

Inference With Difference-in-Differences With a Small Number of Groups

A Review, Simulation Study, and Empirical Application Using SHARE Data

Slawa Rokicki, PhD,*† Jessica Cohen, PhD,‡ Günther Fink, PhD,§ Joshua A. Salomon, PhD,‡
and Mary Beth Landrum, PhD||

Background: Difference-in-differences (DID) estimation has become increasingly popular as an approach to evaluate the effect of a group-level policy on individual-level outcomes. Several statistical methodologies have been proposed to correct for the within-group correlation of model errors resulting from the clustering of data. Little is known about how well these corrections perform with the often small number of groups observed in health research using longitudinal data.

Methods: First, we review the most commonly used modeling solutions in DID estimation for panel data, including generalized estimating equations (GEE), permutation tests, clustered standard errors (CSE), wild cluster bootstrapping, and aggregation. Second, we compare the empirical coverage rates and power of these methods using a Monte Carlo simulation study in scenarios in which we vary the degree of error correlation, the group size balance, and the proportion of treated groups. Third, we provide an empirical example using the Survey of Health, Ageing, and Retirement in Europe.

Results: When the number of groups is small, CSE are systematically biased downwards in scenarios when data are unbalanced or when there is a low proportion of treated groups. This can result in over-rejection of the null even when data are composed of up to 50 groups. Aggregation, permutation tests, bias-adjusted GEE, and wild cluster bootstrap produce coverage rates close to the nominal rate for almost all scenarios, though GEE may suffer from low power.

Conclusions: In DID estimation with a small number of groups, analysis using aggregation, permutation tests, wild cluster bootstrap, or bias-adjusted GEE is recommended.

Key Words: difference-in-differences, clustered standard errors, inference, Monte Carlo simulation, GEE

(*Med Care* 2018;56: 97–105)

Difference-in-differences (DID) estimation has become increasingly popular in the medical and epidemiological literature in recent years.^{1–6} DID is often used to evaluate the effect of a group-level policy on individual-level outcomes. As observations are grouped, errors are correlated across individuals within groups; models that do not account for this correlation will result in misleadingly small standard errors (SEs) and incorrect inference.^{7,8}

DID estimation is often used to analyze the impact of specific policy experiments and interventions. Given that such changes generally occur only in few hospitals, districts, or states, the number of groups in most health-focused DID analyses is small. When the number of clusters is small (generally <50), recent literature has shown that common approaches to correct for correlated errors, such as the cluster-robust sandwich variance estimator, may be biased downwards,^{9–11} resulting in SEs that are too small and confidence intervals (CIs) that are too narrow.

A range of empirical approaches to deal with these challenges have been proposed including bias-adjusted generalized estimating equations (GEE),^{12–15} bootstrapping methods,^{16–18} permutation tests,^{19–21} and aggregation.^{7,10} Although prior work has shown the strength of each approach compared with 1 or 2 alternatives, we attempt to provide a more comprehensive picture of the relative advantages and disadvantages of each approach across a wide range of data scenarios in an effort to offer guidance to applied researchers. In addition, most existing literature has focused on repeated cross-sectional data, which is most commonly used for economic outcomes such as income or hours worked.^{7,16,17,22–24} Although some cross-sectional data are available for health research, medical and epidemiological research more typically focuses on a small number of units repeatedly observed over time in longitudinal datasets.^{3,6,25}

In this paper, we simulate such longitudinal datasets and assess the relative performance of correction methods in

From the *Interfaculty Initiative in Health Policy, Harvard University, Cambridge, MA; †Geary Institute for Public Policy, University College Dublin, Dublin, Ireland; ‡Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA; §Department of Health Care Policy, Harvard Medical School, Boston, MA; and ||Swiss Tropical and Public Health Institute and University of Basel, Basel, Switzerland.

This work was presented by Dr S.R. at the 2016 Irish Economic Association annual meeting and won the Conniffe prize for the best paper by a young economist.

The authors declare no conflict of interest.

Reprints: Slawa Rokicki, PhD, UCD Geary Institute, University College Dublin, Belfield, Dublin 4, Ireland. E-mail: slawa.rokicki@ucd.ie.

Supplemental Digital Content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website, www.lww-medicalcare.com.

Copyright © 2017 Wolters Kluwer Health, Inc. All rights reserved.
ISSN: 0025-7079/18/5601-0097

terms of coverage and power. We first review the most commonly used modeling solutions in DID estimation for panel data, including GEE, permutation tests, clustered standard errors (CSE), wild cluster bootstrapping, and aggregation. Second, we compare the empirical performance of these methods using a Monte Carlo simulation study, testing scenarios in which we vary the degree of error correlation, group size balance, and the proportion of treated groups. We compare both empirical coverage rates and power across all methods. Third, to illustrate the generalizability of our findings to real world settings, we also provide an empirical example using longitudinal data from the Survey of Health, Ageing and Retirement in Europe (SHARE).

MODELING APPROACHES IN DID

Conceptual Review

The main idea of DID is to compare relative trends in treatment and control groups before and after group-level changes.¹ The central aim of DID is causal inference; the basic assumption required for unbiased DID estimates is that of parallel trends in outcomes, that is, the treatment group would have had a trend parallel to the control group in the posttreatment period, had it not been treated. In this article we assume this assumption holds (so that point estimates are unbiased) and then explore various serial correlation scenarios to assess the relative performance of SE corrections proposed in the literature.

Conceptually, the approaches used to account for within-group correlation in outcomes can be divided into 3 broad categories: (1) post hoc adjustments such as CSE, bootstrapping, or permutation tests; (2) explicitly modeling the within-cluster error correlation; and (3) aggregating the data to the group level, thereby eliminating the correlation.

Post hoc Adjustments

Three common post hoc adjustments for SEs in regression models are CSE, cluster bootstrapping, and permutation tests. CSE are a generalization of the White robust covariance sandwich estimator that allow for group-level correlation (clustering) in addition to heteroscedasticity.^{8,26} The technical details for estimating the cluster-robust variance matrix after an ordinary least squares (OLS) regression is shown in Appendix Table 1 (Supplemental Digital Content 1, <http://links.lww.com/MLR/B486>). However, CSE have been shown to perform poorly in scenarios with a small number of clusters because the robust variance estimator is based on a sample variance estimate and residuals tend to underestimate the true error in small samples.^{9,27}

Wild cluster bootstrapping is a modification to the cluster bootstrapping resampling method. Cluster bootstrapping has been shown to be problematic in settings where the treatment variable of interest is binary and cluster invariant.¹⁶ Details of the wild cluster bootstrap procedure are provided in Appendix Table 1 (Supplemental Digital Content 1, <http://links.lww.com/MLR/B486>).

Permutation tests (also called randomization inference) are nonparametric resampling methods.^{19–21,28} They have been more recently applied to quasi-experimental settings.^{23,29–31} The procedure reassigns entire groups to either treatment or

control and recalculates the treatment effect in each reassigned sample, generating a randomization distribution. An exact *P*-value can be calculated as the probability of obtaining a test statistic as far or further from the observed.³¹

Modeling Within-Cluster Error Correlation

There are a number of ways to model within-cluster error correlation including GEE, random effects models, and feasible generalized least squares. Although random effects models and feasible generalized least squares depend on correctly specified error structures, the GEE sandwich estimator is robust with respect to misspecification of the generally unknown covariance structure.^{8,32,33}

There are 2 main problems with the GEE in small samples. First, as with CSE, variance estimates are biased downward; this bias gets larger as the number of groups gets smaller and can be estimated and adjusted for using a Taylor series approximation. Second, the *z*-distribution is a poor approximation of the sampling distribution in small samples and leads to overrejection of the null; a *t*-distribution has been shown to be a better approximation.^{12–15,34,35}

Aggregation

In aggregation, data are collapsed into group cells pre-intervention and post-intervention, thus eliminating the error correlation. Parameters are estimated by first averaging residuals, at the group-time level, from a regression of the outcome on control variables, and using these averaged residuals as the outcome in a group-level DID regression model.⁷ OLS SEs are obtained.

The Additional Problem of Unbalanced Data

Because of a variety of reasons such as differential sampling and attrition, virtually all data available to health researchers tends to be unbalanced, meaning that the number of observations varies across groups and individuals.³⁶ Previous work suggests that in unbalanced data, false discovery rates may be higher than in balanced data for CSE^{17,37,38} as well as for GEE.¹⁵ Carter et al³⁷ demonstrate that the effective number of clusters is reduced when the cluster size varies and provide a measure for calculating this effective number of clusters (G^*) that scales down the true number of clusters (G). MacKinnon and Webb¹⁷ use this measure to produce critical values from the $t(G^*-1)$ distribution and compare false discovery rates with those from the usual $t(G-1)$ distribution. They find that the $t(G^*-1)$ distribution frequently (though not always) reduces rates of false discovery.

In addition, Conley and Taber²³ show that the proportion of treatment groups also impacts false discovery rates in simulation studies. They show that when this proportion is very low (or very high), the treatment effect, though unbiased, is no longer consistent (see full explanation and proof in Conley and Taber²³).

SIMULATION STUDY

We investigated the accuracy of inference for these various approaches by conducting a set of Monte Carlo simulations across a range of scenarios.

We assumed the data generating process (DGP) was known with certainty and given by:

$$Y_{igt} = \beta \text{Trt}_{gt} + u_g + v_i + w_{gt} + \varepsilon_{igt}, \quad (1)$$

with

$$u_g \sim N(0, \sigma_u^2); v_i \sim N(0, \sigma_v^2);$$

$$w_{gt} \sim \text{AR}(1) \text{ with } N(0, \sigma_w^2); \varepsilon_{igt} \sim N(0, \sigma_\varepsilon^2),$$

where Y_{igt} is the outcome for individual i in group g at time t . Trt_{gt} is the indicator for whether the intervention affected group g at time t and β is the DID estimand. u_g and w_{gt} are group-level random effects, while v_i is an individual-level random effect. Via this DGP, the error is correlated within groups and within individuals as normally distributed disturbances, as well as within groups by a first-order autoregressive [AR(1)] process with normal disturbances and an autocorrelation parameter of $\rho = 0.8$. The AR(1) process allows data to be serially correlated across time within groups, as in the way country-specific economic or health conditions evolve progressively over time. Bertrand et al⁷ show that this AR(1) process is too simple to be realistic in panel data; however, they find it is illustrative of the problems in serial correlation and we follow the same process.

Note that if σ_w^2 is 0 or near 0, then individual-level fixed effects will account fully for the within-cluster correlation as the correlation of errors is then driven solely by group- and individual-level processes. However, previous research has shown that the inclusion of group fixed effects in group-year panel data does not eliminate the within-group correlation of the error.^{7,9,24} Thus our DGP induces correlation in the error even after accounting for individual fixed effects.

We tested both low and high correlation scenarios. Similar to Donald and Lang,¹⁰ in the low correlation scenario, we set $\sigma_\varepsilon^2 = 10\sigma_v^2 = 100\sigma_u^2 = 100\sigma_w^2 = 1$. In the high correlation scenario, we set $\sigma_u^2 = \sigma_w^2 = 0.05$, and $\sigma_v^2 = 0.15$. Although our DGP is unique, our intraclass correlations are similar to those of other studies.^{10,13,39}

The list of simulation scenarios is shown in Table 1. We tested both short panels, where we set the number of time points per individual to 4 and long panels where we set the number of time points per individual to 20. The treatment was implemented at the halfway point. We began our simulations with balanced data, where the number of individuals per group was always 30 and the proportion of treated groups was 0.5. Next, we tested the case with unbalanced cluster sizes, where we allowed the number of individuals per group to vary on a uniform distribution between 1 and 59 (for a mean of 30, yielding a coefficient of variation of 0.56). Finally, we tested the case in which the proportion of treated groups was 0.2.

For each scenario, we simulated 1000 datasets under the null treatment effect. We evaluated the performance of the methods detailed below by the coverage rate, the fraction of

simulations in which the 95% CI for β covers the null (in the permutation test and wild cluster bootstrap, we calculated the fraction of simulations in which the P -value is ≥ 0.05). Coverage rates below 0.95 indicate underestimation of SEs and P -values, whereas coverage rates above 0.95 indicate overestimation of SEs; satisfactory performance of models implies that actual coverage rates are close (within the Monte Carlo CI) to the nominal coverage rate of 0.95.

Next, we imposed a treatment effect of 0.6 standard deviations. We again simulated 1000 datasets and we evaluated the performance of the models by the measure of statistical power, the fraction of the simulations that resulted in a significant effect at the 0.05 level.

We tested 6 estimation methods, as follows. We began with the basic DID model:

$$Y_{igt} = a + b_1(\text{GroupTrt}_g * \text{PostTrt}_t) + b_2\text{GroupTrt}_g + b_3\text{GroupTrt}_t + \varepsilon_{igt}, \quad (2)$$

where GroupTrt_g is the indicator for whether the group was treated, PostTrt_t is the indicator for the posttreatment period, and $\text{GroupTrt}_g * \text{PostTrt}_t$ is their interaction. Using this model, we estimated CSE at the group level, wild cluster bootstrap, and permutation tests (see Appendix Table 1, Supplemental Digital Content 1, <http://links.lww.com/MLR/B486> for details). Next we included individual fixed effects, A_i , instead of the intercept, a , and again estimated CSE at the group level. We next collapsed the data into group-time cells and estimated OLS SEs. Finally, we estimated a GEE with the same specification as Equation 2, assuming a normal distribution for the response, the identity as link function, the group as the cluster ID, and an exchangeable working correlation matrix. We adjusted the GEE with small sample bias adjustment and an F -distribution correction as per Fay and Graubard.¹⁴

All simulations were conducted using R, version 3.2.3. The R code needed to implement the methods tested is provided in Supplemental Digital Content 2 (<http://links.lww.com/MLR/B487>).

RESULTS

Simulation Results for Coverage Rates

Figure 1 presents the results of our simulations for all 6 methods in the high correlation scenario when the number of time points per individual is 4. The horizontal line is the nominal coverage of 0.95 and the horizontal dotted lines indicate the Monte Carlo CI. The figure shows coverage rates as the number of groups increases from 5 to 50 for data that are balanced with respect to cluster size, are unbalanced with respect to cluster size, and have a low proportion of treated clusters.

When data were balanced, most models produced coverage rates close to 0.95 as long as the number of groups, G , was at least 7. With short panels (only 4 time points), individual fixed

TABLE 1. Characteristics of Simulation Scenarios

Simulation Scenario	Correlation	Individuals Per Cluster	Proportion of Treated Clusters	Time Points Per Individual
Balanced data	Low and High	30	0.5	4 and 20
Unbalanced cluster size	Low and High	1–59	0.5	4 and 20
Low proportion of treated clusters	Low and High	30	0.2	4 and 20

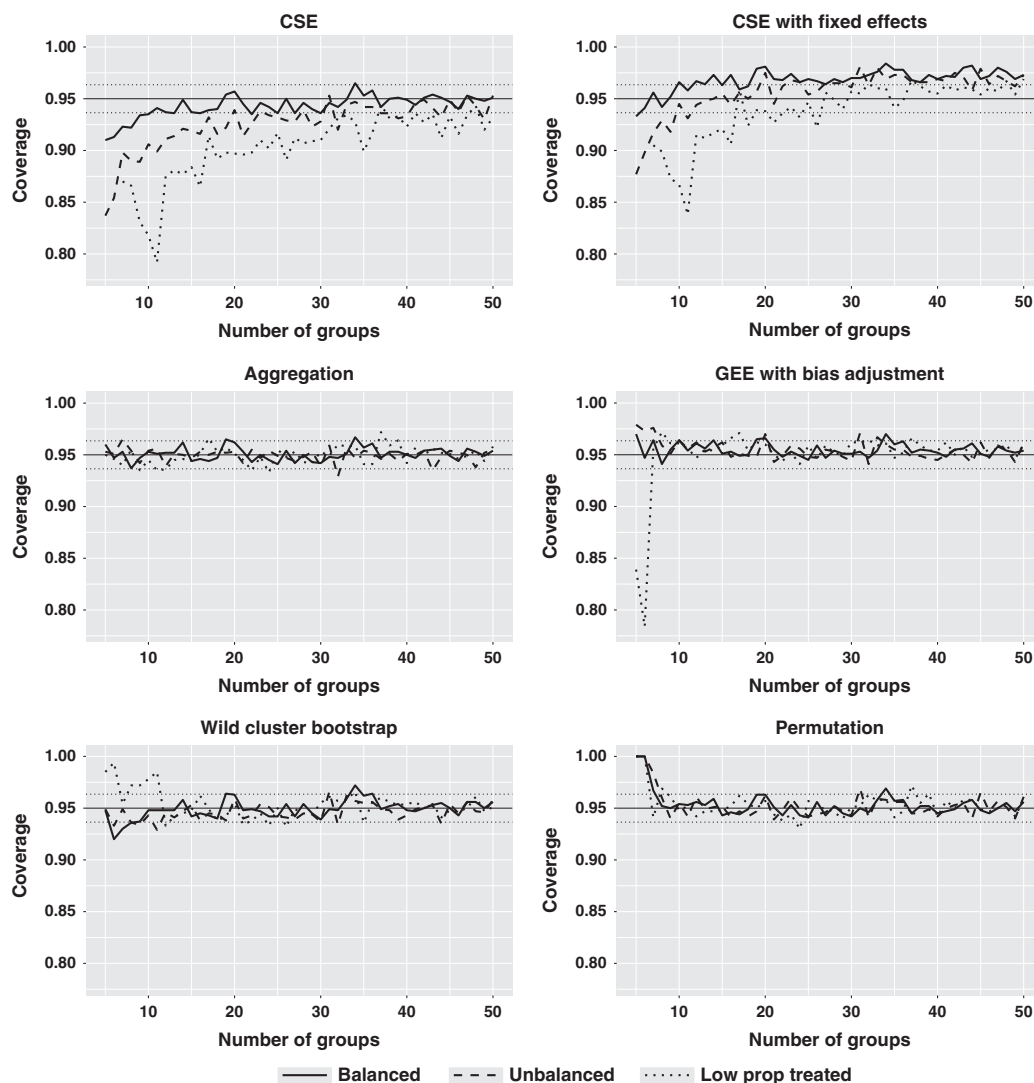


FIGURE 1. Coverage for 6 models as number of groups increases for data that are balanced, unbalanced, and with a low proportion of treated clusters, in the high correlation scenario with 4 time points per individual. Horizontal lines show 0.95, the nominal coverage, and Monte Carlo simulation confidence intervals. For the low proportion of treated case, coverage for CSE is off of the graph for $G=5$ and $G=6$, at 0.68 and 0.64, respectively, and for CSE with individual fixed effects at 0.72 and 0.70, respectively. CSE indicates clustered standard errors; GEE, generalized estimating equations.

effects accounted for most of the variation at the group level and CSE with individual fixed effects produced satisfactory, though slightly conservative, coverage in the balanced case (panel B).

However, results substantially changed when data were unbalanced and when there were a low proportion of treated clusters. In unbalanced data, CSE, even with individual fixed effects, had lower than nominal coverage up to $G=10$. In the low proportion of treated clusters scenario, CSE with fixed effects had lower than nominal coverage even up to $G=18$. It is important to note here that coverage rates do not increase monotonically with G because the finite number of groups did not allow us to keep the proportion of treated clusters constant. For example, when G was 7 the number of treated clusters was 2, resulting in a proportion of about 0.28, whereas when G was 10, the number of treated clusters was still 2 and thus the

proportion was 0.2. The results highlight that both the absolute number of clusters as well as proportion of treated clusters are significant influences on the performance of CSE.

Aggregation (panel C) and permutation (panel F) consistently produced coverage rates very close to 0.95 regardless of balance of data or proportion of treated clusters, aside from permutation when $G < 7$ which produced a coverage of 1 due to the limited number of permutations of the data resulting in P -values necessarily > 0.05 . The adjusted GEE was also consistently satisfactory, aside from the case when $G < 7$ in the low proportion of treated scenario (panel D). This occurred because there was only 1 treated cluster in those cases and the variance matrix estimate of the GEE relies on averaging residuals across clusters.

The wild cluster bootstrap also performed well except in the low proportion of treated clusters scenario, where it produced

conservative coverage rates when $G < 12$ (panel E). This may be due to the limited possible number of transformations of bootstrap residuals when there are few (or almost all) clusters treated; Webb¹⁸ finds that a different weight distribution (such as the Webb 6-point distribution rather than the Rademacher 2-point distribution used here) performs better in very small G .

Results were similar when we increased the number of time points to 20 per individual in the high correlation scenario (Fig. 2). However, in this case, the data were more highly autocorrelated in the AR(1) group-time process, and thus individual fixed effects could no longer control for the correlation in the errors. CSE with fixed effects led to coverage rates considerably below nominal level in balanced data when $G < 9$, in unbalanced data when $G < 22$, and in data with low proportion of treated clusters when $G < 50$.

Other models performed much better. Aggregation, the adjusted GEE, and permutation had coverage rates close to 0.95 regardless of balance of data or proportion of treated clusters, with the minor exceptions mentioned above. The results for the same scenarios with low correlation are shown in Appendix Figures 2, 3 (Supplemental Digital Content 1, <http://links.lww.com/MLR/B486>).

Simulation Results for Statistical Power

We investigated the power of these models to detect a treatment effect at the 0.05 level in scenarios in which the data are unbalanced (Fig. 3A) and have a low proportion of treated clusters (Fig. 3B). All methods resulted in unbiased treatment effects (see Appendix Fig. 4, Supplemental Digital Content 1, <http://links.lww.com/MLR/B486>). The graphs show

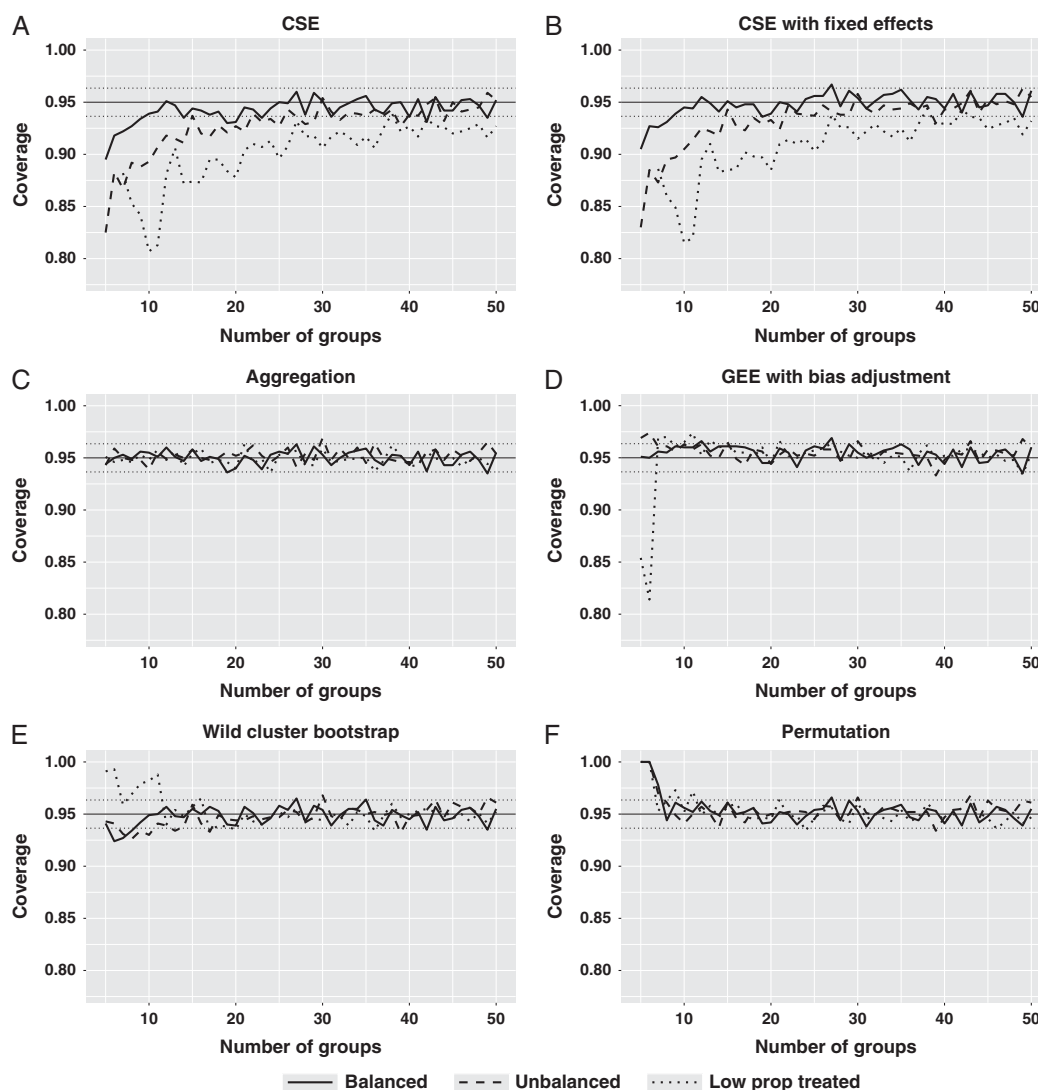


FIGURE 2. Coverage for 6 models as number of groups increases for data that are balanced, unbalanced, and with a low proportion of treated clusters, in the high correlation scenario with 20 time points per individual. Horizontal lines show 0.95, the nominal coverage, and Monte Carlo simulation confidence intervals. For the low proportion of treated case, coverage for CSE is off of the graph for $G = 5$ and $G = 6$, at 0.68 and 0.64, respectively, and for CSE with individual fixed effects at 0.69 and 0.65, respectively. CSE indicates clustered standard errors; GEE, generalized estimating equations.

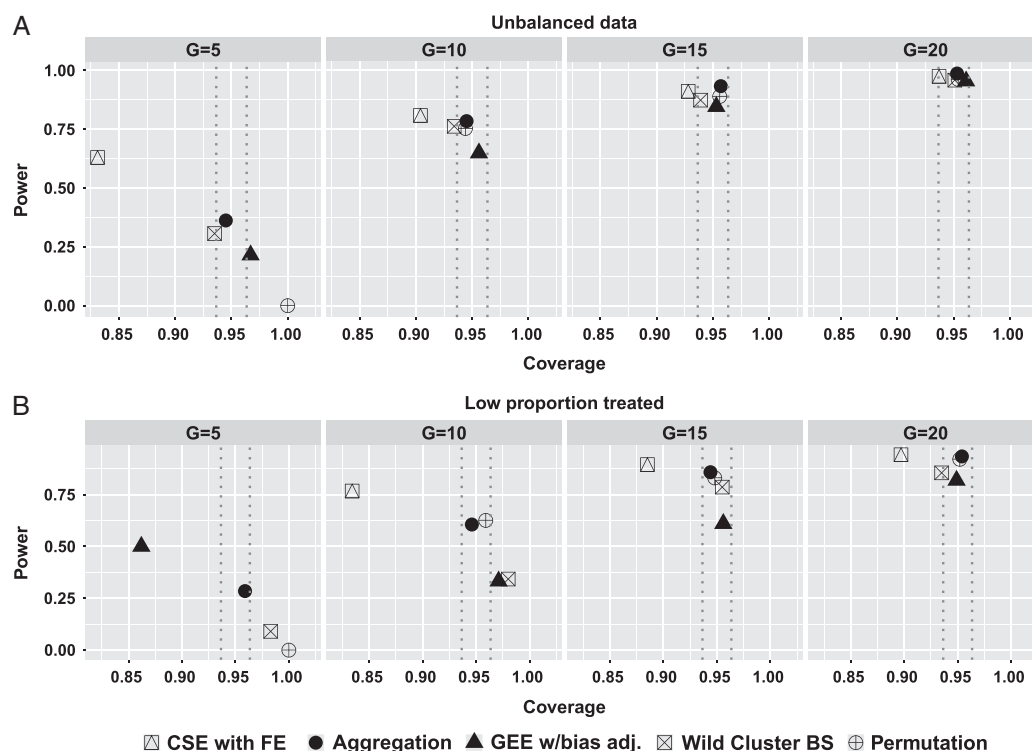


FIGURE 3. Power versus coverage for unbalanced data (panel A) and low proportion of treated clusters (panel B), by number of groups (G). Number of time points for each individual is 20. Dotted lines indicate Monte Carlo confidence intervals for nominal coverage. Monte Carlo confidence intervals for power are not shown to prevent obscurity of results; for each estimate the width of the 95% confidence interval is 0.0196. For Panel B, G=5, CSE with FE coverage is off the graph at 0.69. CSE with FE indicates clustered standard errors with individual fixed effects; Wild Cluster BS, wild cluster bootstrap; GEE w/bias adj, generalized estimating equations with bias adjustment.

coverage rates on the *x*-axis and power on the *y*-axis for 5, 10, 15, and 20 groups. For both data scenarios, we found that aggregation and permutation provided the most power among those models that also met the coverage criterion, though permutation had no power to detect an effect at the 0.05 level when $G=5$ because of the limited number of total possible permutations. As it is more conservative than the other methods,¹⁴ the adjusted GEE was consistently underpowered compared with other methods.

Empirical Example

We investigate the generalizability of the results of our simulations to real world empirical settings using data from SHARE.^{40–42} SHARE is a widely used and cited cross-national longitudinal survey of health and socioeconomic status. The target population for SHARE is persons who are 50 years and older in the respective survey year and their partners of any age. The survey has a longitudinal dimension in that all respondents who have previously participated are eligible to be interviewed in future waves. Recently, DID analyses exploiting country-level differences using SHARE data have been conducted to examine the effect of the recession on elderly informal care receipt,⁴³ maternity leave benefits on mental health,⁴⁴ and health service user fees implementation on health care utilization.⁴⁵ In these analyses, we may be

worried that institutional and cohort factors may drive country-level autocorrelation in DID model errors.

We extract data from the easySHARE combined SHARE dataset and focus on the 9 countries included in all 5 waves.^{40,42} The sample includes 129,764 observations from 54,854 individuals after missing data are excluded.

We first assess the extent of autocorrelation in SHARE health outcomes as compared with our simulated data. Using the procedure outlined in Bertrand et al,⁷ we calculate mean country-wave residuals from a regression of each outcome on country and wave dummies; the autocorrelation coefficients are obtained from a linear regression of the residuals on the lagged residuals. For body mass index (BMI), word recall, and depression scale, the average estimated first-order autocorrelation coefficients are 0.36, 0.24, and 0.38, respectively (Appendix Table 2, Supplemental Digital Content 1, <http://links.lww.com/MLR/B486>). These are quite comparable with the autocorrelation of our simulated data in the high correlation, unbalanced scenario estimated at 0.37. Conversely, for grip strength and subjective wellbeing, the autocorrelation coefficients are near 0. This is perhaps because these measures are not as responsive to country-specific trends over time, so that country and wave fixed effects are effective at eliminating autocorrelation in the residuals.

Next, we assess how similar our simulated results are to results from real data, focusing on the outcome of BMI. The

procedure is as follows: we first resample countries with replacement to get a new sample of 9 countries (preserving the within-country error structure), then we sample 10% of individuals within each country (including all of each individual's measurements). For each sample, we create a placebo intervention that occurs between waves 2 and 4 for some proportion of the countries, and run the same DID models as in the simulated data, but additionally adjusting for sex, age, years of education, and marital status. We evaluate an additional model where we include country and wave fixed effects in the DID regression before applying CSE. We conduct the procedure 1000 times and calculate coverage for all models. We vary the proportion of treated countries, r , from 0.11 to 0.89. The results are shown in Figure 4.

Results are quite similar to those of the simulations with the short panel. CSE, even with country and wave fixed effects, produced lower than nominal coverage and was particularly poor when r was close to 0 or 1. As the panel is relatively short, CSE performed much better when individual fixed effects were included, although coverage was still less than the nominal rate in cases when $r < 0.25$ (ie, number of treated countries < 3). As in the simulations, aggregation and permutation produced coverage rates close to 0.95 regardless of proportion treated. The wild cluster bootstrap performed well, except in the case when r was close to 0 or 1 when it was conservative. GEE also performed well, except in the case of 1 treated or 1 control cluster.

DISCUSSION

In this paper, we reviewed a range of empirical strategies proposed in the recent statistics literature to address the likely high degree of within-group error correlation in longitudinal data used for DID estimation. Our results suggest that CSE, one of the most commonly used strategies, yield CIs that are systematically too narrow in scenarios when data are unbalanced or when there is a low proportion of treated groups. Inclusion of individual fixed effects can somewhat improve coverage rates when applying CSE in short panels; however, they are not effective in longer panels. In contrast, aggregation, the adjusted GEE, and permutation tests consistently produce coverage rates close to the nominal rate of 0.95 regardless of balance of data, aside from the adjusted GEE in the case when there is only 1 treated cluster and permutation in the case when number of groups is < 7 . With a very small number of groups (< 12), the wild cluster bootstrap yields slightly lower than nominal coverage in balanced and unbalanced data, and higher than nominal coverage in the low proportion of treated scenario.

To illustrate the practical relevance of our results, we estimated the same range of models using real data from the SHARE study. We found very similar results for the outcome of BMI: CSE consistently resulted in overrejection of the null. As the panel was relatively short, individual fixed effects were able to reduce the error correlation. However, CSE still resulted in severe overrejection when

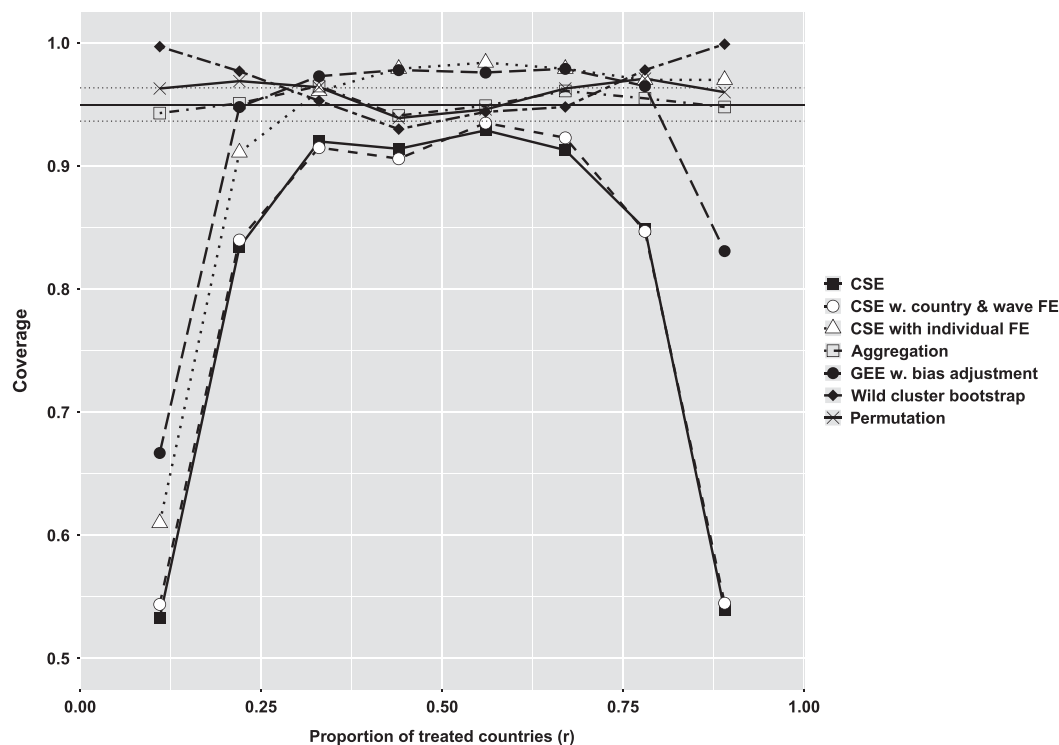


FIGURE 4. Coverage rates for 7 models as proportion of treated countries varies, using SHARE data for outcome of BMI. All models adjusted for sex, age, years of education, and marital status. CSE indicates clustered standard errors; FE, fixed effects; GEE, generalized estimating equations.

the proportion of treated countries was low. In contrast, aggregation and permutation resulted in correct coverage rates in all scenarios.

The main challenge with all methods that seem to work well is power, especially when the number of groups is ≤ 10 . In relative terms, aggregation and permutation seem to perform best in this setting, whereas the power of the bias-adjusted GEE is limited.

This analysis has some limitations. In all simulation studies it is necessary to specify a DGP; we can only be sure that our results hold under the conditions of that unique process. Since in real data we do not observe what the DGP is, we are cautious about generalizing our results. Our empirical example using SHARE data provides some evidence that even under alternative DGPs with different error structures, our results in short panels hold. However, more empirical work using longer panels with more diverse health outcomes and treatment scenarios is necessary.

Nevertheless, these results have important implications for medical and epidemiological research. In real data, it is not possible to know what the true DGP is; researchers should therefore err on the side of caution when applying CSE in DID estimation using longitudinal data with few clusters, particularly when data are not balanced or when there is a low proportion of treated clusters. Reviewers of articles that include small sample clustering should request that authors use appropriate methods, or at minimum compare their findings to either aggregation, permutation tests, GEE with bias adjustment, or the wild cluster bootstrap. Second, although the adjusted GEE provides accurate coverage, it appears to have low power in DID estimation in small samples; researchers may consider permutation or aggregation as alternative methods. Third, as randomized controlled trials are increasingly analyzed using DID, researchers can maximize power and avoid low coverage by designing cluster-randomized trials with equally sized clusters.^{36,39}

Lastly, these findings also have important implications for public policy. Correctly adjusting for correlated data is critical for rigorous evaluation of public programs. Evaluations that find a spurious positive or negative effect of a policy due to inappropriate methodology may promote poor public policy-making.

ACKNOWLEDGMENTS

The authors thank Mark McGovern and Laura Hatfield for their helpful comments and suggestions. This paper uses data from SHARE Waves 1, 2, 3 (SHARELIFE), 4, 5, and 6 (DOIs: 10.6103/SHARE.w1.600, 10.6103/SHARE.w2.600, 10.6103/SHARE.w3.600, 10.6103/SHARE.w4.600, 10.6103/SHARE.w5.600, 10.6103/SHARE.w6.600), see Börsch-Supan and colleagues (2013) for methodological details. The SHARE data collection has been primarily funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), and FP7 (SHARE-PREP: No.211909, SHARE-LEAP: No.227822, SHARE M4: No.261982). Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of

Science, the US National Institute on Aging (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C) and from various national funding sources is gratefully acknowledged (www.share-project.org). This paper uses data from the generated easySHARE dataset (DOI: 10.6103/SHARE.easy.600), see Gruber and colleagues (2014) for methodological details. The easySHARE release 6.0.0 is based on SHARE Waves 1, 2, 3 (SHARELIFE), 4, 5, and 6 (DOIs: 10.6103/SHARE.w1.600, 10.6103/SHARE.w2.600, 10.6103/SHARE.w3.600, 10.6103/SHARE.w4.600, 10.6103/SHARE.w5.600, 10.6103/SHARE.w6.600).

REFERENCES

1. Dimick JB, Ryan AM. Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA*. 2014;312:2401–2402.
2. Estee S, Wickizer T, He L, et al. Evaluation of the Washington State Screening, brief intervention, and referral to treatment project: cost outcomes for medicaid patients screened in hospital emergency departments. *Med Care*. 2010;48:18–24.
3. Shortell SM, Gillies R, Siddique J, et al. Improving chronic illness care: a longitudinal cohort analysis of large physician organizations. *Med Care*. 2009;47:932–939.
4. Zivin K, Pfeiffer PN, Szymanski BR, et al. Initiation of primary care—mental health integration programs in the va health system: associations with psychiatric diagnoses in primary care. *Med Care*. 2010;48:843–851.
5. Werner RM, Duggan M, Duey K, et al. The patient-centered medical home: an evaluation of a single private payer demonstration in new jersey. *Med Care*. 2013;51:487–493.
6. McGovern ME, Herbst K, Tanser F, et al. Do gifts increase consent to home-based HIV testing? a difference-in-differences study in rural KwaZulu-Natal, South Africa. *Int J Epidemiol*. 2016;45:2100–2109.
7. Bertrand M, Duflo E, Mullainathan S. How much should we trust differences-in-differences estimates? *Q J Econ*. 2004;119:249–275.
8. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73:13–22.
9. Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. *J Hum Resour*. 2015;50:317–372.
10. Donald SG, Lang K. Inference with difference-in-differences and other panel data. *Rev Econ Stat*. 2007;89:221–233.
11. McCaffrey DF, Bell RM. Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters. *Stat Med*. 2006;25:4081–4098.
12. Morel JG, Bokossa MC, Neerchal NK. Small sample correction for the variance of GEE estimators. *Biom J*. 2003;45:395–409.
13. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics*. 2001;57:126–134.
14. Fay MP, Graubard BI. Small-sample adjustments for wald-type tests using sandwich estimators. *Biometrics*. 2001;57:1198–1206.
15. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Stat Med*. 2002;21:1429–1441.
16. Cameron AC, Gelbach JB, Miller DL. Bootstrap-based improvements for inference with clustered errors. *Rev Econ Stat*. 2008;90:414–427.
17. MacKinnon JG, Webb MD. Wild bootstrap inference for wildly different cluster sizes. *J Appl Econ*. 2017;32:233–254.
18. Webb MD. Reworking wild bootstrap based inference for clustered errors. Queen's Economics Department Working Paper, Report No.: 1315; 2013. Available at: www.econstor.eu/handle/10419/97480. Accessed January 15, 2016.
19. Fisher RA. *The design of experiments*. Oliver and Boyd: Edinburgh; 1935.
20. Rosenbaum PR. Covariance Adjustment in Randomized Experiments and Observational Studies. *Stat Sci*. 2002;17:286–327.
21. Ernst MD. Permutation methods: a basis for exact inference. *Stat Sci*. 2004;19:676–685.
22. Cameron AC, Miller DL. Robust inference with clustered data. Working Papers, University of California, Department of Economics; 2010. Available at: www.econstor.eu/handle/10419/58373. Accessed June 28, 2017.

23. Conley TG, Taber CR. Inference with “difference in differences” with a small number of policy changes. *Rev Econ Stat*. 2011;93:113–125.
24. Brewer M, Crossley TF, Joyce R. Inference with difference-in-differences revisited, Report No.: ID 2363229. Rochester, NY: Social Science Research Network; 2013. Available at: <https://papers.ssrn.com/abstract=2363229>. Accessed June 28, 2017.
25. Datar A, Sturm R. Physical education in elementary school and body mass index: evidence from the early childhood longitudinal study. *Am J Public Health*. 2004;94:1501–1506.
26. White H. A Heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*. 1980;48:817–838.
27. Wooldridge JM. Cluster-sample methods in applied econometrics. *Am Econ Rev*. 2003;93:133–138.
28. Cohen J, Dupas P. Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment. *Q J Econ*. 2010;125:1–45.
29. Bloom E, Bhushan I, Clingingsmith D, et al. Contracting for health: evidence from Cambodia. Brookings Institution; 2006. Available at: www.webprodserv.brookings.edu/~media/Files/rc/papers/2006/07healthcare_kremer/20060720cambodia.pdf. Accessed June 28, 2017.
30. Ho DE, Imai K. Randomization inference with natural experiments: an analysis of ballot effects in the 2003 California recall election. *J Am Stat Assoc*. 2006;101:888–900.
31. Ryan AM, Burgess JF, Dimick JB. Why we should not be indifferent to specification choices for difference-in-differences. *Health Serv Res*. 2014;50:1211–1235.
32. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*. 1986;42:121–130.
33. Peters TJ, Richards SH, Bankhead CR, et al. Comparison of methods for analysing cluster randomized trials: an example involving a factorial design. *Int J Epidemiol*. 2003;32:840–846.
34. Bell RM, McCaffrey DF. Bias reduction in standard errors for linear regression with multi-stage samples. *Surv Methodol*. 2002;28:169–182.
35. Gunsolley JC, Getchell C, Chinchilli VM. Small sample characteristics of generalized estimating equations. *Commun Stat Simul Comput*. 1995;24:869–878.
36. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol*. 2006;35:1292–1300.
37. Carter AV, Schnepel KT, Steigerwald DG. Asymptotic behavior of a t-test robust to cluster heterogeneity. *Rev Econ Stat*. 2017;99.4:698–709.
38. Imbens GW, Kolesar M. Robust standard errors in small samples: some practical advice. *Rev Econ Stat*. 2016;98:701–712.
39. Rutterford C, Copas A, Eldridge S. Methods for sample size determination in cluster randomized trials. *Int J Epidemiol*. 2015;44:1051–1067.
40. Börsch-Supan A, Gruber S, Hunkler C, et al. easySHARE. Release version: 6.0.0. SHARE-ERIC. Dataset. doi: 10.6103/SHARE.easy.600.
41. Börsch-Supan A, Brandt M, Hunkler C, et al. Data resource profile: the survey of health, ageing and retirement in europe (SHARE). *Int J Epidemiol*. 2013;42:992–1001.
42. Gruber S, Hunkler C, Stuck S. *Generating easySHARE: Guidelines, Structure, Content and Programming*. (SHARE Working Paper Series: 17-2014). Munich: MEA, Max Planck Institute for Social Law and Social Policy; 2014.
43. Costa-Font J, Karlsson M, Oien H. Careful in the crisis? determinants of older people’s informal care receipt in crisis-struck european countries. *Health Econ*. 2016;25 (S2):25.
44. Avendano M, Berkman LF, Brugiavini A, et al. The long-run effect of maternity leave benefits on mental health: evidence from european countries. *Soc Sci Med*. 2015;132:45–53.
45. Kalousova L. Curing over-use by prescribing fees: an evaluation of the effect of user fees’ implementation on healthcare use in the Czech Republic. *Health Policy Plan*. 2015;30:423–431.