Dear Dr. Merino:
Thank you and the editorial team for the giving us the opportunity to address the comments of your statistical consultant, and to the consultant for these thoughtful comments.

Below we provide our responses. The reviewer's comments are in bold, and the responses are in normal type. In the revised manuscript, text that was new in the last revision continues to be highlighted in yellow. The few changes in text for the current revision are highlighted in green.

Whilst the analysis in the manuscript is statistically sound, appropriately extensive and well reported, I have serious concerns about the design and interpretation of this study. There are two key issues which cannot be rectified that undermine the validity of the study:

1) The intervention and control regions are poor matches for each other. Table 2 reports the characteristics in the regions, which show large significant differences in race (12% black vs 28% black) and smaller differences in other key variables. Similarly Figure 1 shows that the outcome rates in the two regions differ significantly throughout the timeframe of the study.

Although the difference in difference analysis looks at changes rather than absolute values (and thus adjusts for differences in baseline outcomes), and is adjusted for characteristics, I have significant residual concerns that the regions being compared are so unalike that the differences are beyond correction in any statistical model.

We very much agree that using appropriate control states in our analyses is a key challenge and is absolutely essential to get right. Our primary analyses employ a difference-in-differences methodology, a very-well accepted and commonly used method for assessing large-scale health policy interventions in the medical and health policy literature. This methodology is robust to level differences between cases and controls in the outcome of interest (ACSC admission rates) provided that longitudinal trends in the outcome prior to the intervention are not different -- this is often referred to as the "parallel trends assumption" (1). That is, if ACSC admission rate trends were already diverging between MA and control states prior to the onset of health reform in MA, our analysis would be biased and a different comparison group should be sought. However, we have addressed this question in two different ways and found that pre-reform trends in the control states were statistically not different from those in MA, supporting the appropriateness of our use of these control states. Below, we provide more detail about our extensive efforts to examine this issue.
First, as presented in the revision last submitted, we conducted an interrupted time series analysis to compare trends in ACSC admission rates (slopes) between MA and control states during only the pre-reform period, and found no differences in trends for the overall composite ACSC rate, the acute composite or the chronic composite rates (all p-values not significant). In addition, we found no difference in trends in these outcomes by race and ethnicity. We did not present the details of this analysis in the last revision but now include them in the revised appendix of the paper (page 49 of the revised manuscript) and show them here below:

<table>
<thead>
<tr>
<th>Measure</th>
<th>Estimated Regression Coefficient from Interrupted Time Series Analysis Examining Pre-reform Trends (slope) in ACSC Admission Rates in MA and Control States</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall composite ACSC</td>
<td>0.018</td>
<td>0.77</td>
</tr>
<tr>
<td>Acute composite ACSC</td>
<td>0.039</td>
<td>0.09</td>
</tr>
<tr>
<td>Chronic ACSC</td>
<td>-0.02</td>
<td>0.68</td>
</tr>
<tr>
<td>Overall Composite ACSC Black-white difference</td>
<td>0.021</td>
<td>0.32</td>
</tr>
<tr>
<td>Overall Composite ACSC Hispanic-white difference</td>
<td>0.26</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Second, we conducted an additional analysis to confirm that trends in ACSC admission rates were not different between MA and controls in the pre-reform period. For this analysis, we divided the pre-reform period into 2 equal periods, “Pre1” and “Pre2”. We then used a difference-in-differences approach identical in form (Poisson regression) and in model covariates to the main analysis presented in the paper (overall composite ACSC rate) except that “Pre1” was treated as the “pre-intervention” period and “Pre2” was treated as the “post-intervention” period. The difference-in-differences estimate from this analysis reflects whether or not there was a different trend in MA and controls in the pre-reform period. The results of this analysis confirm that no such different trends existed (that is, the difference-in-differences estimates were non-significant). We have chosen not to include these analyses in the paper, given the extensive sensitivity analyses already presented, but we would be happy to include them if deemed desirable. We show them below:
<table>
<thead>
<tr>
<th>Outcome</th>
<th>Difference in Change, “Pre1” vs “Pre2” time periods in Massachusetts</th>
<th>95% CI or p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall composite ACSC</td>
<td>0.1%</td>
<td>-1.9% , 2.2%</td>
</tr>
<tr>
<td>Acute composite ACSC</td>
<td>2.9%</td>
<td>-0.5% , 6.3%</td>
</tr>
<tr>
<td>Chronic Composite</td>
<td>-1.2%</td>
<td>-3.8% , 1.4%</td>
</tr>
<tr>
<td>Black-White difference in overall composite ACSC</td>
<td>-2.3%</td>
<td>-8.2% , 3.9%</td>
</tr>
<tr>
<td>Hispanic-White difference in overall composite ACSC</td>
<td>-2.1%</td>
<td>-8.7% , 4.9%</td>
</tr>
</tbody>
</table>

The validity of the difference-in-differences approach does not depend on the similarity (same level) of characteristics between the intervention (MA) and control groups (control states), such as race or income in our analysis. Nonetheless, we share the concerns of the consultant statistician that there could be some way in which such differences, if large, could lead to a different result of the intervention in MA vs control states. The consultant statistician, initial reviewers, and the editorial board raised this issue and correctly pointed to sizeable baseline differences in income and race/ethnicity between controls and MA in Table 1 in the paper. For example, pre-reform mean incomes in MA and control states were $47,343 and 41,018, respectively. The pre-reform proportion of black patients in MA and controls was 11.5% and 27.8% and of Hispanics, 9.2% and 11.7%.

To address the potential impact of these large differences in our main analysis, we employed what has become the standard method of addressing baseline differences in population characteristics in difference-in-differences analyses -- propensity score matching -- to identify a control group that more closely resembles the intervention group with regard to (non-outcome) characteristics. As described in our previous revision (and retained in the current revision), we used logistic regression to calculate propensity scores (based on income and race/ethnicity) for county-level cohorts and then using these scores, identified an alternative control group using only cohorts in control states in the highest quartile of propensity scores (those most closely matching MA cohorts). We presented the results of a sensitivity analysis using this alternative cohort that included cohorts from 63 of the original 150 counties in control states. For the overall composite outcome, the point estimate was 0.0% with 95% CIs, -3.1% and 3.2%. This result did not differ from our main analysis, yet retained reasonably tight
confidence intervals. Estimates for black-white and Hispanic-white differences were similarly not significant. Although not presented in the previous revised paper, this matching resulted in control states that were more closely matched to MA: mean incomes for pre-reform MA and matched controls were about $47,000 and $46,000, respectively. The pre-reform proportion of black patients in MA and matched controls was 12.6% and 21.2% and of Hispanics, 8.6% and 13.7%. Using even more stringent criteria for propensity score matching (top 15th percentile of propensity scores) results in even closer matching of controls with similar ACSC results but with (necessarily) wider confidence intervals and thus we chose to retain the analysis using the top quartile of propensity scores in the revised manuscript. In addition, as pointed out by the statistical consultant, we controlled for multiple potentially important covariates (such as race and income) in all models.

As we stated in our initial response, there are no alternative control states that would be obviously superior to those that were selected for our study. States that are geographically closer (New England states) might have appeal but: 1. they have even smaller minority populations and thus would very substantially lower the statistical power to detect differences in the outcome by race (in an analysis which the reviewer points out already has lower than optimal power) and in some cases would have a greater difference from MA than control states we used with regard to race (black population of Maine, Vermont and New Hampshire are 1.4%, 1.2% and 1.5%, respectively); 2. Many lack state-level discharge data with accurate capture of race/ethnicity. This is in contrast to the control states used in our analysis which have both sizeable minority populations and highly accurate capture of race/ethnicity. In addition, it is hard to imagine other US states that are better matched to MA with regard to other potentially important factors raised by reviewers including geographic proximity to MA, similar medical cultures (percentage teaching and tertiary care hospitals and high proportion of specialists) and similar economies and urbanicity. Lastly, the states we used as controls are identical to those used as controls in prior studies of the impact of MA health reform, including two prior studies from our group that have been published or accepted for publication in BMJ.

Given the demographics and medical cultures and economies of US states, the reality is simply that there are no states that are plausibly closer matches to MA in these important respects. Given the importance of the research question, we believe the challenge is to use control states that, while imperfectly comparable with regard to level differences in characteristics such as race/ethnicity, are the best available and: 1. Demonstrate no violation of the parallel trends assumption and 2. Provide subsets of the population that can be identified as being similar to MA using propensity score matching for which difference-in-differences estimates are robust. The results indicate that our analyses meet these criteria.

2) The study has failed to show impact of the intervention. There is an important difference between claiming evidence of absence of an effect, and absence of evidence. Whilst the study authors have been very careful to use
appropriate words stating “no effects were observed”, I cannot see that the study would have been able to detect large effects even if they had existed.

For example, Table 1 indicated that only 8.8% of the population was uninsured at baseline which decreased to 4.9% after the intervention. Thus, the intervention would only have affected 4% of the population, which places limits on the possible associated change in the event rates. The confidence intervals for the width of the effects are 5.7 percentage points wide for the overall composite ACSC and 6.5 and 7.6 percentage points wide for the acute and chronic ACSCs (Table 3) – these are all wider than the expected effect. For the impact on ethnic disparities the confidence intervals are 20 percentage points wide overall, and 21 and 26 percentage points for the chronic and acute (Table 4) – far too wide to be able to detect any likely effect. Thus it does not appear that the study did not have the statistical power to detect even the largest perceived possible effects if they had existed.

The authors briefly acknowledge this issue in their discussion. Whilst they could include a post hoc power calculation to demonstrate the ability of their study to detect possible effects, I find it hard to believe that it would show that the study could ever have detected anything likely to have occurred.

We appreciate the concern raised that our study may not have adequate power to detect effect sizes as small as would be expected to be induced by the reform. Although greater power would be optimal, for several reasons that are not immediately obvious, we suggest that our analyses are able to exclude effect sizes that might have been expected after reform.

The first issue is the size of the expected magnitude of the change in the proportion of MA residents who gained new insurance under the reform. The statistical consultant points out that the proportion of people discharged from hospitals in MA declined from 8.8 to 4.9 -- about 4% -- and that, therefore, the intervention (gaining insurance) would only have affected 4% of the population. However, the 4% change described is the change in the percentage of MA residents with an ACSC admission, not the percent of MA residents who newly received insurance under the reform. The latter statistic is the one that is relevant here, however. Most studies place the percentage gaining insurance under the reform at 5%-7% but estimates have been as high as 8.2% from MA’s own state surveys (2) -- nearly twice the size of the change estimated from changes in the proportion of ACSCs accounted for by the uninsured (4%) from our Table 1 in the manuscript. Thus, rather than 4% of residents newly gaining insurance, it was likely 5-7% but as much as 8%. Furthermore, the proportions of racial and ethnic minority populations gaining insurance after reform were even larger.

The second issue is whether the post-reform absolute percentage change in ACSC hospitalization rate in MA is set by the absolute percentage of MA residents gaining insurance as a result of the reform—i.e., that a 4% gain in insurance would mean a
maximum expected change in ACSC rates of 4%. While at first this may appear to be the case, it is actually more complicated. The maximum percentage change in the outcome indeed depends on the percentage of the population gaining new insurance; but, crucially, it is also depends on what the difference in ACSC rates were (at baseline) between the insured and uninsured. We cannot calculate these rates directly with our data, but we can get some clue of potential differences from our Table 4, where we see that the baseline ACSC rate differences by race/ethnicity were enormous (annually per 100,000 population, whites=667; blacks=1713; Hispanic=1258). For illustrative purposes: let us assume that the uninsured were 10% of population before reform and dropped to 5% after reform. And let's assume ACSC admission rates were a constant 500 for insured and 2000 for uninsured. Then the average population ACSC rates would be 650 pre-reform and 575 post-reform, a decrease of 11.5%. All of this effect is simply from an increase of 5% in the newly insured; their ACSCs went from the rate for uninsured to the rate for insured. Thus, while we cannot (with our preferred model specification) exclude a significant change in ACSC rates of less than 5.7% (the span of the 95% CIs of our overall composite measure), we can exclude significant changes larger than this which would have been quite plausible, perhaps even expected, and that are of significant policy relevance. The principle illustrated here would, of course, apply equally to our analyses by race/ethnicity. Given the current length of the manuscript and the additional text that would be required to describe this point, we have not modified the manuscript regarding this point but would be happy to do so if deemed important by the editors.

In addition, it is important to point out that based on suggestions made by initial reviewers, we altered our preferred model specification between our initial submission to the initial revision, making the preferred model more conservative in its assumptions, and yielding wider 95%CIs. Specifically, in our initial submission, our models used a random effects (rather than fixed effects) model with no clustering of standard errors. Based in part on comments from and suggestions for additional analyses by reviewers, in our initial revision we adopted as our preferred model specification (in the first revision and retained in the current revision) one with county-level fixed effects and clustering-adjusted standard errors. For comparison, we have now re-run our models with random effects and calculation of standard errors using bootstrap methods and this produces the same point estimates but considerably narrower 95%CI spans. The point estimates and 95% CI for our current preferred model and one with random effects and non-clustered standard errors using bootstrap methods (we now include this in the appendix as a sensitivity analysis in the second revision) are shown below:
Model | Overall Composite ACSC Admissions D-in-D Estimate | Black-White difference for Overall Composite ACSC D-in-D Estimate | Hispanic-White Difference for Overall Composite ACSC D-in-D Estimate
--- | --- | --- | ---
Current Preferred: County fixed effects with clustering-adjusted standard errors, | 1.2 (-1.6, 4.1) | -1.9 (-8.5, 5.1) | 2.0 (-7.5, 12.4)
Alternate: Random effects with standard errors calculating using the bootstrap method. | 1.2 (-0.1, 2.5) | -1.9 (-5.7, 2.0) | 2.0 (-2.5, 6.6)

In fact, other model specifications that reviewers requested and the results of which we reported in the appendix of the initial revision demonstrate considerably narrower 95% CIs than our preferred model specification. Thus, among the multiple model specifications we examined, we selected as our preferred model specification, the one with the widest 95% CI spans—all others had rather tight CIs (we now make this point in the revised discussion (page 21 in the revised manuscript). Given the 95% CI of the preferred model and multiple models explored in sensitivity analyses and the likely size of the impact of the intervention, we believe that our analyses are able exclude effect sizes above levels that might have been expected after reform and that this is novel and important information in evaluating the impact of this landmark reform.

Summary

Racial and ethnic disparities in health outcomes are an enormous problem in the US. The landmark MA health care reform, the basis for the Affordable Act, was a major intervention to expand access to care, particularly for minorities. Despite some acknowledged modest potential limitations of our analysis—statistical power and comparability of controls—we believe that the sensitivity analyses show that these approaches are methodologically strong, that the totality of our many analyses are compelling, and that these data have important policy implications, particularly regarding the challenges of ameliorating the significant racial and ethnic disparities that plague the US health care system.

Thank you again for the opportunity to respond to these thoughtful comments on our manuscript.

Very truly yours,
Danny McCormick

References
