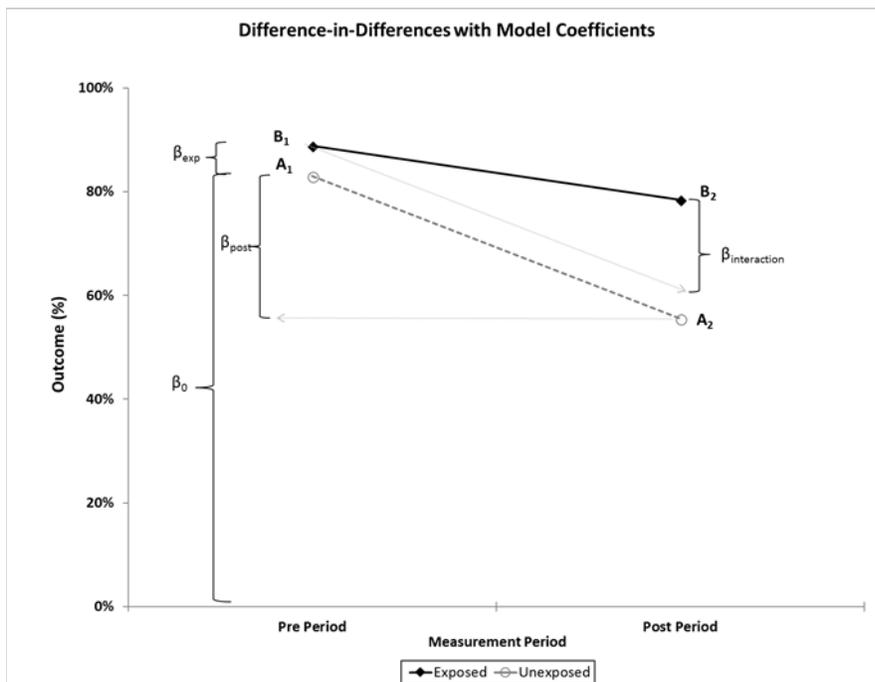


Figure 1. Basic Difference-in-Differences Study Design with Model Coefficients

EXAMPLE DATASET

As in our previous paper, we will use data from a study we performed on the impact of a Value-Based Insurance Design (VBID) pharmacy benefit on medication non-adherence among patients with chronic conditions. Adherence to continuously prescribed medications for diabetes, hypertension and hyperlipidemia is often measured using electronic data on pharmacy fills and refills. For this study, we calculated continuous medication gaps (CMG) which is the percentage of time during a calendar year that

a patient did not have adequate medication on hand. In addition to calculating medication non-adherence, we collected data on patient sex, age, neighborhood socioeconomic status (based on census data), race/ethnicity, preferred language, number and type of chronic conditions, medical facility, and annual out-of-pocket costs for medications.

We identified a cohort of all patients on a non-deductible plan during 2013 whose employers switched to a deductible plan at the beginning of 2014. Our exposed cohort consisted of those members who were offered a new VBID benefit in 2014 as part of the switch. Members of the comparison cohort switched to a deductible plan in 2014 but were not provided with a VBID prescription drug benefit. We then compared medication non-adherence among those with and without the VBID benefit to capture the change between 2013 and 2014. With the D-I-D design, we could identify changes in medication non-adherence due to the VBID plan while removing the influence of the underlying change to a deductible plan and adjusting for confounding, measured or not, at the subject level.

POWER AND SAMPLE SIZE CALCULATIONS

BASIC STEPS

The D-I-D study design can be analyzed using repeated measures ANOVA, generalized linear models with a REPEATED statement (using the GENMOD procedure) or a mixed model with random intercepts (using the MIXED or GLIMMIX procedures), meaning that power calculations for D-I-D models can be performed using the GLMPower procedure in SAS. This procedure is available starting in SAS/STAT® 13, and it calculates power estimates for general linear models including models with repeated measures. Here are the basic steps to performing a power or sample size calculation with PROC GLMPower:

1. Specify the model(s) to be used in the analysis
2. Generate estimates of outcome values and exposure group ratios
3. Estimate standard deviations for outcome measures and correlation between pre and post measures (usually based on pilot data or prior studies)
4. Create an "exemplary dataset" for use with PROC GLMPower
5. Use the exemplary dataset and estimates of standard deviations and correlation to perform prospective power and/or sample size calculations.

Generalized linear or mixed models are generally preferred over the repeated measures ANOVA for D-I-D analysis because both provide parameter estimates. Note that in the D-I-D context, however, all three underlying models are identical. Following the example of Castelloe (Castelloe 2014), we will use PROC MIXED for the power and sample size calculation examples. However, we will demonstrate the PROC GENMOD coding in our later analytic examples.

Step 1: Specifying the Model

For this example, we choose the basic model specification for the planned D-I-D analysis using PROC MIXED:

```
Title 'D-I-D Analysis Power Calculation Planned Model';  
proc mixed data=vbid method=reml;  
  class study_id exposed;  
  model non_adh = post|exposed / solution ddfm=kr;  
  random int/subject=study_id type=un ;  
run;
```

For more detail on using PROC MIXED for D-I-D analysis, see our previous paper (Warton, Parker, and Karter 2016).

Step 2: Estimating Pre and Post Outcomes and Sampling Ratios

We now need to estimate the mean non-adherence values for our pre and post measures in the exposed and unexposed groups as well as our sampling ratio. For this exercise, we assume that we will have two VBID-exposed cohort members for each three non-exposed cohort members (our study actual cohort has a ratio of 1024:1458 as determined during grant preparation). We estimate, based on past studies, that non-adherence (as measured by CMG) before the change to a deductible plan will be 40% (0.40). We want to determine the power to see an increase in non-adherence to 45% (0.45) in those without the VBID plan vs an increase to 42% (0.42) in those with the VBID plan in the post period.

Step 3: Correlation and Standard Deviation Estimates

Prior work with electronic pharmacy data indicates that non-adherence in one year is strongly correlated with non-adherence in the following year. For this power calculation, we estimate a correlation in adherence of 0.8 between consecutive years. Also based on prior data, we expect a standard deviation of 0.30 in measures of CMG.

Step 4: Creating an Exemplary Dataset

Using the techniques nicely presented in John Castelloe's 2014 paper on using PROC GLMPower in repeated measure models (Castelloe 2014), we create a dataset containing the parameters needed for our power and sample size calculations. The variables `non_adh_pre` and `non_adh_post` indicate the mean percent of time without medication on hand in each group in the pre and post periods, respectively, while the `wt` variable captures the ratio between unexposed and exposed (1458:1024 or about 3:2 in our cohort). Based on our estimates in Step 2, we can now create the exemplary dataset containing estimates of mean values and sampling weights for the exposed and unexposed groups to be used in the power calculation:

```
data adh;
input exposed $ non_adh_pre non_adh_post wt;
datalines;
    unexposed .40 .45 3
    exposed .40 .42 2
;
```

Step 5a: Calculating Required Sample Size with PROC GLMPower

For this example, we want to calculate the required sample size at 90% power to identify a D-I-D of 0.03 (or 3%) larger change in mean non-adherence in the non-VBID group compared to the VBID group, under the hypothesis that free medications will offset the increase in non-adherence caused by the rise in deductibles. We obtain these results using PROC GLMPower, setting the POWER option at 0.90:

```
title "VBID D-I-D Required Sample Size: PROC GLMPower";
proc glmpower data = adh;
    class exposed;
    weight wt;
    model non_adh_pre non_adh_post = exposed;
    repeated time;
    power
        mtest = hlt
        alpha = 0.05
        power = 0.9
        ntotal = .
        stddev = 0.3
        matrix("adhcorr") = lear (0.8, 1, 2, 1 2)
        corrmat = "adhcorr";
run;
```

Using our **adh** exemplary dataset, we identify the exposure variable in the CLASS statement and include the ratio of unexposed to exposed using the WEIGHT statement. Our MODEL statement is formatted with the two outcome measures on the left and the exposure after the equals sign. The REPEATED statement with a value of "time" indicates that the outcomes are measured at two different timepoints. In the POWER statement, the HLT value in the MTEST option indicates that we are using the Hotelling-Lawley trace statistic for the overall model F-test. We are requesting sample sizes for a power level of 0.9 with an alpha of 0.05 in the POWER and ALPHA options, respectively, and assuming a standard deviation in our cohort of 0.30 via the STDDEV option. The MATRIX option identifies our correlation structure using the "Linear Exponent AR(1)" or LEAR model method (Simpson et al. 2010). The first value in the parentheses indicates the correlation between measures, the second indicates the decay rate (if any) followed by the number of time points (just 2 for a standard DID model) and the time values (year 1 and year 2). The final CORRMAT option indicates that the "adcorr" matrix defined in the MATRIX statement should be used to account for correlation between year 1 and year 2 measures. The key result from the PROC GLMPower output is circled in Output 1.

Since the D-I-D result of interest is the interaction term between time and exposure, we see that a total sample size of 1755 is adequate to detect a 3% difference-in-differences change in non-adherence between the VBID and non-VBID groups at 90% power. At a 3:2 ratio of unexposed to exposed, we will need 1053 unexposed and 702 exposed cohort members (at a minimum) for this D-I-D analysis as specified.

Computed N Total							
Index	Transformation	Source	Effect	Num DF	Den DF	Actual Power	N Total
1	Time	Intercept	Time	1	323	0.901	325
2	Time	exposed	time*exposed	1	1753	0.900	1755
3	Mean(Dep)	Intercept	Intercept	1	8	0.952	10
4	Mean(Dep)	exposed	Exposed	1	15763	0.900	15765

Output 1 : PROC GLMPower Sample Size Output for VBID D-I-D Study

Step 5b: Calculating Power with PROC GLMPower

If we already have an idea of the potential sample size and the sampling ratio, we can perform a prospective power analysis by setting the POWER option to missing and including the total available cohort size using the NTOTAL option in the PROC GLMPower code. In this case, we estimate a total sample size of 2,482 divided into unexposed and exposed in the previously mentioned 3:2 ratio. Alternatively, we could change the **wt** variable in the ADH dataset to the actual exposed and unexposed sample size numbers, but since our ratio is quite close to 3:2, we will leave it as is. Using the existing **adh** dataset, we make these two small changes to our PROC GLMPower code from above to calculate the power level for this analysis:

```

title "VBID D-I-D Power Calculation: PROC GLMPower";
proc glmpower data = adh;
  class exposed;
  weight wt;
  model non_adh_pre non_adh_post = exposed;
  repeated time;
  power
    mtest = hlt
    alpha = 0.05
    power = .
    ntotal = 2482
    stddev = 0.30
    matrix("adhcorr") = lear (0.8, 1, 2, 1 2)
    corrmatrix = "adhcorr";
run;

```

The results in Output 2 indicate that with our estimated sample size, we will have more than adequate power to detect a difference-in-differences of 0.03 (3%) in adherence.

Computed Power						
Index	Transformation	Source	Effect	Num DF	Den DF	Power
1	Time	Intercept	Time	1	2478	>.999
2	Time	exposed	time*exposed	1	2478	0.971
3	Mean(Dep)	Intercept	Intercept	1	2478	>.999
4	Mean(Dep)	exposed	Exposed	1	2478	0.251

Output 2 : PROC GLMPower Power Calculation Results for VBID D-I-D Study

D-I-D MODELS WITH BINARY OUTCOMES

While it is more intuitive to think of D-I-D models as comparing the change in a continuous outcome over time, D-I-D models are also used in studies with binary outcomes. The basic D-I-D design compares group means in two groups at two different time points. When the outcome measure is continuous, the group means are simply the mean values of the outcome in each group at each of the two time points. Using a mixed model with random intercepts means that each individual in a given exposure group has an intercept that is allowed to vary around the group mean. However, the slopes of individual members' trajectories in a given exposure group are constrained to match the slope of all other group members, which is equal to the slope between the pre and post period group means. When the outcome measure is binary, the pre and post period group means represent the proportion of individuals in each group experiencing the outcome at each of the study time points. These proportions can be estimated as odds or probabilities on the log scale (relative risks) or as proportions (absolute risks) for each group at each time point. Note that in the binary case, the values of the outcome measure are constrained to 0 and 1 for all individuals. Therefore, a random-intercept mixed model for a binary outcome with only two time points is equivalent to a population-level marginal model and can be modeled using PROC GENMOD.

When deciding on a model for a D-I-D analysis with a binary outcome, there are two main questions to answer: First, how common is the outcome of interest? Second, do you want to report absolute or relative risks? For very rare outcomes, the odds ratio from a logistic regression model is an unbiased estimate of the relative risk. For more common outcomes, however, relative risks should be directly estimated using log-binomial or log-Poisson models (Zou 2004; Spiegelman and Hertzmark 2005). In these models, the *log* link is specified, and so coefficient estimates are on the log scale. This means that the difference between relative risks in the exposed and unexposed groups (the D-I-D in relative risks) is calculated by dividing the risk ratio in the exposed group by the risk ratio in the unexposed group. This result is referred to as the "Relative" D-I-D, the ratio of rate ratios, or the RRR. If the magnitudes of the underlying outcome rates are of peripheral interest with the main analytic focus being on the relative change between groups, then calculating relative risks and RRR is a reasonable analytic method.

Usually, however, the magnitudes of the pre to post changes in risk of the outcome between the exposure groups are of interest, and so estimates of absolute risk are required. In these cases, a binomial or Poisson model with the *identity* link yields estimates of absolute risk differences and prevalences in the exposure groups in the pre and post periods. Most of the following examples are based on estimating the absolute risk difference, as it is a more intuitive measure to interpret with a binary outcome.

EXAMPLE D-I-D MODELS FOR BINARY OUTCOMES

Model 1: Binomial Distribution with Identity Link for Absolute Risk Differences

Unfortunately, medication non-adherence is quite common in this study cohort; on average, over 40% of the cohort has inadequate medication on hand for more than 20% of the year. Reporting rates of non-adherence is integral to the study analysis, and so a model that produces estimates of absolute risk differences is desirable. A marginal population-level model (General Estimating Equations, or GEE) generates the same estimates as a mixed model in this case, and so the first model choice in this situation is a binomial model using the identity link with PROC GENMOD. As described in our previous

paper, this procedure requires a dataset in "long" format with two observations (one for each time point) per individual:

```

title "Population-Level Marginal D-I-D Model:PROC GENMOD with dist=binomial
and link=identity";
proc genmod data=vbid descending;
  class study_id exposed (order=formatted ref='non-vbid') / param=ref;
  model non_adh = post|exposed /dist=binomial link=identity type3;
  repeated subject=study_id /type=un;
  format exposed exposed.;
run;

```

In this model specification, we include both **study_id** and **exposed** variables in the class statement. We could also include the **post** variable in the CLASS statement if we wanted to specify the reference level. The MODEL statement includes the standard shorthand for all variables and interaction terms, specifies that the outcome variable distribution is binomial in the DIST option, and that the desired identity link is identified in the LINK option. The TYPE3 option requests Type 3 effects tests which are replaced with joint tests for effects in this model because all combinations of the predictor variables are present in the model (it is fully parameterized).

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept		0.4060	0.0129	0.3808	0.4312	31.57	<.0001
post		-0.0123	0.0106	-0.0332	0.0085	-1.16	0.2451
exposed	VBID	0.0129	0.0201	-0.0264	0.0523	0.64	0.5202
post*exposed	VBID	-0.0170	0.0166	-0.0495	0.0156	-1.02	0.3076

Output 3: D-I-D Coefficient Estimates from PROC GENMOD Model with Binomial Distribution and Identity Link

The results from this model shown in Output 3 (also summarized in Table 1) indicate that the change in non-adherence in the VBID group was 1.7% lower than in the non-VBID group, hinting that the VBID benefit may have helped reduce the impact of the plan deductible change. However, the magnitude of this change is quite small, and it is less than the 3% D-I-D hypothesized in the power calculation. With a p-value of 0.31, therefore, we cannot say that there is a significant difference in adherence change between the two groups.

Model 2: Poisson Distribution with Identity Link for Absolute Risk Differences

It is often the case that binomial models do not converge, and in this situation, using the Poisson model for binary outcomes is an equivalent alternative. The model specification is identical to the binomial model, except that the "binomial" value specified in the DIST option of the MODEL statement is changed to "poisson":

```

title "Population-Level Marginal Model:PROC GENMOD with dist=Poisson and
link=identity";
proc genmod data=vbid descending;
  class study_id exposed (order=formatted ref='non-vbid') / param=ref;
  model poor_adh= post|exposed /dist=poisson link=identity type3;
  repeated subject=study_id /type=un;
  format exposed exposed.;
run;

```

As expected, the results from this model are identical to those produced by the binomial model (Table 1).

Model 3: Binomial Distribution with Log Link for Relative Risk Ratio (RRR)

If the magnitude of outcome rates is not of interest in a D-I-D analysis, relative risk ratios may be obtained easily by modeling a binomial or Poisson distribution with the log link. In the code below, the ESTIMATE statements are used to generate exponentiated coefficients which are the relative risks for each group. The relative risk for the interaction term is the Relative Risk Ratio, often abbreviated as RRR:

```

title "Population-Level Marginal Model (RRR) : PROC GENMOD with dist=binomial
and link=log";
proc genmod data=vbid descending;
  class study_id exposed (order=formatted ref='Non-VBID') / param=ref;
  model non_adh= post|exposed /dist=binomial link=log type3;
  estimate 'RR:EXPOSED' exposed 1 /exp;
  estimate 'RR:POST' post 1 /exp;
  estimate 'RRR:POST*EXPOSED' post*exposed 1 /exp;
  repeated subject=study_id /type=un;
  format exposed exposed.;
run;

```

The results from this model look very different from the first two models (see Table 1), although the conclusions drawn from the results are similar: the value of RRR for non-adherence of 0.959 indicates that the relative change in non-adherence in the VBID group is slightly less than the relative change in non-adherence in the non-VBID group. However, this RRR does not reach a level of statistical significance in this cohort, and so we cannot say that the VBID group trajectory differs from that of the non-VBID group. Note that the intercept term in this model does match the other models. The intercept term in all of the models estimates the mean rate of non-adherence in the unexposed group during the pre-baseline period and is thus the same across all model specifications.

Also note that if the outcome of interest was quite rare, then the RRR could be estimated with the Example 3 code by simply changing the link = log statement to link = logit in the code above. However, with both log-binomial and log-Poisson options available to directly calculate relative risk, there is little reason to do so.

Marginal Population Model	Binomial Distribution Identity Link Absolute Risks		Poisson Distribution Identity Link Absolute Risks		Binomial Distribution Log Link Relative Risk Ratios (Exponentiated)	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Non-Adherence						
Intercept	0.406	(0.381, 0.431)	0.406	(0.381, 0.431)	0.406	(0.382, 0.432)
Post	-0.012	(-0.033, 0.009)	-0.012	(-0.033, 0.009)	0.970	0.920, 1.021
Exposed (VBID)	0.013	(-0.026, 0.052)	0.013	(-0.026, 0.052)	1.038	(0.938, 1.135)
Exposed*Post	-0.017	(-0.050, 0.016)	-0.017	(-0.050, 0.016)	0.959	(0.885, 1.040)

Table 1: Results from PROC GENMOD Marginal Population Models for Non-Adherence in VBID vs Non-VBID Cohorts

BALANCING CASE-MIX DIFFERENCES

One of the advantages of using D-I-D pre-post design in a fixed cohort is that each subject serves as his or her own control, eliminating both measured and unmeasured confounding. However, case mix differences between the exposed and reference groups pose a threat to validity if group characteristics are so dissimilar that changes in the reference group over time are not representative of secular trends in the absence of the exposure. While there is no formal way to actually test this, D-I-D rests on the assumption that the outcome in the exposed group would be the same as seen in the unexposed group if not exposed to the intervention (i.e. the counterfactual) That is, if the group exposed to the changes in the

benefits, had (counterfactually) not been exposed to those benefit changes, we assume the change in medication non-adherence would be similar to that observed in the reference group.

Therefore, it is very important to plan carefully in the design stage to ensure minimal differences between exposed and reference groups. Even with careful planning in a natural experiment or observational study, residual imbalances in case-mix are likely. Common ways to deal with this include adjusting for key covariates in a multiple regression or matching on individual variables that are unbalanced. These may be effective techniques if there are a few imbalances. However, matching on more than 2 or 3 individual variables may not be feasible, and adjusting for many covariates in multiple regression uses degrees of freedom, may reduce statistical power, and constrains the analysis to a conditional model instead of a marginal estimate. This may or may not align with research goals. Conventional regression adjustment estimates the average effect if, at the individual level, a subject's status was changed from exposed to unexposed. In general, quasi-experimental, longitudinal designs are interested in estimating population level marginal effects, the average effect in two populations that are exactly the same in every respect except that one received treatment and one did not, because this mimics the results of a randomized controlled trial (Austin 2011a).

PROPENSITY SCORES AND D-I-D ANALYSIS

One increasingly popular technique in observational research which minimizes case mix differences, allows the estimation of population level marginal effects, and can be implemented within the D-I-D framework is the two-stage propensity score analysis. The first stage of the analysis is to develop a propensity score -- the conditional probability of receiving the treatment or intervention given a set of observed covariates. In the second stage, the propensity score is used to match, stratify, weight or as an individual covariate in the D-I-D model. The practical advantages of propensity scores over covariate adjustment in multiple regression have been comprehensively described by Austin (Austin 2011a).

Stage 1: Developing a Propensity Score Model and PROC PSMATCH

In the first stage of analysis, the analyst must develop a propensity score model. The goal here is to develop a propensity score in which the entire distribution of baseline covariates is independent of exposure status. The propensity score is commonly calculated via a logistic regression, regressing the binary exposure as a function of baseline covariates. Creation of the propensity score is generally a long and iterative modeling process which ends only when a satisfactory covariate balance has been achieved. Machine learning techniques, such as data-adaptive model selection procedures that use cross validation to search through a large space of polynomial models and select the best predictive algorithm, have the potential to streamline the process. These techniques are theoretically superior to iterative approaches because they reduce bias in model specification and increase efficiency, but they require long computing times and high performance hardware with large memory storage capabilities. (Neugebauer, Schmittiel, and van der Laan 2016)

There is a rich literature describing which covariates are appropriate to include in these models. For a good discussion see (Austin 2011a, 2011b). One important concept to keep in mind is to respect time ordering. No variable that was measured AFTER baseline (e.g. the pre or post outcomes of interest) should be used in creating the propensity score model for a D-I-D analysis.

In addition, the analyst must decide how to use the propensity scores in the D-I-D analysis. The four basic options include: (1) Matching--use the propensity score variable to match individual reference subjects to exposed subjects, (2) Stratification—run stratified analyses, usually by quintiles of the propensity score, and then aggregate the results for a weighted treatment effect, (3) Weighting-- use the inverse of the propensity scores as weights to estimate the average treatment effect (ATE) or average effect in the treated (ATT) or (4) Covariate adjustment-- include the propensity score as an individual covariate in the D-I-D model. For a good example illustrating the pros and cons of each method see (Austin 2011b). Discussion of these issues is beyond the scope of the current paper, but empirical observations and Monte Carlo simulations show that matching and weighting are more effective at balancing differences than stratification or covariate adjustment (Austin 2011b).

Once a propensity score model has been developed, the next step is to assess whether satisfactory covariate balance has been achieved between exposed and reference groups via matching, stratification

or weighting. There are several statistical tools available for this assessment; in this paper we present results using standardized differences, ratio of variances and some graphical methods. Briefly, standardized difference is a measure of the difference in means between the exposed and reference groups (or prevalence in the case of a binary categorical variable) divided by the pooled standard deviation. Its domain includes all real numbers. A standardized difference of zero indicates no differences between exposure groups, and the farther away from zero, the greater the differences between groups. While there is no consensus, many accept the criteria of a standardized difference between -0.10 and 0.10 as indicating a negligible difference in the mean or prevalence of a covariate between groups (Austin 2011a). Standardized difference is not influenced by sample size as p-value is. In a large sample, even very small, clinically insignificant differences may be statistically significant using the p-value criteria <0.05 . This limitation is avoided by using standardized differences to assess covariate imbalance (Austin 2011a). In a balanced cohort, the ratio of variances for any individual covariate across exposure status should ideally be 1.0. Examining box plots, quantile-quantile plots, non-parametric density plots or empirical CDF plots by exposure status can also help detect anomalies in the distributions and aid in evaluation of the adequacy of the model. For more in-depth discussion of these methods, see Austin's publications (Austin 2009, 2011a).

The SAS/STAT[®] 14.3 release includes the PSMATCH procedure which facilitates specification of the propensity score model, assesses covariate balance, and can create propensity score weights or a matched sample. To use PROC PSMATCH, all class variables must be coded as binary dummy variables. In addition, consideration must be given to missing data (should there be any). Observations with missing data are discarded by the procedure; therefore, the analyst should create missing indicators and impute values as appropriate. The CLASS statement and either a PSMODEL or PSDATA statement are required for PROC PSMATCH. The binary exposure status variable (which is the outcome for the propensity score model) must appear on the CLASS statement. The PSMODEL statement specifies the dependent and independent variables for the first stage propensity score model. If you have already created propensity scores in a different program, you can use the PSDATA statement to read in precomputed propensity scores. The ASSESS statement assesses variable differences between the exposed and reference groups and is very helpful for the first stage of analysis. In the event you would like to create a matched sample, you can specify the MATCH statement. Several matching algorithms are available including greedy (nearest neighbor), optimal matching and matching with or without replacement. The OUTPUT statement creates an output dataset for all (or matched only) observations with individual propensity scores and weights. If you do not have the latest version of SAS/STAT available but wish to assess covariate balance using standardized differences, a SAS user created macro %stddiff (Yang and Dalton 2012) is available on the internet.

In our example, we ran the following code in PROC PSMATCH to develop a model predicting enrollment in the VBID high deductible insurance plan as a function of all baseline covariates in the full sample of 2482 patients. This code creates an output dataset named **outx** that contains all observations as well as variables for the propensity score, logit propensity score, and weights for the average treatment effect (ATE). On the ASSESS statement we specified that we would like balance diagnostics for the cohort before and after weighting by the ATE, meaning we are trying to weight each exposure group to match the overall population covariate distribution. The VAR option requests diagnostics for specific covariates. Note that we include the reference categories for race (**white**) and comorbidity count (**comorbid0**) in this list, but they are omitted from the PSMODEL statement. Statistics for the propensity score (ps) and logit propensity score (lps) are also requested. The VARINFO option on the ASSESS statement outputs the N, mean, standard deviation, minimum and maximum values by treatment group for each variable specified. It also outputs the treated minus control mean difference. The NLARGESTWGT=6 option prints out the observation number and values of the six largest weights in both the treated and control groups, which is helpful in case we decide to truncate extreme outliers. The PLOTS option specifies plots to output; in this case we request cumulative distribution and standardized difference plots for all variables and boxplots for the propensity score and **age_at_index** variables. The REF=0.10 option following STDDIFF places reference lines at -0.10 and 0.10 on the STDDIFF plot making it very easy to see where balance is inadequate; the default in SAS is -0.25 and 0.25. The NODETAILS option following PLOTS suppresses output on the plots for the common support region; figures are displayed for all observations and the weighted cohort only. The WEIGHT option tells SAS to use weights to calculate the average treatment effect (ATE). Table 2 was created exclusively using output from this code:

```

title "VBID D-I-D: PROC PSMATCH for Propensity Scores";
proc psmatch data=VBID;
  class exposed;
  psmodel exposed (treated=last)= age_at_index female low_ses hispanic
    black asian other race_miss english comorbid1-comorbid3;
  assess ps lps var=(female white age_at_index low_ses hispanic black
    asian other english comorb0-comorb3)
  / varinfo nlargestwgt=6
  plots(nodetails)=(cdfplot boxplot(display=(ps age_at_index))
  stddiff(ref=0.10) )
  weight=atewgt(stabilize=no) ;
output out (obs=all)=outx
  lps=lps
  ps=ps
  atewgt(stabilize=no)=iptw_u ;
run;

```

Characteristic	All				ATE Weighted			
	VBID N=1024	No VBID N=1458	Std Diff	Variance Ratio	VBID N=1024	No VBID N=1458	Std Diff	Variance Ratio
Female	42%	48%	-0.12	0.978	45%	46%	-0.01	0.998
Age, mean (SD)	50.4 (8.3)	52.0 (8.0)	-0.19	1.076	51.4 (7.9)	51.3 (8.4)	0.01	0.883
Race								
Black	14%	12%	0.08	1.193	13%	13%	0.001	1.002
White	64%	57%	0.14	0.943	61%	60%	0.02	0.993
Asian	14%	23%	-0.23	0.690	19%	20%	-0.02	0.972
Other	5%	7%	-0.04	0.849	6%	6%	-0.007	0.974
Hispanic ethnicity	18%	15%	0.09	1.179	17%	16%	0.01	1.027
English speaker	99%	94%	0.32	0.100	97%	96%	0.05	0.778
Lower neighborhood SES	15%	22%	-0.16	0.765	19%	19%	-0.006	0.992
No. of comorbidities								
0	29%	28%	0.01	1.014	28%	28%	0.001	1.001
1	52%	53%	-0.03	1.004	52%	53%	-0.005	1.001
2	17%	17%	0.01	1.021	17%	17%	0.007	1.012
3+	3%	2%	0.04	1.282	2%	3%	-0.003	0.984
Propensity score, mean (SD)	0.45 (0.10)	0.39 (0.12)	0.50	0.623	0.42 (0.11)	0.41 (0.12)	0.03	0.911
Logit Propensity score, mean (SD)	-0.23 (0.43)	-0.53 (0.73)	0.50	0.347	-0.38 (0.55)	-0.40 (0.64)	0.05	0.742

Table 2: Assessment of covariate balance in Stage 1 of the Propensity Score Analysis

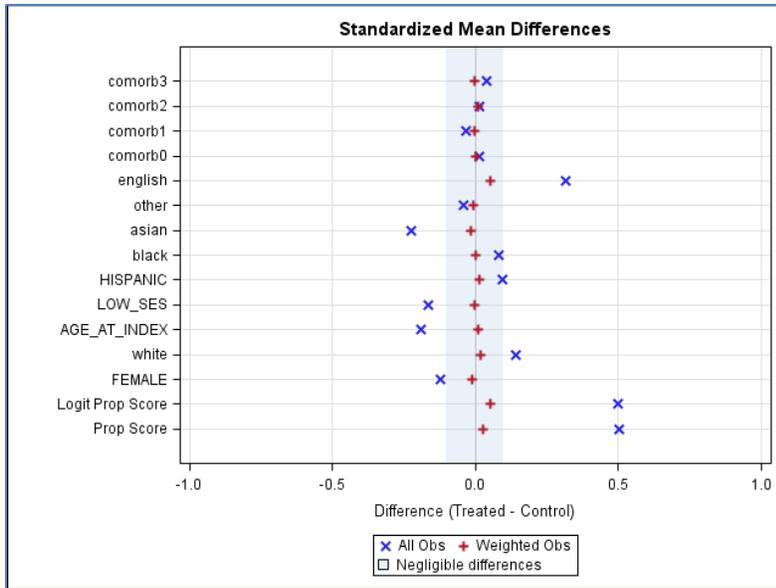


Figure 2: Standardized Mean Difference Plot from PROC PSMATCH

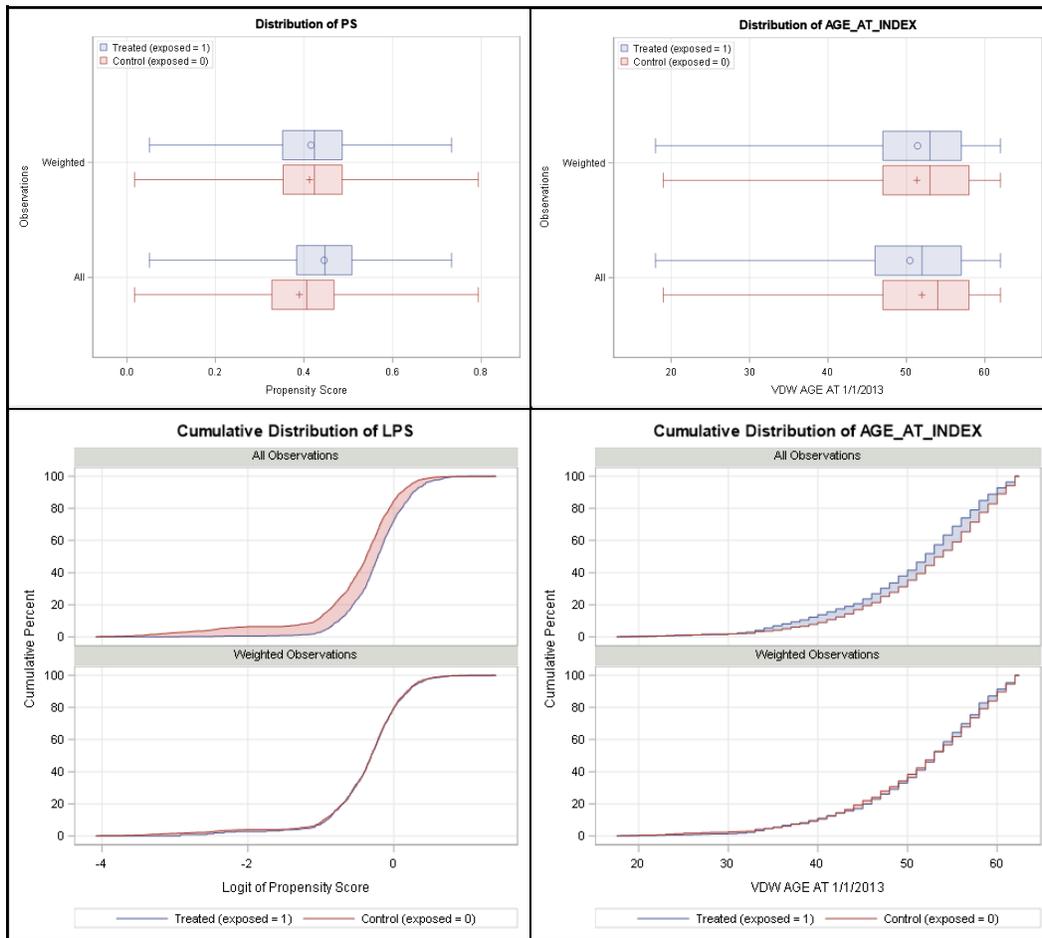


Figure 3: Boxplots and CDF Plots from PROC SMATCH (Partial Output)

Note the case mix differences in gender, age, race, ethnicity, English language proficiency and neighborhood SES in Table 2. In the full cohort, several of the standardized differences are ≥ 0.10 or ≤ -0.10 , suggesting unacceptable balance. In addition, the variance ratio is quite far from 1 for several variables. After weighting, all the imbalances are gone, based on our standardized difference criteria, and the variance ratio is improved (closer to 1) for all variables with the exception of age. Figure 2 visually confirms that after weighting, the differences in balance in the means or prevalence of the analyzed covariates are negligible, as all the standardized differences fall between -0.10 and 0.10.

The boxplots and cumulative distribution plots presented in Figure 3 aid in determining balance in the overall distribution of key continuous variables, including the propensity score and logit propensity score. The logit propensity score is more likely to be normally distributed and, for this reason, it is commonly used when calculating a caliper for propensity score matching. Matching on a single normally distributed variable with a caliper of 0.20 times the pooled standard deviation will eliminate approximately 99% of bias due to measured confounders (Austin 2011a). Note that balance between treated and control groups is greatly improved, and substantive differences are eliminated in the weighted cohort.

The first stage of propensity score analysis is the longest step as it is an iterative process; the covariates may not balance adequately at first. In real world data there are cases where you may never achieve the desired balance and will need to adjust in the analysis phase for residual case mix differences.

Stage 2: Including Propensity Scores in D-I-D Models

Now that we have created the propensity scores and weights and are satisfied with the covariate balance, we can move on to the second stage of the analysis and run the difference-in-differences model.

Here is the code to merge the propensity score weights back into the original data set. Remember that the propensity score model was run only on baseline data, so there should be only one record per subject in the **outx** data set. We sort and merge **outx** back with our **VBID** data set (2 records per subject, pre and post baseline) and rerun the PROC GENMOD code outlined above, but add a WEIGHT statement, to weight the analysis by the ATE weights (**iptw_u**):

```
proc sort
  data=outx (keep=study_id ps lps iptw_u)
  out=ps;
  by study_id;
run;

data xwt;
  merge VBID (in=a)
        ps;
  by study_id;
  if a;
run;

title "D-I-D model for change in medication adherence, Main effects model in ATE weighted cohort";
proc genmod data=xwt descending;
  class study_id exposed (order=formatted ref='Non-VBID') / param=ref;
  model non_adh = post|exposed /dist=binomial link=identity type3;
  repeated subject=study_id /type=un;
  weight iptw_u ;
  format exposed exposed.;
run;
```

Table 3 compares the D-I-D model results from the propensity score weighted model to the results presented earlier in the paper. As in the unadjusted model, we find no evidence that the change in poor medication adherence differs by VBID benefit status. Estimated non-adherence among the VBID group changes from 0.410 (0.408+0.002) before the deductible change to 0.375 (0.408-0.014+0.002-0.021) after the deductible change, representing a 3.5% improvement in non-adherence. The unexposed

reference group experienced a change from 0.408 before the deductible change to 0.394 (0.408-0.014) afterwards, representing a 1.4% improvement in non-adherence. Overall non-adherence decreased in both groups, but the absolute change was 2.1% less in the VBID group compared to the non-VBID group. Note that the intercept and post period coefficient estimates are similar between the unadjusted and propensity score weighted models, but that the estimates changed more dramatically for the terms involving the VBID exposed group. This suggests that the propensity score weighting adjusted for case mix differences resulting in a slightly stronger magnitude of effect, even though it was still not statistically significant.

	Unadjusted		ATE Weighted	
Non-Adherence	Estimate	95% CI	Estimate	95% CI
Intercept	0.406	(0.381, 0.431)	0.408	(0.382, 0.434)
Post	-0.012	(-0.033, 0.009)	-0.014	(-0.035, 0.008)
Exposed (VBID)	0.013	(-0.026, 0.052)	0.002	(-0.044, 0.044)
Exposed*Post (DID estimator)	-0.017	(-0.050, 0.016)	-0.021	(-0.057, 0.015)

Table 3: Comparison of DID model results from the unadjusted and ATE weighted models

DIFFERENCE-IN-DIFFERENCE-IN-DIFFERENCES

The concept of difference-in-difference-in-differences (DDD) is a mouthful but is nothing more than formally comparing two or more D-I-D estimates. This may be referred to as heterogeneity of treatment effects, effect modification or an interaction depending on your discipline. In the current example, since we did not find a significant effect in the main model, it is doubtful that we will find an effect in any particular subgroup. However, since we had a hypothesis *a priori* that the VBID plan would have a stronger effect among patients who face financial hardships, we will examine if the D-I-D among patients living in low SES neighborhoods is the same as the D-I-D among patients who do not live in low SES neighborhoods. Expanding on the D-I-D equation above, we add two additional terms to the DDD regression model: a term estimating the independent effect of SES level on change in medication non-adherence (β_{SES}) and an interaction between SES level, time and the exposure (β_{int2}).

$$Y_{it} = \beta_0 + \beta_{post} * Post + \beta_{exp} * Exposure + \beta_{SES} * SES + \beta_{int1} * Post * Exposure + \beta_{int2} * SES * Post * Exposure + \epsilon_{it}$$

If it fits within the conceptual framework, lower order interaction terms between the effect modifier and time and/or the effect modifier and exposure may also be added to the model. Nevertheless, the parameter estimate β_{int2} will always represent the difference-in-difference-in-differences estimate as shown below:

$$DDD = (\Delta B - \Delta A)_{SES=1} \text{ minus } (\Delta B - \Delta A)_{SES=0}$$

where ΔB is the post minus pre change in the exposed group and ΔA is the post minus pre change in the reference group.

$$(\Delta B - \Delta A)_{SES=1} = [(\beta_0 + \beta_{post} + \beta_{exp} + \beta_{int1} + \beta_{SES} + \beta_{int2}) - [(\beta_0 + \beta_{exp} + \beta_{SES})] - [(\beta_0 + \beta_{post} + \beta_{SES})] - (\beta_0 + \beta_{SES})]$$

$$= [(\beta_{post} + \beta_{int1} + \beta_{int2})] - [(\beta_{post})]$$

$$= \beta_{int1} + \beta_{int2}$$

$$(\Delta B - \Delta A)_{SES=0} = [(\beta_0 + \beta_{post} + \beta_{exp} + \beta_{int1}) - (\beta_0 + \beta_{exp})] - [(\beta_0 + \beta_{post}) - (\beta_0)]$$

$$= [(\beta_{post} + \beta_{int1})] - [(\beta_{post})]$$

$$= \beta_{int1}$$

$$(\Delta B - \Delta A)_{SES=1} - (\Delta B - \Delta A)_{SES=0} = (\beta_{int1} + \beta_{int2}) - (\beta_{int1})$$

$$= \beta_{int2}$$

ASSESSING HETEROGENEITY OF TREATMENT EFFECTS

Figure 4 visually depicts the research question of interest. In the low SES=Yes group, the prevalence of non-adherence drops from 45% to 43% in the exposed and 49% to 47% in the reference group. Therefore, the D-I-D in this group is 0% $[(43-45)-(47-49)]$. In the low SES=No group, the prevalence of non-adherence drops from 41% to 38% in the exposed and 38% to 37% in the reference group. The D-I-D in the SES=No group is -2% $[(38-41)-(37-38)]$. Thus, the unadjusted DDD is 2%. Model results will allow us to infer whether this is a statistically significant difference.

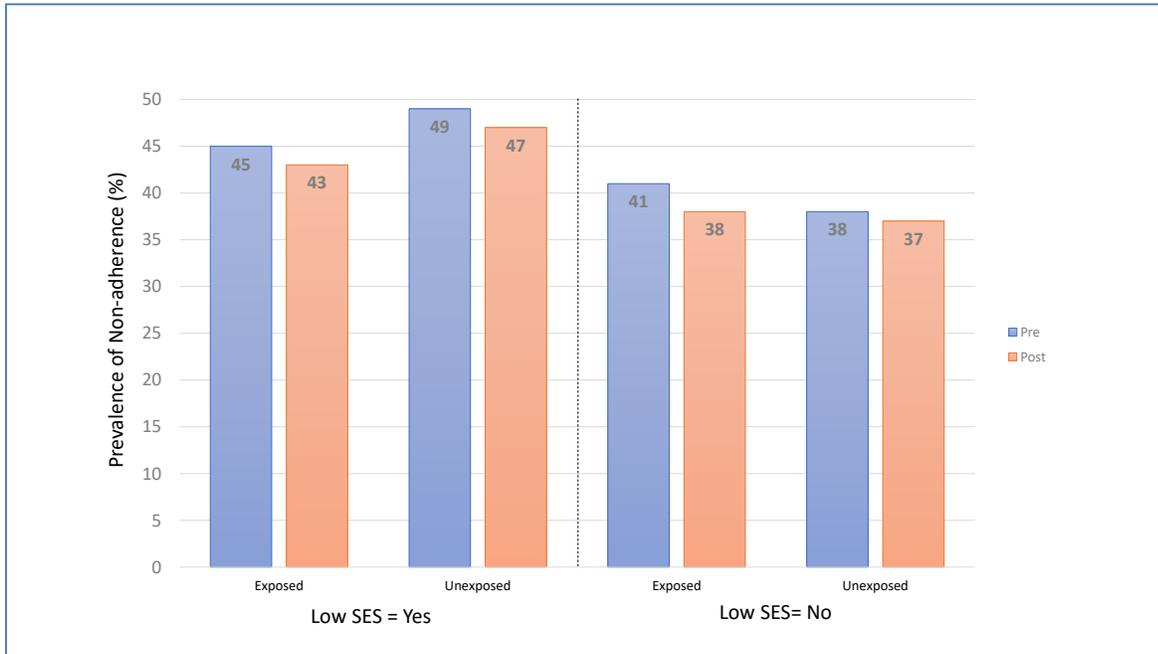


Figure 4 : Unadjusted Non-Adherence Rates in Exposed and Unexposed by Low SES Strata

A DDD model with a binary outcome can be run in SAS various ways. Our example uses PROC GENMOD with a binomial distribution and identity link. Partial output (Output 4) follows below.

```

title "D-I-D model for change in medication adherence, Testing for EM by
SES in ATE weighted cohort";
proc genmod data=xwt descending ;
  class study_id;
  model non_adh = post exposed post*exposed low_ses low_ses*post*exposed
    /dist=binomial link=identity type3;
  repeated subject =study_id /type=un;
  estimate 'low_ses=1, pre, VBID' intercept 1 exposed 1 low_ses 1 ;
  estimate 'low_ses=1, post, VBID' intercept 1 post 1 exposed 1
    post*exposed 1 low_ses 1
    low_ses*post*exposed 1 ;
  estimate 'low_ses=1, pre, NON-VBID' intercept 1 low_ses 1 ;
  estimate 'low_ses=1, post, NON-VBID' intercept 1 post 1 low_ses 1 ;
  estimate 'low_ses=0, pre, VBID' intercept 1 exposed 1 ;
  estimate 'low_ses=0, post, VBID' intercept 1 post 1 exposed 1
    post*exposed 1 ;
  estimate 'low_ses=0, pre, NON-VBID' intercept 1 ;
  estimate 'low_ses=0, post, NON-VBID' intercept 1 post 1 ;
  estimate 'low_ses=1, DID' post*exposed 1 low_ses*post*exposed 1 ;
  estimate 'low_ses=0, DID' post*exposed 1 ;
  weight iptw_u ;
run;

```

Analysis Of GEE Parameter Estimates						
Empirical Standard Error Estimates						
Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept	0.3933	0.0142	0.3655	0.4211	27.74	<.0001
Post	-0.0136	0.0108	-0.0348	0.0076	-1.26	0.2085
Exposed	0.0023	0.0212	-0.0392	0.0438	0.11	0.9125
post*exposed	-0.0148	0.0178	-0.0497	0.0201	-0.83	0.4064
LOW_SES	0.0747	0.0270	0.0219	0.1276	2.77	0.0056
post*exposed*LOW_SES	-0.0328	0.0465	-0.1240	0.0584	-0.71	0.4804

Contrast Estimate Results										
Label	Mean Estimate	Mean		L'Beta Estimate	Standard Error	Alpha	L'Beta		Chi-Square	Pr > ChiSq
		Confidence Limits	Confidence Limits				Confidence Limits	Confidence Limits		
low_ses=1, pre, VBID	0.4703	0.4133	0.5274	0.4703	0.0291	0.05	0.4133	0.5274	261.43	<.0001
low_ses=1, post, VBID	0.4091	0.3297	0.4885	0.4091	0.0405	0.05	0.3297	0.4885	101.99	<.0001
low_ses=1, pre, NON-VBID	0.4680	0.4186	0.5174	0.4680	0.0252	0.05	0.4186	0.5174	344.65	<.0001
low_ses=1, post, NON-VBID	0.4544	0.4050	0.5038	0.4544	0.0252	0.05	0.4050	0.5038	325.22	<.0001
low_ses=0, pre, VBID	0.3956	0.3627	0.4285	0.3956	0.0168	0.05	0.3627	0.4285	554.63	<.0001
low_ses=0, post, VBID	0.3672	0.3338	0.4006	0.3672	0.0171	0.05	0.3338	0.4006	463.67	<.0001
low_ses=0, pre, NON-VBID	0.3933	0.3655	0.4211	0.3933	0.0142	0.05	0.3655	0.4211	769.23	<.0001
low_ses=0, post, NON-VBID	0.3797	0.3520	0.4073	0.3797	0.0141	0.05	0.3520	0.4073	724.76	<.0001
low_ses=1, DID	-0.0476	-0.1391	0.0438	-0.0476	0.0467	0.05	-0.1391	0.0438	1.04	0.3074
low_ses=0, DID	-0.0148	-0.0497	0.0201	-0.0148	0.0178	0.05	-0.0497	0.0201	0.69	0.4064

Output 4 : PROC GENMOD Difference-in-Difference-in-Differences Model Results

In Output 4, the parameter estimate for the **post*exposed*low_ses** interaction is the DDD estimator. We find no significant difference between the D-I-D among those with **low_ses=1** and those with **low_ses=0** (p-value=0.48). The ESTIMATE statements generate the prevalence of non-adherence for each level of **low_ses** at each time point in both exposed and reference groups and also allow the calculation of the D-I-D estimates for each level of the effect modifier. Non-adherence dropped in patients with low SES by 4.8% (-0.0476) while non-adherence dropped by 1.5% among patients without low SES (-0.0148). The difference between those two estimates (3.3%) is the DDD (-0.0328), which is the parameter estimate for the **post*exposed*low_ses** interaction. Note that the D-I-D estimates for both SES groups (particularly for **low_ses=1**) changed in this propensity score weighted model compared to the *unadjusted* rates presented in Figure 4. As in our previous example, this suggests that the propensity score weighting adjusted for case mix differences resulting in a slightly stronger magnitude of effect, although it remained statistically insignificant.

CONCLUSION

This paper introduces four advanced topics in difference-in-differences methodology: power and sample size considerations, options for modeling a binary outcome, balancing case mix differences in exposed and reference groups, and assessment of heterogeneity of treatment effects. We present a practical, step

by step approach for implementation in SAS® that has led to successful publications for our research group. For some examples in peer reviewed healthcare research journals, we refer readers to (Reed et al. 2017) for D-I-D with a continuous outcome, (Parker et al. 2017) for D-I-D with binary outcome, and (Karter et al. 2015) or (Lipska et al. 2018) for DDD and propensity scores in a D-I-D analysis. As the value of D-I-D study designs and analysis in observational research becomes increasingly recognized, it is our hope that this brief, slightly more advanced tutorial gives analysts the information they need to go beyond just getting their feet wet and arms them with the confidence they need to dive right in.

REFERENCES

- Austin, P. C. 2009. 'Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples', *Stat Med*, 28: 3083-107.
- — —. 2011a. 'An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies', *Multivariate behavioral research*, 46: 399-424.
- — —. 2011b. 'A Tutorial and Case Study in Propensity Score Analysis: An Application to Estimating the Effect of In-Hospital Smoking Cessation Counseling on Mortality', *Multivariate Behav Res*, 46: 119-51.
- Castelloe, John. 2014. "Power and Sample Size for MANOVA and Repeated Measures with the GLMPOWER Procedure, Paper SAS030-2014." In *SAS Global Forum 2014 Conference*. Washington, D.C.: SAS Institute Inc.
- Karter, A. J., M. M. Parker, O. K. Duru, D. Schillinger, N. E. Adler, H. H. Moffet, A. S. Adams, J. Chan, W. H. Herman, and J. A. Schmittiel. 2015. 'Impact of a pharmacy benefit change on new use of mail order pharmacy among diabetes patients: the Diabetes Study of Northern California (DISTANCE)', *Health Serv Res*, 50: 537-59.
- Lipska, K. J., M. M. Parker, H. H. Moffet, E. S. Huang, and A. J. Karter. 2018. 'Association of Initiation of Basal Insulin Analogs vs Neutral Protamine Hagedorn Insulin With Hypoglycemia-Related Emergency Department Visits or Hospital Admissions and With Glycemic Control in Patients With Type 2 Diabetes', *JAMA*, 320: 53-62.
- Neugebauer, R., J. A. Schmittiel, and M. J. van der Laan. 2016. 'A Case Study of the Impact of Data-Adaptive Versus Model-Based Estimation of the Propensity Scores on Causal Inferences from Three Inverse Probability Weighting Estimators', *Int J Biostat*, 12: 131-55.
- Parker, M. M., A. Fernandez, H. H. Moffet, R. W. Grant, A. Torreblanca, and A. J. Karter. 2017. 'Association of Patient-Physician Language Concordance and Glycemic Control for Limited-English Proficiency Latinos With Type 2 Diabetes', *JAMA Intern Med*, 177: 380-87.
- Reed, M. E., E. M. Warton, E. Kim, M. D. Solomon, and A. J. Karter. 2017. 'Value-Based Insurance Design Benefit Offsets Reductions in Medication Adherence Associated With Switch To Deductible Plan', *Health Aff (Millwood)*, 36: 516-23.
- Simpson, S. L., L. J. Edwards, K. E. Muller, P. K. Sen, and M. A. Styner. 2010. 'A linear exponent AR(1) family of correlation structures', *Stat Med*, 29: 1825-38.
- Spiegelman, D., and E. Hertzmark. 2005. 'Easy SAS calculations for risk or prevalence ratios and differences', *Am J Epidemiol*, 162: 199-200.
- Warton, EM, MM Parker, and AJ Karter. 2016. "How D-I-D you do that? Basic Difference-in-Differences Models in SAS®." In *Western Users of SAS® Software 2016*. San Francisco, CA: SAS Institute.
- Yang, Dongsheng, and Jarrod E. Dalton. 2012. "A unified approach to measuring the effect size between two groups using SAS Paper 335-2012." In *SAS Global Forum*. Orlando, FL: SAS Institute.
- Zou, G. 2004. 'A modified poisson regression approach to prospective studies with binary data', *Am J Epidemiol*, 159: 702-6.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Margaret Warton
Kaiser Permanente Division of Research
margaret.m.warton@kp.org

Melissa M. Parker
Kaiser Permanente Division of Research
melissa.m.parker@kp.org