Difference-in-Differences Method in Comparative Effectiveness Research: Utility with Unbalanced Groups

Huanxue Zhou¹ · Christopher Taber² · Steve Arcona³ · Yunfeng Li³

Published online: 1 July 2016

© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract

Background Comparative effectiveness research (CER) often includes observational studies utilizing administrative data. Multiple conditioning methods can be used for CER to adjust for group differences, including difference-in-differences (DiD) estimation.

Objective This study presents DiD and demonstrates how to apply this conditioning method to estimate treatment outcomes in the CER setting by utilizing the MarketScan® Databases for multiple sclerosis (MS) patients receiving different therapies.

Methods The sample included 6762 patients, with 363 in the Test Cohort [glatiramer acetate (GA) switched to fingolimod (FTY)] and 6399 in the Control Cohort (GA only, no switch) from a US administrative claims database. A trend analysis was conducted to rule out concerns regarding regression to the mean and to compare relapse rates among treatment cohorts. DiD analysis was used to enable comparisons among the Test and Control Cohorts. Logistic regression was used to estimate the probability of relapse after switching from GA to FTY, and to compare group differences in the pre- and post-index periods.

Results Crude DiD analysis showed that in the pre-index period more patients in the Test Cohort experienced an MS

relapse and had a higher mean number of relapses than in the Control Cohort. During the pre-index period, numeric and relative data for MS relapses in patients in the Test Cohort were significantly higher than in the Control Cohort, while no significant between-group differences emerged during the post-index period. Generalized linear modeling with DiD regression estimation showed that the mean number of MS relapses decreased significantly in the post-index period among patients in the Test Cohort compared with patients in the Control Cohort.

Conclusion In this study, an MS population was utilized to demonstrate how DiD can be applied to estimate treatment effects in a heterogeneous population, where the Test and Control Cohorts varied greatly. The results show that DiD offers a robust method for comparing diverse cohorts when other risk-adjustment methods may not be adequate.

Key Points for Decision Makers

Difference-in-differences (DiD) permits the comparison of differences in outcomes, before and after an intervention, between groups by controlling for bias from unobserved variables that remain fixed over time.

The current study demonstrated the application of the DiD methodology in CER settings to estimate treatment effects in a heterogeneous MS population, where the Test and Control Cohorts varied greatly.

This study has shown that DiD offers a robust comparison of groups, when propensity score matching and other risk-adjustment methods are not suitable. One potential issue is a jump in health outcomes immediately prior to switching the drug.

✉ Yunfeng Li
frank.li@novartis.com

¹ KMK Consulting, Inc., 7 North Tower, 23 Headquarters Plaza, Morristown, NJ 07960, USA

² Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706, USA

³ Outcomes Research Methods and Analytics, US Health Economics and Outcome Research, Novartis Pharmaceuticals Corporation, One Health Plaza 135/584, East Hanover, NJ 07936-1080, USA

1 Introduction

Comparative effectiveness research (CER) has become a cornerstone methodology for health-care decision making, particularly for informing therapeutic options [1]. While randomized controlled trials are the gold standard of CER, observational studies utilizing administrative data are increasingly being conducted to estimate treatment effects between groups [2]. In order to evaluate treatment effects, comparable groups should be established that are well-balanced on multiple factors which may influence outcomes [3, 4].

Although propensity score matching (PSM) is commonly used in CER to create comparable groups, PSM does not apply to studies where treatment and control groups are highly skewed. Applying PSM may result in a small sample size because unmatched patients are dropped from the final sample [2, 4–8]. Furthermore, King and Nielsen [9] argue that PSM should not be used as it “increases imbalance, inefficiency, model dependence, research discretion, and statistical bias at some point in both real data and in data generated to meet the requirements of PSM theory”. Inverse probability of treatment weighting (IPTW) can be applied to balance treatment groups regarding factors that may bias the treatment effect estimates, without losing any patients from the sample [3, 4]. However, treatment effect estimates may be impacted by propensity scores that have very large weights [2]. In addition, approximates of PSM are subject to the completeness of measures related to the differences in patient and clinical characteristics; however, in observational studies, a complete composite of all these differences is not easy to capture because of the lack of related measures. Therefore, these characteristics may not be balanced between different groups, and this may bias the estimate of treatment effects.

Due to the limitations of PSM and IPTW, the difference-in-differences (DiD) method may be an alternative methodology [10]. Historically, DiD has been used in the evaluation of health-care policy, as it allows the researcher to control for background changes in outcomes [10]. DiD estimation permits the comparison of differences in outcomes before and after an intervention (e.g. treatment or health-care policy change) between groups affected and unaffected by the intervention [10, 11]. This methodology is appropriate to use when the interventions involved are “as good as random, conditional on time and group fixed effects” [11]. This means that DiD has the advantage of allowing researchers to estimate treatment effect, while accounting for unobserved variables that are assumed to remain fixed over time [12].

1.1 Study Objective

This retrospective administrative claims database study demonstrates how to apply the DiD method to estimate treatment outcomes in the CER setting. Specifically, this study applied DiD for analyzing treatment effects related to multiple sclerosis (MS) relapses. MS is a unique population, which is characterized by heterogeneity in clinical features and responsiveness to treatment [13]. Using DiD, this study focused on specific methods used to obtain the results regarding treatment effects in patients switching from glatiramer acetate (GA) to fingolimod (FTY) compared with those remaining on GA.

2 Difference-in-Differences (DiD) Methodology

2.1 Review of DiD Methodology

Simple pre- and post-treatment comparisons may be impacted by temporal trends in the outcome variable, or by other events that occurred between the two periods [14]. To overcome this issue, using a quasi-experimental design, DiD can be used when two periods of data are available for the treatment and comparison groups. The DiD estimator measures the treatment effect by looking at the difference between the average outcome in the control and treatment groups, before and after treatment [14].

2.1.1 DiD Assumptions

A key assumption of DiD is known as the ‘parallel trend’ assumption, which supposes that in the absence of treatment, the average outcomes of the treatment group and the comparison group would follow parallel paths over time [14]. This allows DiD to account for unobserved variables, which are assumed to remain fixed over time [12].

2.1.2 DiD Approach

In an evaluation of a treatment effect, a sample of patients are observed before and after a treatment. In the simplest case, if two periods of data (0 and 1) are analyzed and treatment begins in between the two periods of time, the treatment effect can be identified by simply looking at outcomes before and after the treatment, and classify the effect as:

$$\bar{Y}_1 - \bar{Y}_0$$

Herein, $\bar{Y}_1$ is the mean outcome in the period following the treatment and $\bar{Y}_0$ is the mean outcome in the period

prior to the commencement of the treatment. At this point, this can be called or thought of as a matching estimator where these actions match a member of the treatment group to him/herself prior to receiving the treatment. For covariates that do not change over time, perfect balance is present whether or not those variables are included in the dataset.

The problem with this approach is that ‘things’ do change over time. Moreover, the effects of any other event that happened between the two periods are attributed to the treatment. Therefore, in order to account for changes over time, a second group is required. Assume one group, Group A, is administered the treatment between periods 0 and 1 (Let $\bar{Y}_{A1} - \bar{Y}_{A0}$ be the change in the outcome for this group), while a second group, Group B, does not receive the treatment at all (Let $\bar{Y}_{B1} - \bar{Y}_{B0}$ be the difference in outcome for that group).

Under the assumption that $\bar{Y}_{B1} - \bar{Y}_{B0}$ provides a good estimate of what would have happened to Group A had they not received the treatment, the treatment effect can be estimated using the DiD estimate:

$$\hat{\alpha} = (\bar{Y}_{A1} - \bar{Y}_{A0}) - (\bar{Y}_{B1} - \bar{Y}_{B0}).$$

This approach can be formally justified with a fixed effects model.

Let $Y_{it} = \beta_0 + \alpha T_{it} + \delta t + \theta_i + \varepsilon_{it}$ where Y_{it} is the outcome for person i at time t , T_{it} indicates whether person i received the treatment at time t , t is time period (0 or 1) and θ_i is a person fixed effect. As long as $E(\varepsilon_{it}|G_i, t) = 0$ where G_i indicates the group type (either A or B). Then:

$$\begin{aligned} \hat{\alpha} &\approx (E(Y_{it}|G_i = A, t = 1) - E(Y_{it}|G_i = A, t = 0)) \\ &\quad - (E(Y_{it}|G_i = B, t = 1) - E(Y_{it}|G_i = B, t = 0)) \\ &= ([\beta_0 + \alpha + \delta + E(\theta_{it}|G_i = A)] - [\beta_0 + E(\theta_{it}|G_i = A)]) \\ &\quad - ([\beta_0 + \delta + E(\theta_{it}|G_i = B)] - [\beta_0 + E(\theta_{it}|G_i = B)]) \\ &= (\alpha + \delta) - (\delta) \\ &= \alpha \end{aligned}$$

In addition, this approach makes clear what key assumption justifies DiD. The sample analogue of the equation above yields:

$$\hat{\alpha} = \alpha + (\bar{\varepsilon}_{A1} - \bar{\varepsilon}_{A0}) - (\bar{\varepsilon}_{B1} - \bar{\varepsilon}_{B0})$$

Therefore, for consistency we need that:

$$E[(\bar{\varepsilon}_{A1} - \bar{\varepsilon}_{A0}) - (\bar{\varepsilon}_{B1} - \bar{\varepsilon}_{B0})] = 0$$

In practice, the DiD estimate can be obtained either as a simple DiD, by running the fixed effect regression above, or by running a regression of Y_{it} on T_{it} , t , and a dummy variable to belonging to Group A. This can be further generalized to include more time periods (T), more groups (G), and additional covariates as one can run the regression

$$Y_{it} = X'_{it}\beta + \alpha T_{it} + D'_t\delta + G'_i\theta + \varepsilon_{it}$$

where X_{it} represents additional covariates, D_t is a $T \times 1$ vector of dummy variables indicating the time period and G_i is a $G \times 1$ vector of dummy variables indicating the group to which individual i belongs. That is D_t consists of a one in row t and zeros in all other rows while G_i consists of a one in the row corresponding to the group in which individual i belongs, and zeros in all other rows. Expanding this notion to nonlinear models of a linear index such as a logit or a negative binomial is straight forward.

2.1.3 Limitations of Difference-in-Differences Method

The limitations of DiD relate to the need to find similar study groups, as ideally, the only difference should be exposure to the intervention. For instance, according to the common shocks assumption, any event that occurs during or following the intervention, should equally affect each group. Likewise, the parallel trends assumption, outlined above, can be evaluated using a regression model; if the trends between the two groups are significantly different, the analysis may be biased [10]. Therefore, a limitation of this method is in finding treatment and control groups which meet these assumptions [10]. While this approach accounts for unobservable variables that are fixed over time, the biggest issue is that it does not account for unobservable variables that are not fixed over time [15].

2.2 Application of Difference-in-Differences Method

2.2.1 MS Study Sample

The sample was obtained from the Truven Health MarketScan[®] Commercial Claims and Encounters and Medicare Supplemental Databases, one of the largest administrative claims databases in the USA with employer-sponsored and Medicare population with supplemental insurance [16]. Data were de-identified according to the US Insurance Portability and Accountability Act (HIPAA). The study did not involve collection, use, or transmission of individually identifiable data; thus, no Institutional Review Board approval was required.

2.2.2 Patient Cohorts and Study Design

In this study, two treatment cohorts were evaluated: (1) the Test Cohort (patients who had switched from GA to FTY and 2) the Control Cohort (patients who remained on GA) (see Fig. 1 for details on patient selection). The Test Cohort were patients who switched to FTY in the identification period and the Control Cohort were patients who received GA only during the identification period (October

Fig. 1 Patient selection flowchart. ^aFor patients in the Test Cohort, the date of the first FTY claim is the index date; for patients in the Control Cohort, the date of the first GA claim is the index date. ^bFor patients in the Test Cohort, this excludes other MS drugs that are not FTY; for patients in the Control Cohort, this excludes other MS drugs that are not GA. *FTY* fingolimod, *GA* glatiramer acetate, *MS* multiple sclerosis

1, 2010 to September 30, 2012). The index date was defined as the first FTY claim in the Test Cohort, or first GA claim in the Control Cohort. The pre-index period, or baseline period, was defined as the 12 months before the index date, while the post-index period was defined as the 12 months following the index date.

The primary outcome was relapse rate during the post-index period. In this study, an MS relapse was defined using the claims-based algorithm validated by Chastek et al. [17], which involves meeting one of two criteria: a claim with an MS diagnosis code in the primary position at any time during an inpatient hospitalization, or a claim with an MS diagnosis code in the primary or secondary position in an outpatient setting plus a pharmacy or medical claim for a qualifying corticosteroid on the day of, or within 7 days, after the visit [17]. Additionally, the ‘clean period’ between initiation of relapses must be at least 30 days [17].

The study included eight quarters, representing the pre-index and post-index periods; the pre-index period included four quarters of data (4th quarter prior to index, 3rd quarter prior to index, 2nd quarter prior to index, 1st quarter prior to index; labeled as: −4, −3, −2, −1, respectively in Fig. 2), likewise, the four quarters of data in

the post-index period were the 1st quarter post index, 2nd quarter post index, 3rd quarter post index and the 4th quarter post the index. These quarters are labeled as: 1, 2, 3, and 4, respectively in Fig. 2.

In the first phase of this study, IPTW analyses were applied (see Appendix Table 4. Balance Check Propensity Score Weighted Baseline Measures in appendices). The preliminary analysis revealed that the patient populations varied greatly across the two treatment cohorts; therefore, PSM and other risk-adjustment methods would not have been suitable for further analysis. DiD analysis was used to enable robust comparison of the Test and Control Cohorts.

2.2.3 MS Study Patient Baseline Characteristics

The analysis included data from 6762 patients, including 363 (5.4 %) in the Test Cohort, 6399 (94.6 %) in Control Cohort. For reasons we will discuss later, we eliminated data from −1Q prior to switching drugs, the quarter immediately prior to the index date, as described in the section below (see Sect. 2.3).

Baseline demographic and clinical characteristics varied between the Test Cohort and the Control Cohort (Table 1).

Fig. 2 Patients (%) with MS relapses during the pre- and post-index period by quarter.

^aThe index date was defined as the first claim for FTY in the Test Cohort and as the date of first claim for the Control Cohort. FTY fingolimod, GA glatiramer acetate, MS multiple sclerosis

While no significant differences in gender or type of insurance plan were reported, on average, patients in the Test Cohort were significantly younger than those in the Control Cohort ($p = 0.0000$; Table 1). Patients in the Test Cohort had a significantly higher mean (SD) number of medications than those in the Control Cohort [8.0 (5.1) vs. 7.3 (5.5), $p = 0.0099$, respectively]. Furthermore, a significantly larger percentage of patients in the Test Cohort had MRI scans than the Control Cohort (71.9 vs. 47.5 %, $p = 0.0000$), and patients in the Test Cohort had significantly more MRI scans than those in the Control Cohort [1.0 (0.9) vs. 0.6 (0.8), $p = 0.0000$; Table 1]. Overall, the Test Cohort had a higher percentage of patients with MS symptoms (78.8 %) compared to the Control Cohort (69.5 %; $p = 0.0002$). Specifically, compared to the Control Cohort, a significantly higher percentage of patients in the Test Cohort experienced pain ($p = 0.0206$), fatigue ($p = 0.0051$), gait, balance and coordination ($p = 0.0140$), other emotional changes ($p = 0.0159$) and other symptoms ($p = 0.0000$; Table 1). For both continuous and categorical measures of medication adherence, values for patients in the Test Cohort were significantly lower than were those for patients in the Control Cohort (Table 1).

2.2.4 Trend Analysis

Before conducting the DiD analyses, a trend analysis was conducted to investigate the parallel trends assumption, the key assumption of DiD.

Ideally, in the absence of treatment, the trends in outcomes would be parallel between the treatment and

control groups. While it is impossible to test this assumption after the treatment has been administered, it is feasible to test it in the prior periods. The common practice is to examine the outcomes of interest graphically with multiple points of time to see whether the common trend assumption remains in the periods before the treatment is administered. As shown in Fig. 2, we see MS relapses, either measured by the mean number of MS relapses or proportion of patients who experienced an MS relapse, were close to parallel between the Test Cohort and the Control Cohort during the pre-index period, except in the quarter immediately prior to the switch, labeled as -1Q in Fig. 2.

In -1Q, we observed a peak in relapse rates, an issue referred to as the Ashenfelter's Dip in the economics literature [18, 19]. We suspect that the MS relapse may be a major driver for patients to switch medications, which relates to our model above. This suggests an issue to be addressed. To explore this issue mathematically, we have four quarters before and four quarters after the treatment has been implemented so we can write the DiD estimator as:

$$\begin{aligned} \hat{\alpha} = & ([\bar{Y}_{A1} + \bar{Y}_{A2} + \bar{Y}_{A3} + \bar{Y}_{A4}] - [\bar{Y}_{A-4} - \bar{Y}_{A-3} - \bar{Y}_{A-2} - \bar{Y}_{A-1}]) \\ & - ([\bar{Y}_{B1} + \bar{Y}_{B2} + \bar{Y}_{B3} + \bar{Y}_{B4}] - [\bar{Y}_{B-4} - \bar{Y}_{B-3} - \bar{Y}_{B-2} - \bar{Y}_{B-1}]) \\ = & \alpha + ([\bar{e}_{A1} + \bar{e}_{A2} + \bar{e}_{A3} + \bar{e}_{A4}] - [\bar{e}_{A-4} + \bar{e}_{A-3} + \bar{e}_{A-2} + \bar{e}_{A-1}]) \\ & - ([\bar{e}_{B1} + \bar{e}_{B2} + \bar{e}_{B3} + \bar{e}_{B4}] - [\bar{e}_{B-4} + \bar{e}_{B-3} + \bar{e}_{B-2} + \bar{e}_{B-1}]) \end{aligned}$$

For consistency, we need the expected value of the $\bar{e}$ terms to be zero; however, in Fig. 2 it looks like $\bar{e}_{A-1}$ is a large number. It suggests that timing of MS relapse may be

Table 1 Summary of pre-index period (baseline) demographics and clinical characteristics for Test Cohort and Control Cohort

Characteristic	Test cohort (<i>n</i> = 363) ^d	Control cohort (<i>n</i> = 6399) ^e	Between-group OR/difference		
			Mean	95 % CI	<i>p</i> value ^a
<i>Including −1Q before index date</i>					
Mean (SD) age, y	47.3 (10.1)	49.7 (10.5)	−2.37	−3.44, −1.30	0.0000
Age group, <i>n</i> (%)					
18–34	43 (11.8)	541 (8.5)			0.0000
35–44	95 (26.2)	1446 (22.6)			
45–54	143 (39.4)	2205 (34.5)			
≥55	82 (22.6)	2207 (34.5)			
Female gender, <i>n</i> (%)	283 (78.0)	4905 (76.7)			0.5660
Plan type, <i>n</i> (%)					0.3334
Fee for service	308 (84.8)	5304 (82.9)			
HMO and POS capitation	55 (15.2)	1095 (17.1)			
Region, <i>n</i> (%)					0.0227
Northeast	51 (14.0)	1241 (19.4)			
North central	101 (27.8)	1800 (28.1)			
South	134 (36.9)	1961 (30.6)			
West	77 (21.2)	1397 (21.8)			
Mean (SD) medications	8.0 (5.1)	7.3 (5.5)	0.71	0.17, 1.25	0.0099
Patients w/MRI services, <i>n</i> (%)	261 (71.9)	3038 (47.5)	2.83	2.24, 3.58	0.0000
Mean (SD) MRI services	1.0 (0.9)	0.6 (0.8)	0.40	0.31, 0.50	0.0000
Patients w/MS symptoms, <i>n</i> (%)	286 (78.8)	4447 (69.5)	1.63	1.26, 2.11	0.0002
Pain	178 (49.0)	2742 (42.9)	1.28	1.04, 1.59	0.0206
Depression	52 (14.3)	924 (14.4)	0.99	0.73, 1.34	0.9518
Fatigue	65 (17.9)	820 (12.8)	1.48	1.12, 1.96	0.0051
Gait, balance and coordination	54 (14.9)	687 (10.7)	1.45	1.08, 1.96	0.0140
Headache	44 (12.1)	648 (10.1)	1.22	0.88, 1.70	0.2226
Bladder dysfunction	34 (9.4)	626 (9.8)	0.95	0.66, 1.37	0.7948
Other emotional change	47 (12.9)	586 (9.2)	1.48	1.07, 2.03	0.0159
Numbness	41 (11.3)	566 (8.8)	1.31	0.94, 1.84	0.1122
Others ^b	165 (45.5)	2113 (33.0)	1.69	1.37, 2.09	0.0000
Mean (SD) compliance (modified) ^c	0.68 (0.28)	0.84 (0.21)	−0.16	−0.19, −0.13	0.0000
Modified compliance ^c group, <i>n</i> (%)					0.0000
<50 %	93 (25.6)	577 (9.0)			
50–80 %	104 (28.7)	1139 (17.8)			
≥80 %	166 (45.7)	4683 (73.2)			

CI confidence interval, GA glatiramer acetate, FTY fingolimod, HMO Health Maintenance Organization, MRI magnetic resonance imaging, Max maximum, MS multiple sclerosis OR odds ratio, POS point-of-service, Q quarter, Rx prescription, SD standard deviation, Tx treatment, w with, y year

^a Chi-square test (Fisher's exact test is employed when ≥ 20 % of the cells have an expected value < 5) for categorical/dummy variables and 2-sided 2-sample *t* test with unequal variance for continuous variables

^b Other includes bowel dysfunction, visual symptoms, sexual dysfunction, dizziness and vertigo, muscle weakness/spasm/spasticity, cognitive function, speech disorders, swallowing problems, hearing loss, seizures, tremor, itching

^c Calculated as: [total Rx days of supply − Max (0, (estimated next Tx date − study end date))]/(index date − date of first GA in baseline period)

^d The Test Cohort included patients who switched from GA to FTY

^e Control Cohort included patients who remained on GA

related to switching to FTY. Doctors do not switch their patients at random times, they are likely to switch them following a relapse. The problem is if a high value of $\bar{e}_{A-1}$

induces the doctor to switch drugs then we would expect that $\bar{e}_{A-1} - \bar{e}_{B-1} > 0$ which would lead us to overstate the effect of the drug.

There is no perfect way to address this problem. On the one hand, if the switch to FTY was only related to $\bar{e}_{A-1}$ and not any of the other error terms, then by throwing out data from $-1Q$ we can get a consistent estimate of α . On the other hand, if there is positive serial correlation in $\bar{e}_{AQ}$ then we would expect some of the shock to persist. If this is the case, then excluding data from period $-1Q$ likely leads us to understate the effect of the drug. In this instance, we can think of the two specifications (excluding and including data from $-1Q$) as providing upper and lower bounds on the effect.

3 Results

3.1 Crude DiD Estimate

First, a crude DiD estimate was applied to estimate treatment effects. As shown in Table 2, when including data from $-1Q$, in the pre-index period 109 (30.0 %) of the patients in the Test Cohort had an MS relapse, compared to 898 (14.0 %) in the Control Cohort. Overall, in the post-index period, 50 patients (13.8 %) in the Test Cohort experienced an MS Relapse, compared to 739 patients (11.5 %) in the Control Cohort (Table 2). In terms of the frequency of MS relapses, the mean (SD) number of relapses in the pre-index period was 0.34 (0.58) and 0.17 (0.47), for the Test and Control Cohorts, respectively, while in the post-index period, the mean (SD) was 0.18 (0.52) and 0.14 (0.43) in the Test and Control Cohorts,

respectively (data not shown). Similarly, when excluding data from $-1Q$, a higher number of patients in the Test Cohort experienced an MS relapse than in the Control Cohort in the pre-index period (Table 2).

3.2 Logistic Regression Models

Following the trend and crude DiD analyses, logistic regression was utilized to statistically test the parameters relating to group differences in the pre- and post-index periods to understand whether treatment can reduce relapse rates. Using the number of relapses while taking the medication (herein, presented by the proportion of patients experiencing a relapse) as the dependent variable, logistic regression was used to estimate the probability of experiencing a relapse while taking FTY or GA, and group differences in the pre- and post-index periods between the Test and Control Cohorts were compared.

Including data from $-1Q$, during the pre-index period, the overall risk of MS relapse was significantly higher in patients in the Test Cohort than for the Control Cohort (OR = 2.63, 95 % CI: 2.08, 3.33, $p = 0.0000$). However, after switching, the overall risk of MS relapse was not significantly different between the Test and Control Cohorts (OR = 1.22, 95 % CI: 0.90, 1.67, $p = 0.1994$) (Table 2). Likewise, when excluding data from $-1Q$, the overall risk of an MS relapse was significantly higher among patients in the Test Cohort than for the Control Cohort in the pre-index period (OR = 2.04, 95 % CI: 1.56, 2.65, $p = 0.0000$).

Table 2 Crude DiD outcomes and odds ratio by logistic regression

	Test cohort ^a	Control cohort ^b	Between-group odds ratio/difference			
	(<i>n</i> = 363)	(<i>n</i> = 6399)	Mean	95 % CI	<i>p</i> Value ^c	-2 Log L ^d
Including -1Q before index date						
Patients with MS relapse, <i>n</i> (%)						
Overall in pre-period	109 (30.0)	898 (14.0)	2.63	2.08, 3.33	0.0000	5634
Overall in post-period	50 (13.8)	739 (11.5)	1.22	0.90, 1.67	0.1994	4870
Excluding -1Q before index date						
Patients with MS relapse, <i>n</i> (%)						
Overall in pre-period	76 (20.9)	736 (11.5)	2.04	1.56, 2.65	0.0000	4939
Overall in post-period	50 (13.8)	739 (11.5)	1.22	0.90, 1.67	0.1994	4870

CI confidence interval, DiD difference-in-differences, FTY fingolimod, GA glatiramer acetate, MS multiple sclerosis; Q quarter, OR odds ratio, SD standard deviation

^a The Test Cohort included patients who switched from GA to FTY

^b Control Cohort included patients who remained on GA

^c Chi-square test (Fisher's exact test is employed when ≥ 20 % of the cells have an expected value < 5) for categorical/dummy variables and 2-sided 2-sample t test with unequal variance for continuous variables

^d There were two subpopulations and two model parameters in each logistic regression model and residual degrees of freedom is zero. Therefore, the Pearson and deviance goodness-of-fit tests cannot be obtained

3.3 DiD Regression Estimation

Finally, using DiD regression estimation by including an interaction between time (pre-index vs. post-index period) and cohorts (Test vs. Control Cohort) into explanatory variables and the count of number of patients with MS relapse as dependent variables, treatment effects by switching from GA to FTY were estimated while controlling for time effects. The purpose of this analysis was two-fold: first, it was used to estimate the magnitude of treatment effects, by controlling for the time period to see how much treatment contributes to the outcomes; and secondly, it was used to test if these differences were statistically significant. The results showed that the mean number of MS relapses decreased significantly from the pre- to the post-index period for Test Cohort, compared with the Control Cohort. As mentioned previously, the MS relapse rate made a significant jump in the quarter prior to switching to FTY for the Test Cohort, implicating Ashenfelter's Dip. To handle this issue, two separate analyses were conducted, one including data from $-1Q$ and another excluding data from $-1Q$. The analysis showed that the MS relapse rate decreased by 36 % [$1 - \exp(-0.44)$] in the Test Cohort from the pre- to post-index period ($p = 0.0007$, Table 3) when including data from $-1Q$ while the MS relapse rate decreased by 25 % [$1 - \exp(-0.29)$; $p = 0.0276$, Table 3] when excluding data from $-1Q$. Thus applying our bounding argument, we conclude that the MS relapse rate decreased by between 25 and 36 %.

4 Discussion

To the best of our knowledge, the DiD method has not previously been used in a CER setting to examine treatment effects on health outcomes. The current study

provides a unique example to demonstrate the application of DiD, evaluating treatment effects of two MS therapies on the number of relapses experienced in two patient cohorts: the Test Cohort and the Control Cohort. The preliminary analysis of the Test and Control Cohorts showed that the patient populations varied significantly on several demographic and clinical characteristics; therefore, PSM and other risk-adjustment methods would not have been adequate.

A trend analysis was conducted to rule out concerns regarding regression to the mean and to compare the relapse rates among the Test and Control Cohorts. The trend analysis showed that the mean number of MS relapses, and the proportion of patients experiencing an MS relapse, were significantly higher in the Test Cohort compared to the Control Cohort during the pre-index period. This change represents a problem known as Ashenfelter's Dip. In the economics literature, the Ashenfelter's Dip refers to the decline in mean earnings among participants in government training programs just prior to program entry (e.g. adult education programs), which may bias before-after estimates in program evaluation, where pre- and post-program earnings are compared [18, 19]. In the current study, the Ashenfelter's Dip may have important consequences in measuring treatment effects, as before-after comparisons may overstate or understate the impact of treatment [19]. Evidence of the Ashenfelter's Dip among the Test Cohort is not surprising, as it suggests that a switch in medication may be due to the timing of an MS relapse. In order to provide an estimate of the upper and lower bounds of the treatment effect, analyses were conducted including and excluding data from $-1Q$.

Including data from $-1Q$, the crude DiD analysis showed that a higher percentage of patients in the Test Cohort had experienced an MS relapse than in the Control Cohort in the pre-index period, as well as a higher mean

Table 3 Negative binomial DID model for the number of MS relapses

Parameter	Estimate	SE	95% CI	<i>p</i> value
Test Cohort ^a vs Control Cohort ^b (including $-1Q$ before index date)				
Intercept	-1.78	0.02	-1.82, -1.73	<0.0001
Cohort: GA → FTY (reference = Control Cohort)	0.71	0.08	0.55, 0.87	<0.0001
Time: post-period (reference = pre-index period)	-0.19	0.04	-0.26, -0.12	<0.0001
Cohort × time	-0.44	0.13	-0.69, -0.18	0.0007
Test Cohort ^a vs. Control Cohort ^b (excluding $-1Q$ before index date)				
Intercept	-2.04	0.03	-2.09, -1.99	<0.0001
Test Cohort (reference = control cohort)	0.56	0.09	0.39, 0.74	<0.0001
Time: post-period (reference = pre-index period)	0.08	0.04	0.01, 0.15	0.0303
Cohort × time	-0.29	0.13	-0.55, -0.03	0.0276

CI confidence interval, DiD difference-in-differences, FTY fingolimod, GA glatiramer acetate, MS multiple sclerosis; *Q* quarter, *S* standard error

^a The Test Cohort included patients who switched from GA to FTY

^b Control Cohort included patients who remained on GA

number of relapses. Logistic regression was used to estimate the probability of experiencing a relapse while taking FTY or GA, and to compare group differences in the pre- and post-index periods. Overall, for the duration of the pre-index period, both numeric and relative data for MS relapse in patients in the Test Cohort were significantly higher than in the Control Cohort, while no significant between-group differences emerged during the post-index period. Finally, differences in the number of relapses while on FTY or GA were estimated using generalized linear modeling with a DiD regression model, which showed that while patients in the Test Cohort experienced significantly more MS relapses, the interaction term for time $\times$ treatment cohort showed that the mean number of MS relapses decreased significantly in the post-index period and compared with patients in the Control Cohort.

As an alternative to other methods (e.g. PSM or IPTW), DiD allows the researcher to control bias from unobserved variables that remain fixed over time and which are correlated with outcomes [12]. DiD is most often used to look at interventions, programs, or health-care policy changes. In one review [11], the most commonly used variables include employment/wages, other market variables, and health outcomes. Several papers utilizing DiD have examined health-care policy and health outcomes [20–27]. For example, Dimick and Ryan [10] highlighted two articles [28, 29] utilizing DiD to evaluate changes following the 2011 Accreditation Council for Graduate Medical Education duty hour reforms. From the pharmacology perspective, DiD has been used to evaluate patterns of oral hypoglycemic agents (e.g. discontinuation) following the publication of a meta-analysis on adverse events with specific medications [30].

4.1 Limitations

There are limitations associated with the utilization of administrative data, as these databases are created to manage health-care transactions rather than for research purposes. Variation in patient characteristics covered by different types of health insurance plans may be present; therefore, the findings of this study may not be generaliz-

able outside of MS patients covered by commercial health insurance in the USA.

5 Conclusion

The current study demonstrated the application of DiD methodology in CER settings to estimate treatment effects in a heterogeneous MS population, where the Test and Control Cohorts varied greatly. Our study has shown that DiD offers a more appropriate comparison when PSM and other risk-adjustment methods are not deemed to be adequate

Acknowledgements Michelle A. Adams, BSJ, MA of Write All Inc. and Brittany Gerber, MA of Medlior Health Outcomes Research Ltd. provided medical writing and editorial assistance for this manuscript.

Author contributions All listed authors met the criteria for authorship set for by the International Committee for Medical Journal Editors (ICMJE).

Compliance with Ethical Standards H Zhou is an Analyst at KMK Consulting Inc. and works as a consultant for Novartis Pharmaceuticals Corporation. Y. Li and S. Arcona are employees of Novartis Pharmaceuticals Corporation. C. Taber is an Economist at the Department of Economics, University of Wisconsin, and received consulting fees for his expertise from Novartis Pharmaceuticals Corporation. Funding for this project was provided by Novartis Pharmaceuticals Corporation, East Hanover, NJ. Publication of the study results was not contingent upon sponsor's approval and operated independently of funders.

Open Access This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix

See Table 4.

Table 4 Balance check propensity score weighted baseline measures

Variable	Test cohort (GA to FTY)	Control cohort (GA only)	Odds ratio (difference) (test vs. control)		
			Mean	95 % CI	<i>p</i> value
Age group, years (%)					0.2228
18–34	7.1	8.7			
35–44	19.3	22.8			
45–54	36.2	34.7			
55+	37.4	33.8			
Female gender (%)	80.7	76.8			0.09
Plan type (%)					0.23
Fee for service	80.5	83.0			
HMO and POS capitation	19.5	17.0			
Region (%)					0.87
Northeast	19.3	19.1			
North central	26.3	28.1			
South	31.2	31.0			
West	23.2	21.8			
Medication burden, mean ^a	7.90	7.35	0.54	0.01, 1.08	0.05
Patients with MRI service (%)	50.7	48.8	1.08	0.87, 1.33	0.49
Patients with relapse (%)	14.4	14.9	0.97	0.71, 1.30	0.82
MS symptoms (%)					
Pain	49.4	43.2	1.28	1.04, 1.59	0.02
Depression	15.3	14.4	1.07	0.80, 1.44	0.64
Fatigue	12.5	13.0	0.95	0.69, 1.31	0.78
Walking (gait), balance, coordination problems	12.5	10.9	1.17	0.85, 1.61	0.34
Headache	9.7	10.2	0.95	0.66, 1.35	0.76
Bladder dysfunction	9.2	9.7	0.94	0.65, 1.35	0.73
Other emotional changes	10.7	9.4	1.16	0.83, 1.64	0.39
Numbness	11.0	9.0	1.25	0.89, 1.75	0.20
Others	36.4	33.7	1.12	0.90, 1.40	0.29
Modified PDC group ^b					
<50 %	10.7	9.9			0.84
50 to ≤80 %	17.6	18.4			
≥80 %	71.8	71.7			

PDC proportion of days covered, MS multiple sclerosis, MRI magnetic resonance imaging, HMO Health Maintenance Organization, POS point of service, GA glatiramer acetate, FTY fingolimod, PDC proportion of days covered

^a Excluding DMTs

^b Modified PDC is calculated by [total Rx days of supply – max (0, (estimated next Tx date – study end date))] / (index date – date of first GA in baseline period)

References

1. Concato J, Lawler EV, Lew RA, et al. Observational methods in comparative effectiveness research. *Am J Med.* 2010;123(12 Suppl 1):e16–23.
2. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar Behav Res.* 2011;46(3):399–424.
3. Curtis LH, Hammill BG, Eisenstein EL, et al. Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Medical Care.* 2007;45(10 Supl 2):S103–7.
4. Lanehart RE, Rodriguez de Gil P, Kim ES, Bellara AP, Kromrey JD, Lee SR. Paper 314-2012: propensity score analysis and assessment of propensity score approaches using SAS procedures. *SAS Global Forum* 2012; 2012.
5. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med.* 1997;127(8 Pt 2):757–63.
6. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Serv Outcomes Res Methodol.* 2001;2:259–78.
7. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med.* 2008;27(12):2037–49.

8. Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biometrical J.* 2009;51(1): 171–84.
9. King G, Nielsen R. Why propensity scores should not be used for matching. 2016. Available at: <http://gking.harvard.edu/publications/why-propensity-scores-should-not-be-used-formatching>. Accessed 7 Sept 2015.
10. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *Jama.* 2014;312(22):2401–2.
11. Bertrand M, Duflo E, Mullainathan S. How much should we trust difference-in-differences estimates? *Q J Econ.* 2004;119:249–75.
12. Crown WH. Propensity-score matching in economic analyses: comparison with regression models, instrumental variables, residual inclusion, differences-in-differences, and decomposition methods. *Appl Health Econ Health Policy.* 2014;12(1):7–18.
13. Disanto G, Berlanga AJ, Handel AE, et al. Heterogeneity in multiple sclerosis: scratching the surface of a complex disease. *Autoimmun Dis.* 2010;2011:932351.
14. Abadie A. Semiparametric difference-in-differences estimators. *Rev Econ Stud.* 2005;72(1):1–19.
15. Meyer B. Natural and quas-experiments in economics. *J Bus Econ Stat.* 1995;13(2):151–61.
16. Butler Quint J. White paper health research data for the real world: the MarketScan® Databases. Ann Arbor: Truven Health Analytics Inc; 2015.
17. Chastek BJ, Oleen Burkey M, Lopez-Bresnahan MV. Medical chart validation of an algorithm for identifying multiple sclerosis relapse in healthcare claims. *J Med Econ.* 2010;13(4):618–25.
18. Ashenfelter O. Estimating the effect of training programs on earnings. *Rev Econ Stat.* 1978;60:47–57.
19. Heckman JJ, Smith JA. The pre-programme earnings dip and the determinants of participation in a social programme. Implications for simple programme evaluation strategies. *Econ J.* 1999;109: 313–48.
20. Weiss J, Makonnen R, Sula D. Shifting management of a community volunteer system for improved child health outcomes: results from an operations research study in Burundi. *BMC Health Serv Res.* 2015;15(Suppl 1):S2.
21. Brenner S, Muula AS, Robyn PJ, et al. Design of an impact evaluation using a mixed methods model—an explanatory assessment of the effects of results-based financing mechanisms on maternal healthcare services in Malawi. *BMC Health Serv Res.* 2014;14:180.
22. Colla CH, Lewis VA, Gottlieb DJ, et al. Cancer spending and accountable care organizations: evidence from the Physician Group Practice Demonstration. *Healthcare (Amst).* 2013;1(3–4): 100–7.
23. Dubay L, Kenney G. Expanding public health insurance to parents: effects on children's coverage under Medicaid. *Health Serv Res.* 2003;38(5):1283–301.
24. McAdam-Marx C, Dahal A, Jennings B, et al. The effect of a diabetes collaborative care management program on clinical and economic outcomes in patients with type 2 diabetes. *J Manag Care Spec Pharm.* 2015;21(6):452–68.
25. Pereira SK, Kumar P, Dutt V, et al. Protocol for the evaluation of a social franchising model to improve maternal health in Uttar Pradesh, India. *Implement Sci.* 2015;10:77.
26. Salinas-Rodriguez A, Torres-Pereda Mdel P, Manrique-Espinoza B, et al. Impact of the non-contributory social pension program 70 y mas on older adults' mental well-being. *PLoS One.* 2014;9(11):e113085.
27. Siddiqui M, Roberts ET, Pollack CE. The effect of emergency department copayments for Medicaid beneficiaries following the Deficit Reduction Act of 2005. *JAMA Intern Med.* 2015;175(3): 393–8.
28. Rajaram R, Chung JW, Jones AT, et al. Association of the 2011 ACGME resident duty hour reform with general surgery patient outcomes and with resident examination performance. *Jama.* 2014;312(22):2374–84.
29. Patel MS, Volpp KG, Small DS, et al. Association of the 2011 ACGME resident duty hour reforms with mortality and readmissions among hospitalized Medicare patients. *Jama.* 2014; 312(22):2364–73.
30. Jain RMC, Lee H, Wong W. Use of rosiglitazone and pioglitazone immediately after the cardiovascular risk warnings. *Res Soc Adm Pharm.* 2012;8(1):47–59.