Online Appendix for "Messaging and the Mandate: The Impact of Consumer Experience on Health Insurance Enrollment Through Exchanges"

By Natalie Cox, Benjamin Handel, Jonathan Kolstad, and Neale Mahoney*

A. Summary Statistics

Table A1 shows demographic summary statistics across campaigns and treatments. We observe a total of 662,713 unique visitors to the site. We only observe demographic information for 310,210 of these observations, and these visitors are more likely to submit an application. The first column of Table A1 shows mean demographic characteristics for the control group across the three campaigns for the subsample the submits demographic information. The average visitor is slightly more than 40 years old, has 1.6 family members and is slightly more likely to be female than male. Figure A1 displays the distribution of users in the experiment by self-reported age and gender. Most users are in the range of 20-60 years old, and both genders seem well represented across age groups.

The remaining columns of Table A1 conduct randomization tests to confirm that validity of the experiment. In particular, the columns show the results of regressions of the different demographic variables on an indicator for whether the visitor was assigned Treatment B or Treatment C. The table shows no correlation between treatment assignment and the demographics, confirming the randomized nature of the experiment.

Table A2 looks at whether the demographic composition and behavior of visitors to the site changes as we approach the deadlines (using data pooled across treatments and campaigns). The table shows that the average visitor tends to be older and has a higher likelihood of submitting an application as we approach the enrollment deadline, although the effects are not particularly strong.

B. Treatment Effect Heterogeneity

We begin our analysis by simply comparing the outcome across treatment and control at the aggregate level. That is we pool the entire sample across demographic observables and days, and simply compare those in treatments with a

* Cox: Economics Department, UC Berkeley, address1, email1. Surname2: Economics Department, UC Berkeley and NBER, address2, email2. Kolstad: Wharton School, University of Pennsylvania and NBER, email3. Mahoney: University of Chicago Booth School of Business and NBER.

Table A1—Randomization Test

	Panel A: Campaign 1						
	Control	*Treatment B*			*Treatment C*		
	Mean	TE	SE	*p*	TE	SE	*p*
Age	40.21	-0.04	0.08	0.60	0.06	0.10	0.53
Family Size	1.66	0.00	0.01	0.83	-0.01	0.01	0.24
Female (Percent)	52.7	0.1	0.3	0.63	0.2	0.4	0.62
N=	125,994	127,543			67,121		
	Panel B: Campaign 2						
	Control	*Treatment B*			*Treatment C*		
	Mean	TE	SE	p	TE	SE	p
Age	40.88	-0.05	0.09	0.54	-0.09	0.09	0.33
Family Size	1.63	0.00	0.01	0.99	0.00	0.01	0.73
Female (Percent)	53.1	-0.3	0.3	0.32	-0.7	0.3	0.02
N=	113,036	111,085			111,357		
	Panel C: Campaign 3						
	Control	*Treatment B*			*Treatment C*		
Age	42.88	-0.62	0.32	0.05	-0.05	0.31	0.86
Family Size	1.61	-0.04	0.02	0.09	-0.03	0.02	0.13
Female (Percent)	54.1	0.3	1.1	0.83	-0.3	1.1	0.78
N=	6,732	6,655			7,425		

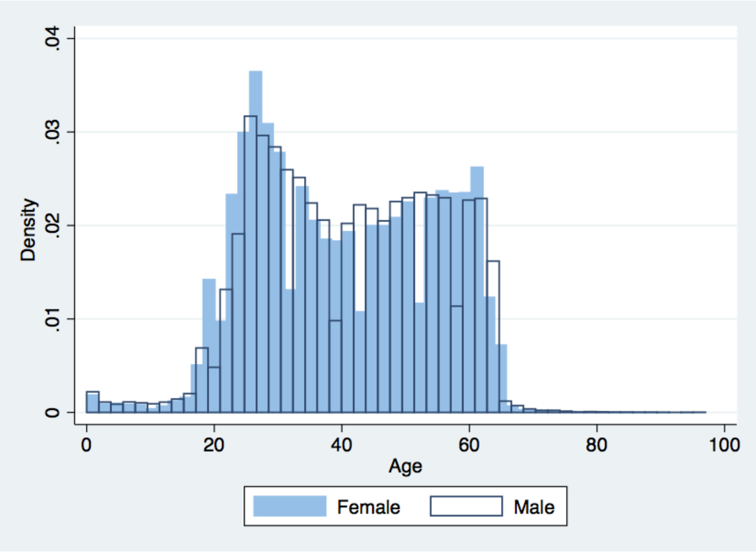


Figure A1. Distribution of Age, by Gender

Table A2—Characteristics of Visitors and Conversion Rate over Time

	Pooled Campaigns (N = 310,210)							
	Age		*Female*		*Family Size*		*Conversion Rate*	
Days Before Deadline	Mean	SE	Mean	SE	Mean	SE	Mean	SE
t-4	41.01	0.09	0.54	0.003	1.67	0.007	0.029	0.001
t-5	40.65	0.05	0.53	0.002	1.64	0.004	0.028	0.000
t-6	40.81	0.07	0.52	0.002	1.62	0.005	0.026	0.000
t-7	41.21	0.07	0.52	0.003	1.61	0.005	0.024	0.000
t-8	40.76	0.06	0.53	0.002	1.66	0.005	0.026	0.000
t-9	40.32	0.06	0.54	0.002	1.65	0.004	0.026	0.000
t-10	38.41	0.15	0.54	0.006	1.66	0.013	0.024	0.001

countdown and those in treatments with additional messaging. Table A3 presents the mean conversion rates for these scenarios. The control group had a mean conversion rate of the sample period of 2.63 percent. Treatment group B—customers who saw the countdown messaging with reference to the penalty of signing up beyond the deadline—had a mean conversion rate of 2.62 percent. Treatment C—customers who saw the countdown but with relatively more benign messaging—had a mean conversion rate of 2.70 percent. While the treatments do appear to have had a small impact, the differences are both small in magnitude and treatment B is statistically insignificant while treatment C is significant at the 10 percent level.

We next turn to the impact of messaging over time. While there are many reasons to consider how the types of customers shopping and the ways in which messaging might affect their behavior might change over time, we are particularly interested in the time dimension because (i) the treatment messaging emphasized time to the deadline using a countdown clock and (ii) given some of the difficulties with HealthCare.gov there were repeated changes to the actual deadline to get coverage in order to avoid paying the individual penalty. These features could certainly have led to many different responses depending on procrastination, beliefs about enforcement, etc. We do not explicitly model this but rather focus on whether there are differences in the impact of treatments early in the rollout from 10 to 7 days from the relevant deadline compared to the closer dates from 6 to 4 days prior to the deadline. The second panel of Table A3 shows the differential impact of treatments B and C in each of these periods. We first note that conversion rates were lower even in the control group in the early period, potentially reflecting some search behavior or procrastination. In the early period the deadline messaging appears, if anything, to have reduced the conversion rate, although the effects are small and not significantly different from zero. On the other hand, as the deadline drew near we see conversion rates rise in the control group but also a significant increase in the impact of treatment C on conversions.

Table A3—Treatment Effects for Targeted Groups

	No Targeting					
	Assign All to Control		B vs. Control		C vs. Control	
	Mean	SE	TE	SE	TE	SE
Conversion Rate (Percentage Points)	2.63	0.03	-0.01	0.06	0.06*	0.07
N	662,713					
	Time Targeting					
	Assign All to Control		B vs. Control		C vs. Control	
Time-Specific Conversion Rates (Percentage Points)	Mean	SE	TE	SE	TE	SE
Late Period Conversion Rate	2.70	0.04	0.06†	0.06	0.16 **†	0.06
Early Period Conversion Rate	2.57	0.03	-0.01	0.05	-0.06	0.06
N	662,713					
	Demographic Targeting					
	Assign All to Control		B vs. Control		C vs. Control	
Demographic Cell-Specific Conversion Rates (Percentage Points)	Mean	SE	TE	SE	TE	SE
Female, non-Single, >40	3.11	0.16	0.03	0.22	0.34	0.25
Female, non-Single, <40	3.40	0.12	-0.04	0.17	-0.04	0.18
Female, Single, >40	4.48	0.15	-0.03	0.21	0.11	0.24
Female, Single, <40	7.24	0.16	0.42 *	0.23	0.59 **	0.26
Male, non-Single, >40	3.26	0.20	-0.20	0.28	0.33†	0.32
Male, non-Single, <40	3.86	0.22	-0.17	0.28	-0.37	0.32
Male, Single, >40	4.17	0.10	-0.08	0.15	-0.05	0.16
Male, Single, <40	7.31	0.13	0.35 *	0.19	0.37 *	0.21
N	310,105					
	Time and Demographic Targeting					
	Assign All to Control		B vs. Control		C vs. Control	
Demographic/Time-Specific Conversion Rates (Percentage Points)	Mean	SE	TE	SE	TE	SE
Late, Female, non-Single, >40	3.14	0.24	0.08	0.34	0.51	0.34
Late, Female, non-Single, <40	3.47	0.12	-0.04	0.17	-0.04	0.18
Late, Female, Single, >40	4.42	0.22	0.03	0.30	0.36	0.33
Late, Female, Single, <40	7.46	0.25	0.52	0.36	1.01 **	0.38
Late, Male, non-Single, >40	3.41	0.30	-0.29	0.41	0.39	0.47
Late, Male, non-Single, <40	3.98	0.34	-0.04	0.44	-0.37	0.47
Late, Male, Single, >40	4.04	0.15	0.08	0.22	0.14	0.23
Late, Male, Single, <40	7.48	0.20	0.45	0.30	0.53 *	0.30
Early, Female, non-Single, >40	3.09	0.21	-0.02	0.28	0.11	0.36
Early, Female, non-Single, <40	3.34	0.12	-0.03	0.17	-0.07	0.18
Early, Female, Single, >40	4.53	0.20	-0.07	0.28	-0.16	0.35
Early, Female, Single, <40	7.08	0.21	0.33	0.30	-0.03	0.34
Early, Male, non-Single, >40	3.14	0.27	-0.14	0.37	0.17	0.44
Early, Male, non-Single, <40	3.77	0.28	-0.28	0.37	-0.43	0.43
Early, Male, Single, >40	4.27	0.15	-0.21	0.20	-0.22	0.23
Early, Male, Single, <40	7.18	0.18	0.28	0.25	0.09	0.30

Note: * denotes significance in difference between treatment and control conversion rates.
† denotes significance in difference between treatment B and treatment C conversion rates.

While still relatively small, 0.16 percentage points, the effect is positive and significant at the 5 percent level. These differences underscore the importance not only of the ability to try new customer experiences but to understand how they might impact different customers differently. In this case, the optimal strategy would appear to be avoiding messaging with a countdown until close to the actual deadline itself. We return to the optimal combination below.

In addition to time targeting, customers can be targeted based on potentially detailed demographic information. In the setting we study, this can occur if customers search for plan options and supply demographic detail. More generally, when customers arrive at web pages they bring cookies from prior sites visited and firms can potentially predict demographic details based on this information to target messaging (or target based on website history directly). In either case, we are interested in whether the impact of the different treatments varies along the dimension of easily observed demographics. To study this question we estimate the impact of each treatment within demographic groups defined by age (above or below the median age of 40), gender and single versus family.

The results, in the third panel of Table A3 demonstrate that there are some groups for which the impact of different messaging seems to be both economically meaningful and statistically significant. We find, for example, that both men and women under 40 who are single have a significant positive response to treatment B and C relative to control. The impact is 0.35 and 0.42 percentage points respectively or a 6 percent and 5 percent increase in conversion rates due to deadline messaging. The deadline messaging does not have a statistically significant impact among other demographic groups, families overall and younger single populations.

Finally, we study variation in the impact of messaging based on geography. In this case, we study the differential impact of the treatments across states. States are a natural segmentation unit in our setting for a number of reasons: insurance markets are regulated at the state level, plan offerings are generally made at the state level and political sentiment vis a vis health reform also varies across states (e.g. red versus blue states). In this study we do not explore these mechanisms explicitly. Instead, we estimate the variation across states by allowing a differential treatment effect for each message across each state, as described in our estimation approach. Figure A2 shows the distribution of these treatment effects across the various states in our sample, controlling for individual specific characteristics like age, gender, and family size. The upper figure is the distribution for the "pooled" treatment effect, which combines observations experiencing both treatments B and C, while the bottom histograms separate the two effects. While the majority of states have a near zero response to the deadline messaging, there are a number of states (approximately 20 percent) for whom the treatment increases or decreases the conversion rate by more than 1 percentage point (note that this is very large in comparison to a 0.3 to 0.4 percentage point treatment effect for specific demographic groups, as discussed above), with the effect being as large as 3 to 4 percentage points for some outlying states.

C. Bias Adjustment

Let i denote the day × demographic group × state × treatment cells and g denote the "groups" at which we allow for heterogeneous treatment effects. Using the experiment, we are able to recover *estimates* of the treatment effect $\hat{\beta}_g$ for each group g. We would like to recover the conversation rate from the "optimal" targeting scheme, which we define as the scheme that assigns treatment to a group if and only if it increases conversions relative to the control. This is given by $r^\star = \sum_g s_g \max\{\beta_g, 0\}$, where s_g is the fraction of visitors in demographic group g.

By the Central Limit Theorem, we know the treatment effect estimates are distributed $\sqrt{n_g}\left(\hat{\beta}_g - \beta_g\right) \sim N(0, \sigma_\beta^2)$ where n_g is the number of observations in group g and σ_β is the standard deviation of the estimate. It is convenient to rewrite the estimated treatment effect as the sum of the true effect and a normally distributed error term, $\hat{\beta}_g = \beta_g + \epsilon_{\beta_g}$, where $\epsilon_{\beta_g} \sim N(0, \frac{\sigma_\beta^2}{n_g})$.

Consider the adjusted estimator $\widehat{r^\star}$ that assigns the *estimated* treatment minus a bias-adjustment term to a group if and only if the *estimated* treatment effect is positive:

$$\widehat{r^\star} = \sum_g s_g 1(\widehat{\beta_g} > 0)\left\{\widehat{\beta_g} - E[\epsilon_{\beta_g} \mid \widehat{\beta_g} > 0]\right\}.$$

where $E[\epsilon_{\beta_g} \mid \widehat{\beta_g} > 0]$ is the term that adjusts the estimate. The expected value of this object is given by

$$\begin{aligned} E_{\epsilon_g}[\widehat{r^\star}] &= \sum_g s_g E_{\epsilon_g}\left[1(\widehat{\beta_g} > 0)\left\{\widehat{\beta_g} - E[\epsilon_{\beta_g} \mid \widehat{\beta_g} > 0]\right\}\right] \\ &= \sum_g s_g P(\widehat{\beta_g} > 0) E_{\epsilon_g}\left[1(\widehat{\beta_g} > 0)\left\{\widehat{\beta_g} - E[\epsilon_{\beta_g} \mid \widehat{\beta_g} > 0]\right\} \mid \widehat{\beta_g} > 0\right]. \end{aligned}$$

We can further simply the expectation term to be

$$\begin{aligned} &E_{\epsilon_g}\left[1(\widehat{\beta_g} > 0)\left\{\widehat{\beta_g} - E[\epsilon_{\beta_g} \mid \widehat{\beta_g} > 0]\right\} |\widehat{\beta_g} > 0\right] \\ &= E_{\epsilon_g}[\widehat{\beta_g}|\widehat{\beta_g} > 0] - E[\epsilon_{\beta_g} \mid \widehat{\beta_g} > 0] \\ &= E_{\epsilon_g}[\beta_g + \epsilon_{\beta_g}|\widehat{\beta_g} > 0] - E[\epsilon_{\beta_g} \mid \widehat{\beta_g} > 0] \\ &= \beta_g. \end{aligned}$$

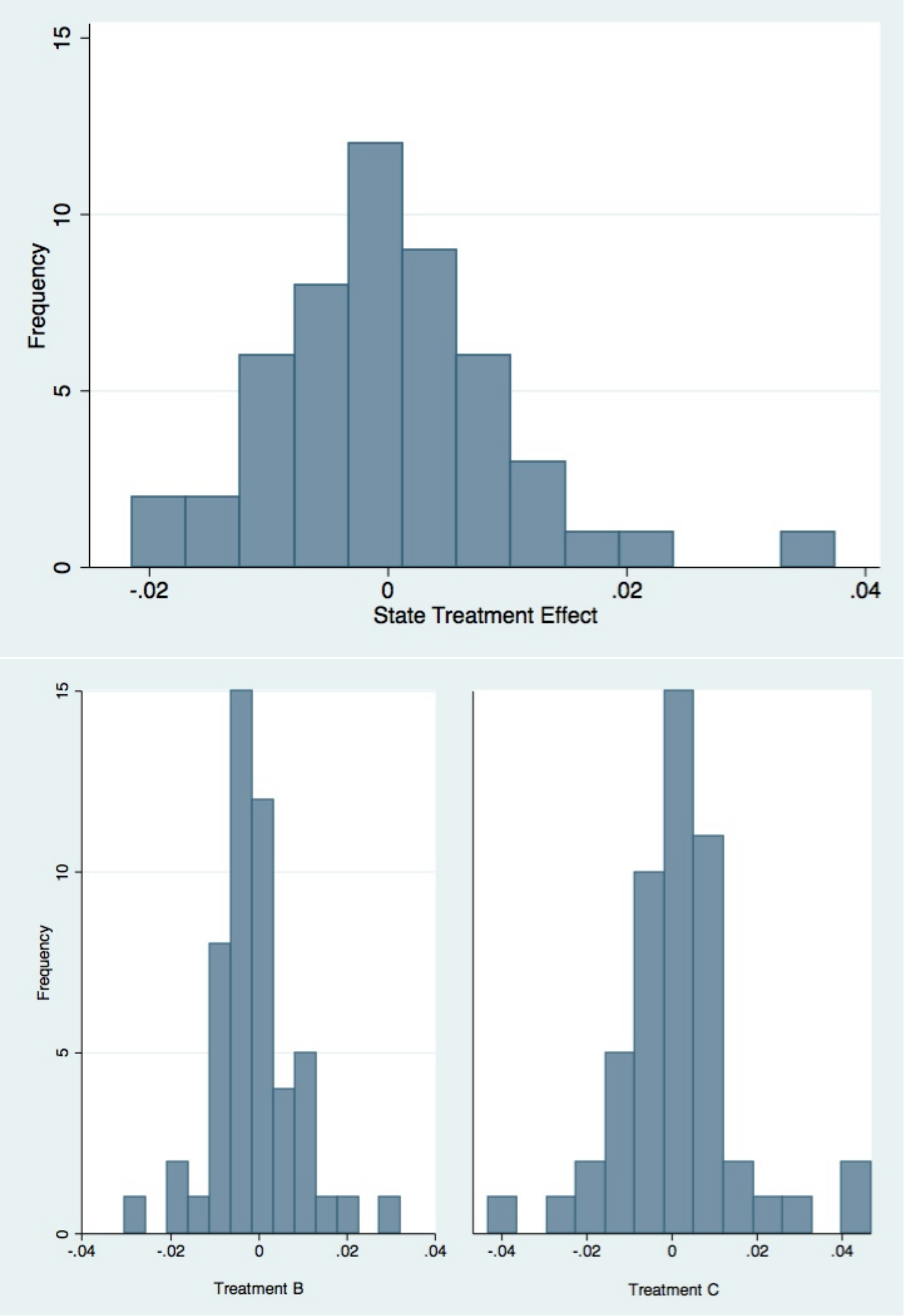


Figure A2. Distribution of State Treatment Effects, Single Pooled Treatment (top) and Separated Treatments (bottom)

The probability term can be simplified to be

$$P(\widehat{\beta_g} > 0) = P(\beta_g + \epsilon_{\beta_g} > 0) = \Phi(\frac{\beta_g}{\frac{\sigma_{\beta_g}}{\sqrt{n_g}}}).$$

Putting these pieces together gives us

$$E_{\epsilon_g}[\widehat{r^\star}] = \sum_g s_g \Phi(\frac{\beta_g}{\frac{\sigma_{\beta_g}}{\sqrt{n_g}}})\beta_g.$$

There are a few things to note about our estimator compared to the true effect of optimal targeting given by $\sum_g \max\{\beta_g, 0\} = \sum_g 1(\beta_g > 0)\beta_g$. First, our estimator is downward biased:

$$E_{\epsilon_g}[\widehat{r^\star}] = \sum_g s_g \Phi(\frac{\beta_g}{\frac{\sigma_{\beta_g}}{\sqrt{n_g}}})\beta_g < \sum_g 1(\beta_g > 0)\beta_g = r^\star$$

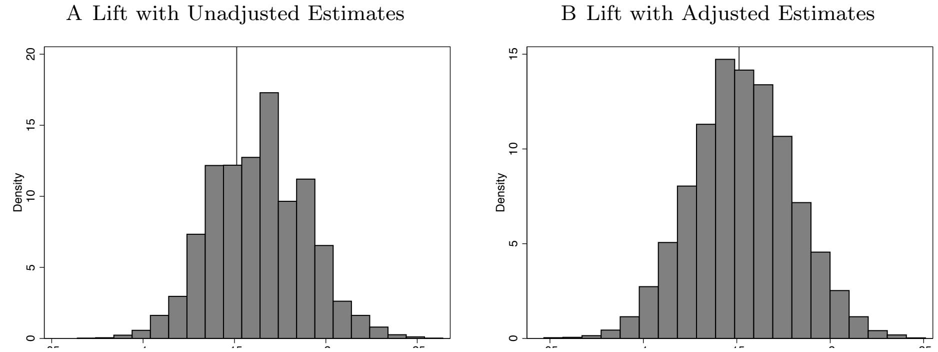
This can be observed by comparing $\Phi(\frac{\beta_g}{\sigma_{\beta_g}/\sqrt{n_g}})\beta_g$ to $1(\beta_g > 0)\beta_g$ over the negative and positive ranges of β_g. When β_g is negative, $\Phi(\frac{\beta_g}{\sigma_{\beta_g}/\sqrt{n_g}})\beta_g < 1(\beta_g > 0)\beta_g = 0$ and therefore the term on the left hand side of the inequality is smaller. When β_g is positive, $\Phi(\frac{\beta_g}{\sigma_{\beta_g}/\sqrt{n_g}})\beta_g < 1(\beta_g > 0)\beta_g = \beta_g$ and the term on the left had side is smaller as well.

Second, the estimator is consistent in the sense that this bias goes to zero for large n_g. This follows from the fact that $\Phi(\frac{\beta_g}{\sigma_{\beta_g}/\sqrt{n_g}})$ converges to $1(\beta_g > 0)$ for large enough n_g.

Third, at least in our setting, we have conducted Monte Carlo simulations that suggest that bias is small. In particular, we've conducted a Monte Carlo with the following parameters:

- Number of demographic groups: 10
- Observations per demographic group: 50
- Conversion rate: Control is 50 percent: Group-specific treatment effect is draw from uniform distribution with 0-100 percent support. This means that targeting raises the conversion rate by approximately 12.5 percentage points.
- Number of Monte Carlos runs: 10,000

Panel A of Figure A3 shows a histogram of the estimated “lift” or increase in conversion rate when the estimates are unadjusted for this source of bias. Panel B shows the lift with the adjusted estimates. The true estimate is 15.12 percent and is indicated by the vertical line in the figures. The average unadjusted estimate is 16.31 percent and the average adjusted is 15.30 percent.


A Lift with Unadjusted Estimates

B Lift with Adjusted Estimates

Figure A3. Monte Carlo Simulations

Adjusting for this source of bias when there are two treatment arms is conceptually similar. Let $\widehat{\beta_g}$ be the estimated effect of Treatment B and $\widehat{\gamma_g}$ be the estimated effect of Treatment C. With two treatments, the adjusted conversion rate from optimal targeting is given by

$$\widehat{r^\star} = \sum_g s_g \Bigg[1(\widehat{\beta_g} > \max\{\widehat{\gamma_g}, 0\}) \left\{ \widehat{\beta_g} - E[\epsilon_{\beta_g} \mid \widehat{\beta_g} > \max\{\widehat{\gamma_g}, 0\}] \right\} + \\ 1(\widehat{\gamma_g} > \max\{\widehat{\beta_g}, 0\}) \left\{ \widehat{\gamma_g} - E[\epsilon_{\gamma_g} \mid \widehat{\gamma_g} > \max\{\widehat{\beta_g}, 0\}] \right\} \Bigg].$$

We have not been able to derive analytical expressions for the conditional expectation terms $E[\epsilon_{\beta_g} \mid \widehat{\beta_g} > \max\{\widehat{\gamma_g}, 0\}]$ and $E[\epsilon_{\gamma_g} \mid \widehat{\gamma_g} > \max\{\widehat{\beta_g}, 0\}]$. We therefore estimate these terms by numerically integrating over the distributions of β_g and γ_g, as described below.

D. Bootstrap Standard Errors

To assess the impact of each treatment and targeting rule we estimate bootstrapped standard errors. While there are a number of general advantages, we are particularly interested in estimating bootstrapped standard errors in our setting given the potential for spurious findings when we study heterogeneity in the treatment effect. To do so we first divide our population of the control and treatment group into cells (with "cells" in our case referring to the demographic, time, or geographic-specific groups that we are interested in analyzing and targeting). We then apply a block bootstrap approach at the cell level and draw entire clusters with replacement 500 times. We then estimate the maximum treatment effect for each cell in each sample, and adjust this estimate given the level of noise in the sample and heterogeneity in the model of interest using numerical simulation. We repeat this process for all 500 bootstrap samples, forming a distribution of estimated treatment effects which yields a 95 % confidence interval.

Specifically, the steps of the bootstrap procedure to construct a confidence interval for our adjusted optimal targeting treatment effect, are:

1) Draw a sample from our population, with replacement, stratifying at the cell-level of targeting, g.

2) Estimate the maximum treatment effect for each cell (aka each targeted group), $\hat{\beta}_g$ within the sample.

3) For each cell group and treatment effect estimate, simulate 300 draws from a distribution $\sim N(\hat{\beta}_g, \sigma^2_{\hat{\beta}})$. Use these draws to numerically calculate $\mathbb{P}(\hat{\beta}_g > 0)$ and $\mathbb{E}\left[\hat{\beta}_g \mid \hat{\beta}_g > 0\right]$.

4) Using the probability and expected value obtained from step 3, calculate the adjusted treatment effect $\tilde{\beta}_g = \mathbb{P}(\hat{\beta}_g > 0)\mathbb{E}\left[\hat{\beta}_g \mid \hat{\beta}_g > 0\right]$.

5) Take the weighted average of $\hat{\beta}_g$ over all targeted groups to recover overall average treatment effect: $\hat{r}^* = \sum_g s_g 1(\widehat{\beta_g} > 0)\tilde{\beta}_g$

This process is repeated 500 times to create a distribution of $\hat{r}^*$, which is then used to calculate a 95 percent CI.

When we allow for heterogeneity this procedure accounts for potentially spurious findings in small cells flexibly. Suppose, for example, that we were spuriously finding large effects in only a small number of cells and these biased our main findings upwards. When we estimate bootstrapped standard errors those cells will be excluded for a number of runs and, therefore, we could find lower bound estimates without those cells that would be close to the control group.