Did Massachusetts Health Care Reform Lower Mortality? No According to Randomization Inference

Robert KAESTNER

ABSTRACT
In an earlier article, Sommers, Long, and Baicker concluded that health care reform in Massachusetts was associated with a significant decrease in mortality. I replicate the findings from this study and present p-values for the parameter estimates reported by Sommers, Long, and Baicker that are based on an alternative and valid approach to inference referred to as randomization inference. I find that estimates of the treatment effects produced by Sommers, Long, and Baicker are not statistically significant when p-values are based on randomization inference methods. Indeed, the p-values of the estimates reported in Sommers, Long, and Baicker derived by the randomization inference method range from 0.22 to 0.78. Therefore, the authors’ conclusion that health reform in Massachusetts was associated with a decline in mortality is not justified. The Sommers, Long, and Baicker analysis is largely uninformative with respect to the true effect of reform on mortality because it does not have adequate statistical power to detect plausible effect sizes.

Introduction
Health care reform in Massachusetts, and the consequences of that reform, has garnered a great deal of attention from policy analysts and the media because the Massachusetts reform is widely considered a forerunner and model of the broader federal reform enacted as part of the Affordable Care Act (ACA). There have been several analyses of the consequences of the Massachusetts reform, and one of the most important was by Sommers, Long, and Baicker (2014) who examined whether health care reform in Massachusetts was associated with changes in mortality. Based on the results of their analysis, Sommers, Long, and Baicker (2014) concluded: “Health reform in Massachusetts was associated with significant reductions in all-cause mortality and deaths from causes amenable to health care” (p. 585).

If true, the findings reported by Sommers, Long, and Baicker (2014) are quite important, as there is relatively little evidence of a causal link between health insurance coverage and mortality. Moreover, because of its similarity with the Affordable Care Act, the consequences of health care reform in Massachusetts are often used as predictors of the consequences of the Affordable Care Act. The significance of the Sommers, Long, and Baicker (2014) finding in the public sphere is reflected in the widespread media attention the article received (e.g., “Mortality Drop Seen to Follow ’06 Health Law, The New York Times, May 5, 2014).

However, it would be equally important to the policy debate if the Sommers, Long, and Baicker (2014) was not accurate, and there are plausible reasons to expect that the inferences drawn by Sommers, Long, and Baicker (2014) were misleading. To study the question of whether health care reform in Massachusetts was associated with mortality, Sommers, Long, and Baicker (2014) used a pre and post-test with comparison group research design, which is commonly referred to as a difference-in-differences approach. In this approach, Massachusetts was the treatment state and other states were control states. One problem with this approach is that standard methods of statistical inference are likely to be invalid in this context. Conley and Taber (2011) showed that when there is only one or two treated units (e.g., states) and many control units, standard errors of estimates using common methods such as cluster-robust standard errors (CRSE) may be severely biased. Notably, standard errors of estimates in Sommers, Long, and Baicker (2014) were constructed using CRSE. Thus, it is worth assessing whether the inferences drawn from the study remain the same when more appropriate methods of inference are used.

In this article, I replicate the findings of Sommers, Long, and Baicker (2014) and present p-values for the parameter estimates that are based on an alternative and valid approach to inference referred to as randomization inference (also referred to as permutation test). Randomization inference is a nonparametric approach to estimating the significance level of treatment effects that is appropriate even in cases where there is only one treatment group (Fisher 1935; Rosenbaum 2002; Conley and Taber 2011). Randomization inference is an approach that has been widely used in medical science and also, but much less so, in social science and policy analysis applications (Good 2000, 2005; Imbens and Rosenbaum 2005; Ho and Imai 2006; Abadie, Diamond, and Hainmueller 2010; Cohen and Dupas 2010; Bloom et al. 2013; Courtemanche and Zapata 2014; Ryan et al. 2014).

I find that estimates of the treatment effects produced by Sommers, Long, and Baicker (2014) are not statistically significant when p-values are based on randomization inference methods. Indeed, the p-values of the estimates reported in Sommers, Long, and Baicker (2014) derived by the randomization...
inference method range from 0.22 to 0.78. Therefore, the authors’ conclusion that health reform in Massachusetts was associated with a decline in mortality is not justified. The Sommers, Long, and Baicker (2014) analysis is largely uninformative with respect to the true effect of reform on mortality because it does not have adequate statistical power to detect plausible effect sizes. Interestingly, a similar analysis applied to whether Massachusetts healthcare reform decreased the proportion of persons uninsured indicates that inferences associated with previously reported findings, for example, in Long (2008) and Long, Stockley, and Yemane (2009), remain valid even when derived from randomization inference methods.

**Brief Background**

The potential problems with standard methods of conducting statistical tests in difference-in-differences (including experimental designs) analyses have been well studied by statisticians from several disciplines including economics (e.g., Moulton 1990; Bertrand, Duflo, and Mullainathan 2004; Donald and Lang 2007; Cameron, Gelbach, and Miller 2008; Conley and Taber 2011). As this literature shows, problems related to inference may be particularly severe when there are a few clusters (i.e., treated and untreated units), few clusters that are treated, and/or clusters with greatly varying numbers of observations. In these cases, the common approach of constructing standard errors using methods such as CRSE is not reliable (Conley and Taber 2011; MacKinnon and Webb 2014; Cameron and Miller in press). To address these problems, several approaches have been proposed (Cameron, Gelbach, and Miller 2008; Conley and Taber 2011; MacKinnon and Webb 2014; Webb 2014).

In the analysis of Massachusetts health care reform, the problem with the standard method of inference (CRSE) stems from the fact that there is only one treated group. In such cases, Conley and Taber (2011) showed that estimates of the treatment effect from a difference-in-differences design are unbiased, but not consistent. The inconsistency is because there is only one treated group and a fixed number of periods. Therefore, the “noise” of the estimate of the treatment effect does not go toward zero. As a result, standard methods of inference are incorrect because they do not account for the inconsistency of the estimate and use a confidence interval based on a consistent parameter estimate (i.e., confidence interval is too small). Conley and Taber (2011) provided the proof.

To address this problem, Conley and Taber (2011) proposed a method that accounts for the inconsistency by using the distribution of “noise” derived from the control units as an estimate of the distribution of the “noise” that is a component of the inconsistent estimate of the treatment effect. Because the difference between the estimated difference-in-differences parameter and the true estimate is equal to the “noise,” the distribution of the “noise” derived from the control groups is the null distribution—the distribution of the estimate under the assumption that the null hypothesis of no effect is true. Given the null distribution, inferences about the inconsistent parameter estimate can be obtained by comparing the difference-in-difference estimate to the null distribution. Rejecting the null is indicated when the difference-in-difference estimate is in the tails of the null distribution, which implies that the difference-in-differences estimate of the treatment effect is more than just “noise” and therefore “unusual” in a statistical sense.

The Conley and Taber (2011) approach is closely related to randomization inference, which has been widely used in biostatistics and, to a lesser extent, in social sciences (Gail et al. 1996; Good 2000; Braun and Feng 2001; Imbens and Rosenbaum 2005; Ho and Imai 2006; Small, Ten Have, and Rosenbaum 2008; Abadie, Diamond, and Hainmueller 2010; Cohen and Dupas 2010; Bloom et al. 2013; Courtemanche and Zapata 2014). In fact, Conley and Taber (2011) showed that their approach is asymptotically equivalent to randomization inference.

In my analysis, I use both the randomization inference approach and the Conley and Taber (2011) approach, although I focus mainly on the randomization inference method because of its intuitive, conceptual appeal and ease of implementation. (Conley and Taber [2009] presented Monte Carlo results that show that the two approaches produce very similar results.) The intuition of the randomization inference approach is straightforward. If the null hypothesis (i.e., no effect) of the (quasi) experiment is true for all units (i.e., sharp null), then the estimate of the treatment effect does not depend on whether an observation is assigned as a treatment or a control unit. One could switch treatment status labels, reestimate, and obtain the null result. Therefore, it is possible to randomly assign treatment status to observations in the sample and calculate a treatment effect with the expectation that the estimate would be zero. If this random switching of treatment status is done many times, then it will produce a distribution of estimates of the treatment effect under the null hypothesis of no effect, or what is referred to as the null distribution. Given the null distribution, it is possible to assess whether the actual treatment effect obtained from the (quasi) experiment is “surprising”—does the estimate lie in the tails of the null distribution? The p-value of the actual (quasi) experimental estimate is obtained by calculating the proportion of estimates from the null distribution that are larger (in absolute value) than the estimate obtained from the experiment. If all possible permutations of treatment status are included in the analysis, then randomization inference produces an exact test statistic, but if only a sample of possible permutations is included, the test statistic is approximate. The only assumption of randomization inference is that observations (groups) are exchangeable—that the joint distribution of observations is invariant with respect to realignment of treatment status (Good 2005). Randomization will ensure this to be the case, and the differences-in-differences assumes this to be the case. (Conditioning on variables is fine as Gail, Tan, and Piantadosi [1988] and Small et al. [2008] show. In this case, exchangeability refers to the residual after adjusting for covariates.)

The upshot of this brief discussion is that the standard errors and inferences reported by Sommers, Long, and Baicker (2014) are likely to be biased because there is only one treatment group state among the 47 states used in the analysis. As a result, the difference-in-differences estimate of the treatment effect is inconsistent and valid inference requires taking this inconsistency into account. (Inferences will be valid if the error term has the properties of a classical ordinary least-square regression [e.g., normally distributed iid].)
Replication and Randomization Inference

Sommers, Long, and Baicker (2014) conducted two analyses that differ by the unit of observation used. In one, Sommers, Long, and Baicker (2014) used a sample of 513 counties in the United States as a control group for the 14 counties in Massachusetts. I refer to this analysis as the county-level analysis. The second analysis aggregated the county level information to the state and I refer to this as the state-level analysis.

County-Level Analysis

To conduct the randomization inference analysis, I followed Sommers, Long, and Baicker (2014) and used data from vital statistics on deaths that are reported at the county level. These data are available from the Centers for Disease Control (http://www.cdc.gov/nchs/data_access/cmfr.htm). As in Sommers, Long, and Baicker (2014), counties are the geographic unit that determines assignment of treatment status. The 14 counties in Massachusetts were treatment counties; Massachusetts passed health care reform in 2006 and other states presumably had no (major) change in policy. Sommers, Long, and Baicker (2014) selected control units from counties in the rest of the United States using a propensity score matching procedure that matched counties on several dimensions measured at baseline: demographic and socioeconomic characteristics, rates of uninsured, and mortality rates. The top quartile of counties from the propensity score distribution was selected as control units. The propensity score procedure yielded 513 counties in 46 states. I used these counties as controls. (We thank Benjamin Sommers for providing the county FIPS codes for the counties used in Sommers, Long, and Baicker 2014.) Note, however, that the unit of observation in Sommers, Long, and Baicker (2014) was not the county, but the county-age-gender-race-year cell. For example, a typical value of the dependent variable was the number of deaths among white, men, ages 20 to 34 in Norfolk County, MA in 2007.

In most analyses, Sommers, Long, and Baicker (2014) used a general linear model with a negative binomial distribution and a log link function with several covariates, although the authors also reported estimates obtained using ordinary least-square (OLS) methods. (Some models did not use covariates and simply controlled for state (county), year, and interaction between treatment and year.) I was able to replicate exactly most of Sommers, Long, and Baicker (2014) estimates. To obtain the p-values using randomization inference methods, I randomly assigned treatment status to 14 (the number of counties in Massachusetts) out of 527 counties that made up all the treatment and control counties and obtained estimates of the treatment effect using the same statistical models as Sommers, Long, and Baicker (2014). I did the random assignment of treatment status 1000 times. (I also used 2000 replications to assess whether the p-values obtained were sensitive to the number of replications. They were not.) After the 1000 randomized estimates were obtained, a p-value was calculated as the proportion of 1000 estimates that were larger in absolute value than the estimate reported in Sommers, Long, and Baicker (2014). As an alternative to randomization inference, I also used the method proposed by Conley and Taber (2011). Results were very similar using this method, which is not surprising because the two methods are asymptotically equivalent (Conley and Taber 2011). Given the similarity of estimates, I do not report the Conley and Taber (2011) estimates.

State-Level Analysis

Sommers, Long, and Baicker (2014) also conducted analyses using the county-level data aggregated to the state-year level as the unit of observation. To conduct the randomization inference analysis in this case, I assigned each of the 46 control states (determined by the 513 control counties) to be the treatment group and the remaining states (including Massachusetts) to be controls and estimated the regression models used by Sommers, Long, and Baicker (2014) model 47 times. I then calculated the p-value as the proportion of the 47 estimates (including Massachusetts) that were larger in absolute value than the estimate reported in Sommers, Long, and Baicker (2014).

Results

Table 1 presents the results. The first column describes the sample, for example, whether the unit of analysis is county (county-age-gender-race-year) or state. The second column shows the method of estimation: general linear model (GLM) in which the dependent variable is the mortality count, or ordinary least squares (OLS) in which the dependent variable is the mortality rate (multiplied by 100). Estimates from GLM models are interpreted as percentage changes in the number of deaths. The third column indicates whether the model included covariates or not; all models include year and state fixed effects and an interaction between treatment status and post reform period. Covariates are those from Sommers, Long, and Baicker (2014) and include age, race, and county-level variables. The fourth column reports the estimate published in Sommers, Long, and Baicker (2014), and the fifth column reports my replication of that estimate. As noted earlier, I was able to replicate most of the estimates almost exactly, but not all. In particular, for models that included separate estimates of the treatment effect by subgroup (e.g., age), I was not able to replicate exactly the Sommers, Long, and Baicker (2014) study, but my estimates are close to the published estimates. Columns six and seven report the published p-values and the p-values of the replication. Finally, the last column reports the p-values from the randomization inference method.

I begin the discussion with reference to the top panel of estimates in Table 1, which reports estimates for the full sample when the unit of analysis is county-age-gender-race-year. As a comparison of columns four and five, and columns six and seven, indicate, I was able to replicate the Sommers, Long, and Baicker (2014) study almost exactly. The published p-values for the estimates in the top panel indicate that health care reform in Massachusetts significantly decreased mortality with p-values ranging from 0.003 to 0.04. Note too, that estimates are nontrivial in size—health reform was associated with a 3% to 4% decrease in mortality, which given that only a small fraction of the population was affected (i.e., gained health insurance coverage), for example, less than 10%, imply very large effects (e.g., 30% to 40%) of treatment (gaining health insurance) on mortality.
The p-values from the randomization inference procedure suggest that the estimates obtained by Sommers, Long, and Baicker (2014) are not statistically significant. The p-values range from 0.277 to 0.420 and are far above common levels of statistical significance. Very similar results were obtained using the Conley and Taber (2011) method, which is not surprising given that it is asymptotically equivalent to randomization inference. (I applied the Conley and Taber [2011] method only to estimates obtained using OLS.) In contrast, the use of wild, cluster bootstrap method proposed by Cameron, Gelbach, and Miller (2008) yielded p-values (not reported) of the same order of magnitude as those reported in Sommers, Long, and Baicker (2014), which is also not surprising as this method does not account for the inconsistency of the difference-in-differences estimate.

The magnitudes of the p-values obtained by randomization inference also suggest that the Sommers, Long, and Baicker (2014) analysis lacked adequate statistical power to detect effect sizes that are plausible, as the estimates reported in the article are quite large as previously discussed. Also note that the p-values are approximately the same from models with and without covariates, which differs somewhat from the results reported in Ryan et al. (2014). In that article, randomization inference performed poorly when adjustment was made for covariates, but that was in a model in which treatment status was correlated with baseline covariates by definition.

The middle panel of Table 1 shows estimates and p-values for the analyses based on state-level aggregates. Again, I was able to replicate the estimates published in Sommers, Long, and Baicker (2014) almost exactly. Here too, the p-values based on randomization inference are several times larger than those reported in Sommers, Long, and Baicker (2014) and not close to indicating statistically significant estimates with a range from 0.404 to 0.723.

Finally, the bottom panel of Table 1 reports estimates from models that allow the treatment effect to differ by subgroup, for example, by age. Notably, and following the approach of Sommers, Long, and Baicker (2014), estimates in this panel are not from regression models using stratified samples, but from regression models that use the full sample, but allow the treatment effect to differ by subgroup. For these models, I was not able to replicate exactly the estimates in Sommers, Long, and Baicker (2014), although my estimates are quite close to those of Sommers, Long, and Baicker (2014). Similar to previous results, the p-values based on randomization inference are orders of magnitude greater than the standard errors from the replication (column 7) and the standard errors reported (column 6) in Sommers, Long, and Baicker (2014). The randomization inference p-values range from 0.427 to 0.783.

### Conclusion

In a recent article, Sommers, Long, and Baicker (2014) evaluated whether health reform in 2006 in Massachusetts, which resulted in a significant decrease in the proportion of persons without health insurance, affected mortality. Based on their analysis, Sommers, Long, and Baicker (2014) concluded that health reform did in fact decrease mortality. If true, this is a finding with substantial public policy significance because of the lack of evidence that health insurance improves health. Indeed, since the results from the seminal experimental analysis of the RAND Health Insurance Experiment, researchers have not produced much credible evidence that health insurance improves health despite the strong intuition underlying the hypothesis. Because of this dearth of evidence, and the importance of health care reform in Massachusetts vis-à-vis the Affordable Care Act, the Sommers, Long, and Baicker (2014) article garnered widespread media attention and has been used by senior government officials as evidence of a likely benefit of the ACA. (For example, see report of Jason Furman, head of the White House Council of Economic Advisers: https://www.whitehouse.gov/blog/2014/02/06/six-economic-benefits-affordable-care-act)

One potential problem with the Sommers, Long, and Baicker (2014) analysis was that inferences were based on an approach that was unlikely to be valid given the empirical setting. The natural experiment that was the basis of the analysis was characterized by only one treatment state. In these settings, common approaches to inference have been shown to over reject the null hypothesis (Conley and Taber 2011). I assessed the extent of the potential inference problem by calculating p-values for estimates in Sommers, Long, and Baicker (2014) using randomization inference methods. The p-values derived from this approach are

#### Table 1 Replication of Sommers, Long, and Baicker (2014) with randomization inference p-values

| Units of sample | Method | Covariates | Sommers, Long, and Baicker (2014) estimate | Replication estimate | Sommers, Long, and Baicker (2014) p-value | Replication p-value | Randomization inference p-value |
|----------------|--------|------------|------------------------------------------|---------------------|----------------------------------------|-------------------|--------------------------------|
| Counties       | GLM    | No         | −4.2                                     | −4.2                | 0.032                                  | 0.031             | 0.277                           |
|                | GLM    | Yes        | −2.9                                     | −2.9                | 0.003                                  | 0.003             | 0.420                           |
|                | OLS    | No         | −11.4                                    | −11.5               | 0.040                                  | 0.039             | 0.364                           |
|                | OLS    | Yes        | −6.4                                     | −6.5                | 0.006                                  | 0.005             | 0.344                           |
| States         | GLM    | No         | −4.1                                     | −4.1                | 0.025                                  | 0.025             | 0.404                           |
|                | GLM    | Yes        | −2.3                                     | −2.2                | 0.103                                  | 0.141             | 0.723                           |
| Counties       | GLM    | No         | −4.1                                     | −4.1                | 0.025                                  | 0.025             | 0.404                           |
|                | GLM    | Yes        | −2.3                                     | −2.2                | 0.103                                  | 0.141             | 0.723                           |
| Non-Latino     | GLM    | Yes        | −3.6                                     | −3.4                | 0.030                                  | 0.022             | 0.596                           |
| Latino or Non-White | GLM    | Yes        | −4.6                                     | −4.5                | <0.001                                 | <0.001            | 0.783                           |
| Age 20–34      | GLM    | Yes        | −3.0                                     | −3.8                | <0.001                                 | 0.010             | 0.540                           |
| Age 35–64      | GLM    | Yes        | −2.2                                     | −2.7                | 0.008                                  | 0.006             | 0.631                           |
| Low-income     | GLM    | Yes        | −1.8                                     | −1.7                | 0.120                                  | 0.283             | 0.777                           |
| High-income    | GLM    | Yes        | −2.3                                     | −2.2                | 0.103                                  | 0.141             | 0.723                           |

NOTES: GLM: general linear model; negative binomial with log link. OLS: ordinary least squares. Covariates include.

The p-values from the randomization inference procedure suggest that the estimates obtained by Sommers, Long, and Baicker (2014) are not statistically significant. The p-values range from 0.277 to 0.420 and are far above common levels of statistical significance. Very similar results were obtained using the Conley and Taber (2011) method, which is not surprising given that it is asymptotically equivalent to randomization inference. (I applied the Conley and Taber [2011] method only to estimates obtained using OLS.) In contrast, the use of wild, cluster bootstrap method proposed by Cameron, Gelbach, and Miller (2008) yielded p-values (not reported) of the same order of magnitude as those reported in Sommers, Long, and Baicker (2014), which is also not surprising as this method does not account for the inconsistency of the difference-in-differences estimate.

The magnitudes of the p-values obtained by randomization inference also suggest that the Sommers, Long, and Baicker (2014) analysis lacked adequate statistical power to detect effect sizes that are plausible, as the estimates reported in the article are quite large as previously discussed. Also note that the p-values are approximately the same from models with and without covariates, which differs somewhat from the results reported in Ryan et al. (2014). In that article, randomization inference performed poorly when adjustment was made for covariates, but that was in a model in which treatment status was correlated with baseline covariates by definition.

The middle panel of Table 1 shows estimates and p-values for the analyses based on state-level aggregates. Again, I was able to replicate the estimates published in Sommers, Long, and Baicker (2014) almost exactly. Here too, the p-values based on randomization inference are several times larger than those reported in Sommers, Long, and Baicker (2014) and not close to indicating statistically significant estimates with a range from 0.404 to 0.723.

Finally, the bottom panel of Table 1 reports estimates from models that allow the treatment effect to differ by subgroup, for example, by age. Notably, and following the approach of Sommers, Long, and Baicker (2014), estimates in this panel are not from regression models using stratified samples, but from regression models that use the full sample, but allow the treatment effect to differ by subgroup. For these models, I was not able to replicate exactly the estimates in Sommers, Long, and Baicker (2014), although my estimates are quite close to those of Sommers, Long, and Baicker (2014). Similar to previous results, the p-values based on randomization inference are orders of magnitude greater than the standard errors from the replication (column 7) and the standard errors reported (column 6) in Sommers, Long, and Baicker (2014). The randomization inference p-values range from 0.427 to 0.783.

### Conclusion

In a recent article, Sommers, Long, and Baicker (2014) evaluated whether health reform in 2006 in Massachusetts, which resulted in a significant decrease in the proportion of persons without health insurance, affected mortality. Based on their analysis, Sommers, Long, and Baicker (2014) concluded that health reform did in fact decrease mortality. If true, this is a finding with substantial public policy significance because of the lack of evidence that health insurance improves health. Indeed, since the results from the seminal experimental analysis of the RAND Health Insurance Experiment, researchers have not produced much credible evidence that health insurance improves health despite the strong intuition underlying the hypothesis. Because of this dearth of evidence, and the importance of health care reform in Massachusetts vis-à-vis the Affordable Care Act, the Sommers, Long, and Baicker (2014) article garnered widespread media attention and has been used by senior government officials as evidence of a likely benefit of the ACA. (For example, see report of Jason Furman, head of the White House Council of Economic Advisers: https://www.whitehouse.gov/blog/2014/02/06/six-economic-benefits-affordable-care-act)

One potential problem with the Sommers, Long, and Baicker (2014) analysis was that inferences were based on an approach that was unlikely to be valid given the empirical setting. The natural experiment that was the basis of the analysis was characterized by only one treatment state. In these settings, common approaches to inference have been shown to over reject the null hypothesis (Conley and Taber 2011). I assessed the extent of the potential inference problem by calculating p-values for estimates in Sommers, Long, and Baicker (2014) using randomization inference methods. The p-values derived from this approach are
very large and range from 0.0277 to 0.783. These \( p \)-values indicate that none of the estimates reported in Sommers, Long, and Baicker (2014) are statistically significant at commonly accepted levels of significance. Moreover, given the relatively large magnitudes of the estimates reported in Sommers, Long, and Baicker (2014), the large \( p \)-values associated with these estimates suggest that the Sommers, Long, and Baicker (2014) analysis lacked adequate statistical power to detect plausible effect sizes.

To investigate whether a similar problem affected previous studies of the effect of reform in Massachusetts on the proportion of the population that was uninsured, which used the same difference-in-difference approach, I also conducted analyses of this question and used randomization inference. To do so, I used data from the American Community Survey from 2005 to 2008 and conducted analyses at the state level treating Massachusetts as the treated state and all others as the control states. This approach is broadly consistent with previous studies of the question (Long 2008; Long, Stockley, and Yemane 2009). In this case, inferences reported in previous studies, for example, Long (2008) and Long, Stockley, and Yemane (2009), remain valid even when based on randomization inference. My estimate indicated that Massachusetts reform was associated with approximately a five percentage point (0.051 to be exact) decrease in the percentage uninsured, which is similar to the 5.8 to 6.6 percentage point decreases reported in Long (2008) and Long, Stockley, and Yemane (2009), respectively, and the \( p \)-value of my estimate was 0.0196.

To summarize, while the Sommers, Long, and Baicker (2014) analysis was careful and thorough, the authors failed to recognize a serious problem with respect to inference that arises in the context of the difference-in-differences approach used in their analysis. Correcting this oversight suggests that estimates of the treatment effect were not statistically significant and that their study had inadequate statistical power to detect plausibly sized effects. Most importantly, the question of whether Massachusetts health reform affected mortality remains unanswered. A similar reversal of inference was reported in Courtemanche and Zapata (2014). In that study, the authors examined whether Massachusetts health reform was associated with an improvement in self-rated, general health, and other self-reported measures of health. When standard methods of inference were used, estimates of the effect of reform in Massachusetts were statistically significant. However, when Courtemanche and Zapata (2014) conducted a randomization inference analysis, they reported \( p \)-values indicating that their estimates pertaining to general, self-assessed health were no longer statistically significant (\( p \)-value = 0.16). (Courtemanche and Zapata [2014] actually reported the one-tailed test of significance, which is incorrect. The two-tailed test is what I used above, and verified by Charles Courtemanche in personal correspondence.)

From a public policy perspective, results from previous studies as to whether Massachusetts health care reform improved health that used difference-in-difference methods, as virtually all have done, should be viewed with appropriate skepticism. It remains unknown whether this reform improved health and the use of results of these studies to predict the effects of the ACA on health is not well founded. The results of this study also illustrate the conceptual appeal of randomization inference and its ease of implementation. Given the importance of policy analysis and the potential pitfalls of quasi-experimental methods such as those outlined in Meyer (1995) and Ryan et al. (2014), more policy analysts should probably make use of randomization inference methods.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2010), "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program," Journal of the American Statistical Association, 105, 493–505. [1,2]

Bertrand, M., Duflo, E., and Mullainathan, S. (2004), "How Much Should We Trust Differences-In-Differences Estimates?" The Quarterly Journal of Economics, 119, 249–275. [2]

Bloom, N., Eifert, B., Mahajan, A., McKenzie, D., and Roberts, J. (2013), "Does Management Matter? Evidence from India," The Quarterly Journal of Economics, 128, 1–51. [1,2]

Braun, T.M., and Feng, Z. (2001), "Optimal Permutation Tests for the Analysis of Group Randomized Trials," Journal of the American Statistical Association, 96, 1424–1432. [2]

Cameron, A.C., Gelbach, J.B., and Miller, D.L. (2008), "Bootstrap-Based Improvements for Inference With Clustered Errors," The Review of Economics and Statistics, 90, 414–427. [2,4]

Cameron, A.C., and Miller, D.L. (in press), "A Practitioner’s Guide to Cluster-Robust Inference," Journal of Human Resources. [2]

Cohen, I., and Dupas, P. (2010), "Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment," The Quarterly Journal of Economics, 125, 1–45. [1,2]

Conley, T.G., and Taber, C.R. (2011), "Inferences with ‘Difference in Differences’ with a Small Number of Policy Changes," The Review of Economics and Statistics, 93, 113–125. [1,2,3,4]

Courtemanche, C.J., and Zapata, D. (2014), "Does Universal Coverage Improve Health? The Massachusetts Experience," Journal of Policy Analysis and Management, 33, 36–69. [1,2,5]

Donald, S.G., and Lang, K. (2007), "Inference With Difference-in-Differences and Other Panel Data," The Review of Economics and Statistics, 89, 221–233. [2]

Fisher, R.A. (1935), The Design of Experiments, Edinburgh, London: Oliver and Boyd. Available at http://catal og.hathitrust.org/Record/001306228 [1]

Gail, M.H., Mark, S.D., Carroll, R.J., Green, S.B., and Pee, D. (1996), "On Design Considerations and Randomization-Based Inference for Community Intervention Trials," Statistics in Medicine, 15, 1069–1092. [2]

Gail, M.H., Tan, W.Y., and Piantadosi, S. (1988), "Tests for No Treatment Effect in Randomized Clinical Trials," Biometrika, 75, 57–64. [2]

Good, P.I. (2000), Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses, New York: Springer. [1,2]

——— (2005), Permutation, Parametric, and Bootstrap Tests of Hypotheses (3rd ed.), New York: Springer. [1,2]

Ho, D.E., and Imai, K. (2006), "Randomization Inference With Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election," Journal of the American Statistical Association, 101, 888–900. [1,2]

Imbens, G.W., and Rosenbaum, P.R. (2005), "Robust, Accurate Confidence Intervals With a Weak Instrument: Quarter of Birth and Education," Journal of the Royal Statistical Society, Series A, 168, 109–126. [1,2]

Long, S.K. (2008), "On the Road to Universal Coverage: Impacts of Reform in Massachusetts at One Year," Health Affairs, 27, w270–w284. [2,5]

Long, S.K., Stockley, K., and Yemane, A. (2009), "Another Look at the Impact of Health Reform in Massachusetts: Evidence Using New Data and a Stronger Model," American Economic Review, Papers and Proceedings, 99, 508–511. [2,5]

MacKinnon, J.G., and Webb, M.D. (2014), "Wild Bootstrap Inference for Wildly Different Cluster Sizes," Working Paper 1314, Department of Economics, Queen’s University. Available at https://ideas.repec.org/p/qed/wpaper/1314.html [2]
Moulton, B.R. (1990), “An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Unit,” *The Review of Economics and Statistics*, 72, 334–338. [2]

Meyer, B. (1995), “Natural and Quasi-Experiments in Economics,” *Journal of Business and Economic Statistics*, 12, 151–161. [5]

Rosenbaum, P.R. (2002), “Covariance Adjustment in Randomized Experiments and Observational Studies,” *Statistical Science*, 17, 286–327. [1]

Ryan A. M., Burgess J. F., and Dimick J. B. (2015), “Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences,” *Health Services Research*, 50, 1211–35. [1,4,5]

Small, D.S, Ten Have, T.R., and Rosenbaum, P.R. (2008), “Randomization Inference in a Group–Randomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance, and Quantile Effects,” *Journal of the American Statistical Association*, 103, 271–279. [2]

Sommers, B.D., Long, S.K., and Baicker, K. (2014), “Changes in Mortality After Massachusetts Health Care Reform: A Quasi-Experimental Study,” *Annals of Internal Medicine*, 160, 585–593. [1,2,3,4,5]

Webb, M.D. (2014), “Reworking Wild Bootstrap Based Inference for Clustered Errors,” Working Paper 1315, Department of Economics, Queen’s University. Available at https://ideas.repec.org/p/qed/wpaper/1315.html [2]