Appendix 1. Methods for Patient-Level Characteristics

Patient demographic and clinical covariates were measured and discretized as follows: age in years, body mass index in kg/m^2, self-identified race/ethnicity, previous births, abortions, miscarriages and Cesarean sections, and gestational age on the day of mifepristone administration. Gestational age was further discretized into categories: ≤49 days (the baseline or reference level), 50 - 56 days inclusive, and 57 - 63 days inclusive. Categories of race/ethnicity were white, black, Asian, other, and not reported. Previous births, abortions, miscarriages, and Cesarean sections were categorized as binary variables (0 vs. 1 or more). Body mass index (BMI) was discretized into categories: underweight (≤ 18.5 kg/m2), normal weight (18.5 - 24.9 kg/m2), overweight (25.0 - 29.9 kg/m2), and obese (≥ 30.0 kg/m2). Age was treated as a continuous variable. Patient- level information is shown in Table 1 of the main paper.

Lohr PA, Starling JE, Scott JG, Aiken ARA. Simultaneous compared with interval medical abortion regimens where home use is restricted. Obstet Gynecol 2018; 131.
The authors provided this information as a supplement to their article.

Appendix 2. Methods for Unadjusted Primary-Outcome Analysis

To compare the effectiveness of the simultaneous versus interval regimen in the overall sample, as well as for each gestational age group, we proceeded as follows. We calculated unadjusted success rates for both the simultaneous and interval groups, yielding point estimates and 95% confidence intervals for unadjusted success rates and the unadjusted relative risk. We computed two-sided p-values against the null hypothesis that the log of the relative risk equals zero (equivalently, that RR = 1), under the assumption that the estimator of log relative risk has an asymptotic normal distribution around its true value. These confidence intervals and p-values are shown in the top half of Table 2 in the main paper.

Appendix 3. Methods for Adjusted Primary-Outcome Analysis

To account for the fact that patients self-assigned to a medical abortion regimen, we used logistic regression with propensity-score adjustment to estimate a relative risk of success. This is defined as the population-average ratio of two success probabilities:

$$RR = \frac{P(successful\ abortion\ under\ simultaneous\ regimen)}{P(successful\ abortion\ under\ interval\ regimen)}$$

This involved three steps:

1. Estimating propensity scores.
2. Using the propensity scores in a logistic regression model to estimate the relationship between abortion regimen and the probability of successful abortion, adjusting for patient self-assignment to regimen.
3. Post-processing the output of the logistic regression model to produce an estimate and a confidence interval for the relative risk of success—that is, the population average ratio of success probabilities given in the above equation.

We will describe each of these steps in turn.

Estimation of propensity scores. We begin by calculating a propensity score for each woman, estimating her probability of self-assignment to the simultaneous group (Appendices 10, 11), given her individual characteristics. We compare the clinical and demographic characteristics of women in simultaneous and interval groups and identify covariates that are significantly different between the subjects selecting the simultaneous versus the interval treatment (Table 1). These covariates were used as input features to predict propensity scores. Propensity scores are often estimated using a logistic-regression model. However, this involves strong parametric assumptions about the relationship between subject-level features and the probability of self-assignment to the simultaneous regimen, and incorrect model specification can produced biased estimates of average treatment effects. We therefore construct a random forest model instead of a logistic regression model (1, 2). Using random forests to estimate propensity scores avoids strong parametric assumptions and can account for complex interactions and

nonlinear effects among the covariates. Having fit the model, we then discretized the propensity scores into quintiles for subsequent use in our logistic-regression model for treatment success.

Logistic regression modeling. The binary response variable in our logistic regression was an indicator for early medical abortion success (0 = failure, 1 = success). The predictor variables were an indicator z for abortion regimen (0 = interval, 1 = simultaneous); a set of dummy variables encoding gestational-age category; and a set of dummy variables encoding the patient's propensity-score quintile, as estimated by our random- forest model. The interval group includes all EMAs with a 24- and 48-hour interval between receipt of mifepristone and receipt of misoprostol. Occasionally 72-hour intervals occur, but none were present in the dataset.

We fit both an overall regression model, as well as three separate regression models for each gestational age category (≤49 days, 50-56 days inclusive, 57-63 days inclusive.) In each logistic regression, we perform covariate adjustment using quintiles of the propensity score as a categorical predictor in each model (3, 4, 5, 6, 7). Propensity score quintiles were used in order to allow for the possibility of a nonlinear relationship between the propensity score and the probability of successful abortion. In the over-all model, we include main effects for treatment, gestational age, and propensity score quintiles as covariates. In the models for each gestational age group, we include main effects for treatment and propensity score quintiles. Thus our overall model takes the form:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 z_i + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 g_{1i} + \beta_7 g_{2i}$$

where p_i represents the probability of successful EMA for the ith subject, z_i is an indicator for the subject self-selecting to the simultaneous treatment (0 = interval, 1 = simultaneous), x_{ki} represents the an indicator for the ith observation's propensity score belonging to the kth propensity score quintile for $k \in \{2, \dots, 5\}$ and g_{ji} represents the jth gestational age level $j \in \{1,2\}$ for observation i. The first propensity score quintile and gestational age ≤ 49 are reference levels. Gestational age is included as a covariate only in the overall model. Therefore the terms involving β_6 and β_7 are absent in the models for the three gestational age stratified models, since these models are only fit to subsets of patients within a specific gestational-age category.

Post-processing the logistic regression model to obtain a relative risk estimate. We use the method described in the papers by Greenland (8, 9) to obtain relative risk estimates from our logistic regression models. We briefly describe this process here.

For each of the four fitted logistic regression models (one overall, three specific to a gestational age group), we can obtain fitted probabilities $p_1(x_i)$ and $p_0(x_i)$. These are the predicted success probabilities arising from the logistic regression model, assuming that participant i selected to the simultaneous or interval regimen, respectively. Only the treatment indicator z_i differs in the two fitted probabilities; all other covariates are identical and equal to their observed values (9). Explicitly, for the overall model we have:

$$p_0(x_i) = \frac{\exp(\beta_0 + \beta_1 \cdot 0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 g_{1i} + \beta_7 g_{2i})}{1 + \exp(\beta_0 + \beta_1 \cdot 0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 g_{1i} + \beta_7 g_{2i})}$$

with $z_i = 0$. This represents the fitted model's answer to the question: what would have been the predicted success probability for subject i, assuming that she had self-assigned to the interval regimen? Similarly

$$p_1(x_i) = \frac{\exp(\beta_0 + \beta_1 \cdot 1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 g_{1i} + \beta_7 g_{2i})}{1 + \exp(\beta_0 + \beta_1 \cdot 1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 g_{1i} + \beta_7 g_{2i})}$$

with $z_i = 1$. This represents the fitted model's answer to the question: what would have been the predicted success probability for subject i, assuming that she had self-assigned to the simultaneous regimen? We also have similar expressions for the models specific to each gestational age group, without the terms involving g_{1i} and g_{2i}.

We can use these fitted probabilities to obtain the estimated relative risk $r(x_i)$ for each subject in the data set, for each model, as the ratio of the two predicted success probabilities:

$$r(x_i) = \frac{p_1(x_i)}{p_0(x_i)}$$

Similarly, we can also estimate a population-average relative risk for the simultaneous versus interval group from each model by averaging the subject-level estimated risks (8,9). For model $k \in \{0,1,\ldots,3\}$, with n_k observations (where $k = 0$ refers to the overall model and $k = 1,2,3$ to the gestational-age-specific models), the standardized risk under the simultaneous regimen is

$$R_{k1} = \frac{1}{n_k}\sum_{i=1}^{n_k} p_1(x_i) \qquad (1)$$

Similarly, the standardized risk under the interval regimen is

$$R_{k0} = \frac{1}{n_k}\sum_{i=1}^{n_k} p_0(x_i) \qquad (2)$$

These quantities are shown as adjusted risks in Table 2 in the main paper. The population-average relative risk estimate for model k is now the ratio of these two standardized risks:

$$RR_k = \frac{R_{k1}}{R_{k0}} \qquad (3)$$

These quantities are also shown as adjusted relative risks in Table 2 in the main paper.

We calculate the adjusted number needed to treat (NNT) for each model as below, representing the number of patients who must select the interval regimen before there is one more successful EMA than if patients were treated with the simultaneous regimen.

$$NNT_k = \frac{1}{(R_{k0} - R_{k1})} \qquad (4)$$

Greenland provides standard error formulas for the adjusted relative risk based on the delta method (8, 9), and use these to obtain standard errors for the log of the relative risk, $\log(RR_k)$, and for NNT_k, for

each of the models. Included in Table 2 are p-values for testing the null hypothesis $H_0: \log(RR_K) = 0$.

Using point estimates and standard errors of the log relative risk, we also perform the Tukey Range Test for multiple comparison of means. Specifically, define

$$q_{k,l} = \frac{\log RR_k - \log RR_l}{se(\log RR_k - \log RR_l)}$$

The quantity $q_{k,l}$ is the standardized difference in log relative risk for gestational-age group k versus group l. In the denominator, se represents the standard error of the numerator, obtained through the Pythagorean identify for the standard error of the sum of two independent random variables:

$$se(\log RR_k - \log RR_l) = \sqrt{se(\log RR_k)^{\,2} + se(\log RR_l)^{\,2}}$$

Note that independence follows from the fact that the gestational-age specific relative risks RR_k are obtained from separate, non-overlapping subsets of the overall sample.

Now define

$$Q = max_{k,l \in \{1,2,3\}}(q_{k,l})$$

as the maximum standardized difference in the log relative risks. Under the null hypothesis that the relative risk of success for the simultaneous regimen is the same across gestational-age group ($k = 1, 2, 3$), the statistic Q should asymptotically behave the difference between the maximum and minimum of three draws from a standard normal distribution. It should therefore follow Tukey's studentized range distribution with 3 groups and degrees of freedom (df) equal to ∞. However, we actually use a Monte Carlo simulation to form the sampling distribution of Q under the null hypothesis that relative risk is constant across gestational age groups. This allows us to avoid relying on an asymptotic approximation. This range test is how we obtain the p-value for constant relative risk across gestational age groups reported in the Results section of the main paper.

Appendix 4. Results and Interpretation

Propensity score estimation via our random forest model yields propensity scores which have fairly similar distributions between the simultaneous and interval groups (Appendices 10, 11). The density plots of the two treatment groups' propensity scores explain why we do not see large changes between the unadjusted results and the adjusted results in Table 2 of the main paper.

Model fits for the four logistic regressions including propensity score quintiles are shown in the tables below (Appendices 5, 6, 7, 8). The figures below (Appendices 12, 13, 14, 15) give confidence intervals for the logistic regression covariates. All of the logistic regressions show a significant negative estimate for the coefficient on the simultaneous-treatment indicator variable (z_i). The general lack of significance of the regression coefficients for the propensity score quintiles are consistent with our finding that the propensity score densities of the simultaneous and interval regimens are similar.

In conducting the Tukey range test to assess whether the three population-average relative risk estimates for the gestational age levels are different, we observed the following standardized differences: 1.763 for comparing the first and second gestational age groups, 1.008 for comparing the first and third, and 0.23 for comparing the second and third. The p-value for the Tukey range test is 0.264, so we fail to reject the null hypothesis that all three relative risk estimates are equal (Appendix 9). Note that this p-value is derived from the sampling distribution of the standardized range statistic Q, as defined in Equation 5, that we obtained by Monte Carlo simulation. If we were instead to use the asymptotic approximation to the distribution of Q based on normal theory, the p-value would be slightly smaller, at p = 0.182.

We provide these adjusted RR and NNT estimates and confidence intervals in Table 2. Adjusting for self-assignment results in a small increase in the NNT point estimates, with slightly wider confidence intervals.

We found that although the simultaneous regimen was slightly less effective than the interval (with overall 97.3% relative effectiveness), the relative effectiveness did not decline significantly as gestational age progressed, indicating lack of a gestational-age interaction with the treatment variable. Clinicians should feel comfortable that simultaneous administration may be prescribed at any of the gestational

age groups without any greater loss of effectiveness with simultaneous than with interval administration, as gestational age increases.

As with any informed consent discussion, accurate information should be provided about the risks and benefits of the medical abortion regimens available, and women should be advised on the pertinent signs and symptoms that indicate potential failure and which should prompt contact with the provider.

Appendix 5. Factors Associated With Successful Early Medical Abortion With Simultaneous or Interval Administration of Mifepristone and Misoprostol at British Pregnancy Advisory Service From May 1, 2015, to April 30, 2016 (n = 28,901)

Coefficient	Estimate	95% Confidence Interval	p-value	Significance
(Intercept)	3.78	(3.584, 3.992)	<0.001	***
Simultaneous Regimen	-0.69	(-0.887, -0.449)	<0.001	***
GA 50-56 days	-0.41	(-0.538, -0.286)	<0.001	***
GA 57-63 days	-0.75	(-0.881, -0.622)	<0.001	***
PS quintile 2	0.28	(0.096, 0.458)	0.003	**
PS quintile 3	0.16	(-0.016, 0.338)	0.075	+
PS quintile 4	-0.05	(-0.224, 0.116)	0.537	
PS quintile 5	-0.17	(-0.337, -0.002)	0.048	*

Interval regimen, gestational age (GA) ≤ 49 days, and propensity-score (PS) quintile 1 are the reference categories.

Appendix 6. Factors Associated With Successful Early Medical Abortion With Simultaneous or Interval Administration of Mifepristone and Misoprostol at British Pregnancy Advisory Service from May 1, 2015, to April 30, 2016, for Gestational Age Until 49 Days (n = 16,021)

Coefficient	Estimate	95% Confidence Interval	p-value	Significance
(Intercept)	3.83	(3.512, 4.178)	<0.001	***
Simultaneous Regimen	-0.77	(-1.12, -0.441)	<0.001	***
PS quintile 2	0.35	(0.068, 0.634)	0.016	*
PS quintile 3	0.17	(-0.092, 0.441)	0.199	
PS quintile 4	-0.00	(-0.259, 0.25)	0.977	
PS quintile 5	-0.16	(-0.409, 0.086)	0.206	

Appendix 7. Factors Associated With Successful Early Medical Abortion With Simultaneous or Interval Administration of Mifepristone and Misoprostol at British Pregnancy Advisory Service from May 1, 2015, to April 30, 2016, for Gestational Age 50–56 Days (n = 7,637)

Coefficient	Estimate	95% Confidence Interval	p-value	Significance
(Intercept)	3.49	(3.152, 3.869)	<0.001	***
Simultaneous Regimen	-0.82	(-1.196, -0.475)	<0.001	***
PS quintile 2	0.29	(-0.04, 0.619)	0.086	+
PS quintile 3	0.09	(-0.23, 0.41)	0.580	
PS quintile 4	-0.01	(-0.33, 0.305)	0.943	
PS quintile 5	-0.17	(-0.478, 0.137)	0.283	

Appendix 8. Factors Associated With Successful Early Medical Abortion With Simultaneous or Interval Administration of Mifepristone and Misoprostol at British Pregnancy Advisory Service from May 1, 2015, to April 30, 2016, for Gestational Age 57-63 Days (n = 5,243)

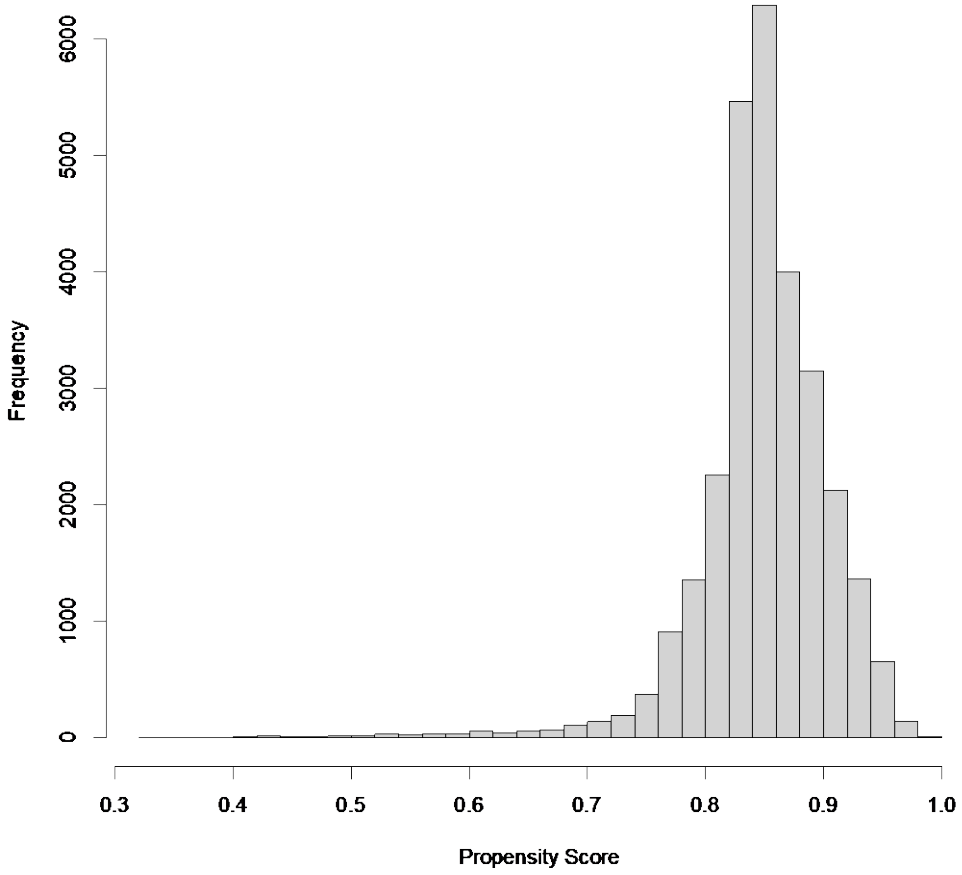
Coefficient	Estimate	95% Confidence Interval	p-value	Significance
(Intercept)	2.91	(2.596, 3.239)	<0.001	***
Simultaneous Regimen	-0.50	(-0.829, -0.191)	0.002	**
PS quintile 2	0.16	(-0.178, 0.503)	0.347	
PS quintile 3	0.23	(-0.124, 0.586)	0.203	
PS quintile 4	-0.17	(-0.506, 0.152)	0.297	
PS quintile 5	-0.18	(-0.517, 0.155)	0.297	

Appendix 9. Detailed Results for Tukey's Range Test to Compare Population Average Adjusted Log Relative Risk Across Gestational Age Groups, Rounded to the Second Decimal Place

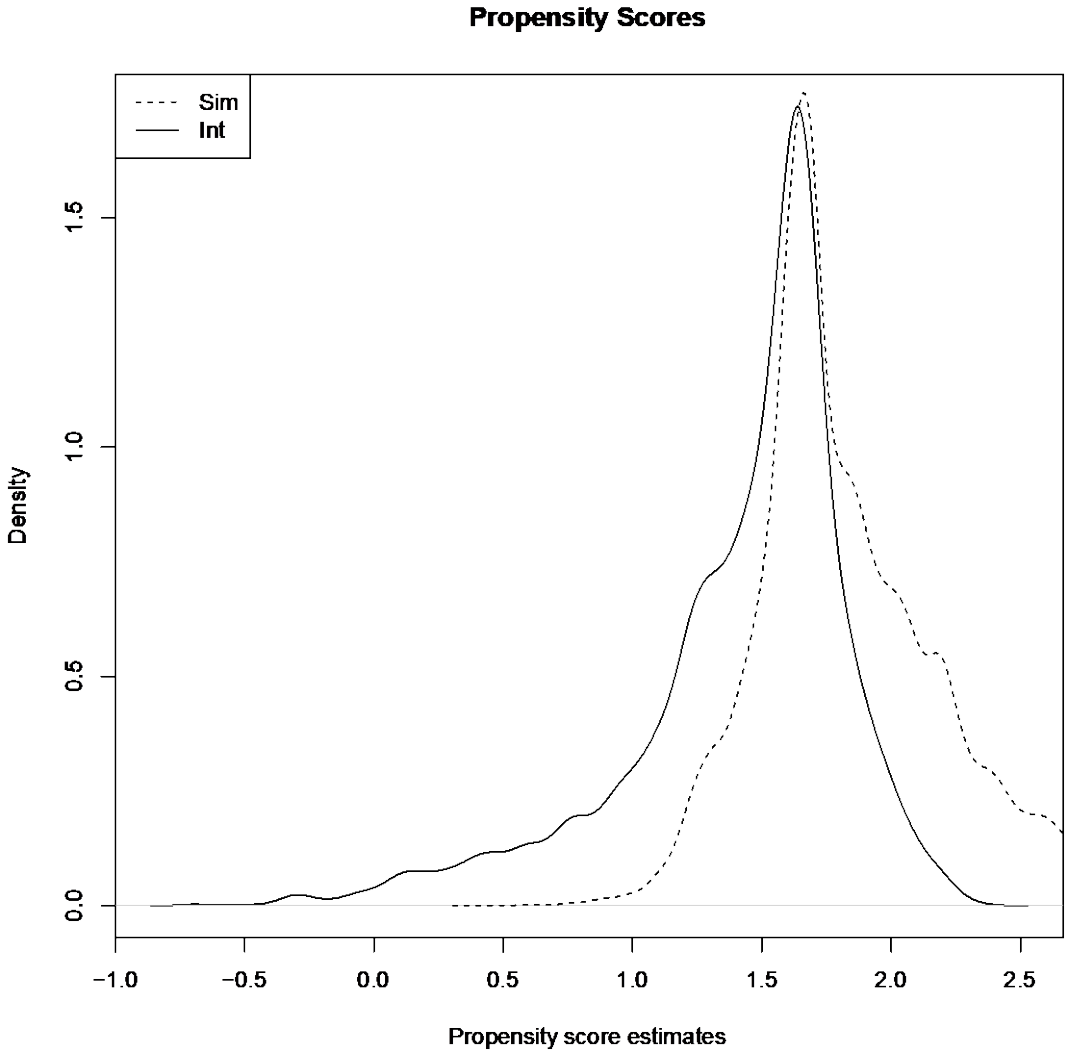
Comparison	$\Delta \log RR$	$SE(\Delta \log RR)$	Std Mean Diffs
GA ≤49 days versus GA 50-56 days	0.01	0.01	1.76
GA ≤49 days versus GA 57-63 days	0.01	0.01	1.01
GA 50-56 days versus GA 57-63 days	0.00	0.01	0.23

$\Delta \log RR$ represents the difference in log relative risk between the two categories.

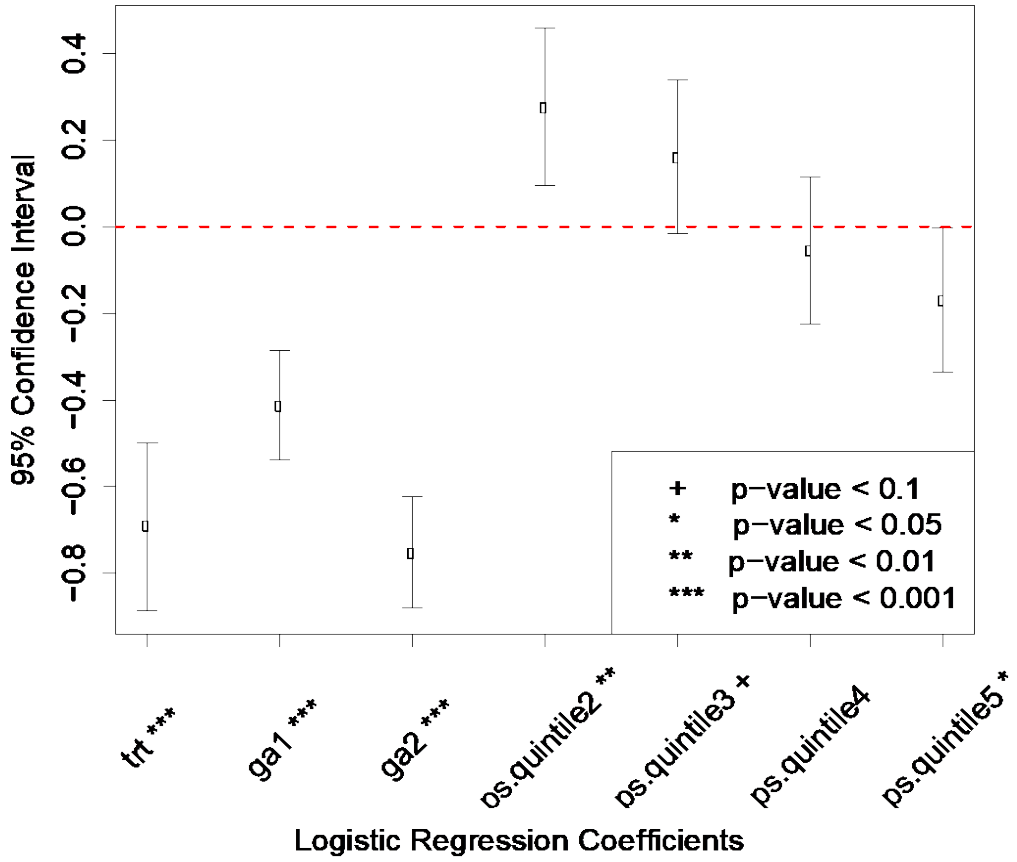
Appendix 10. Histogram of estimated propensity scores.



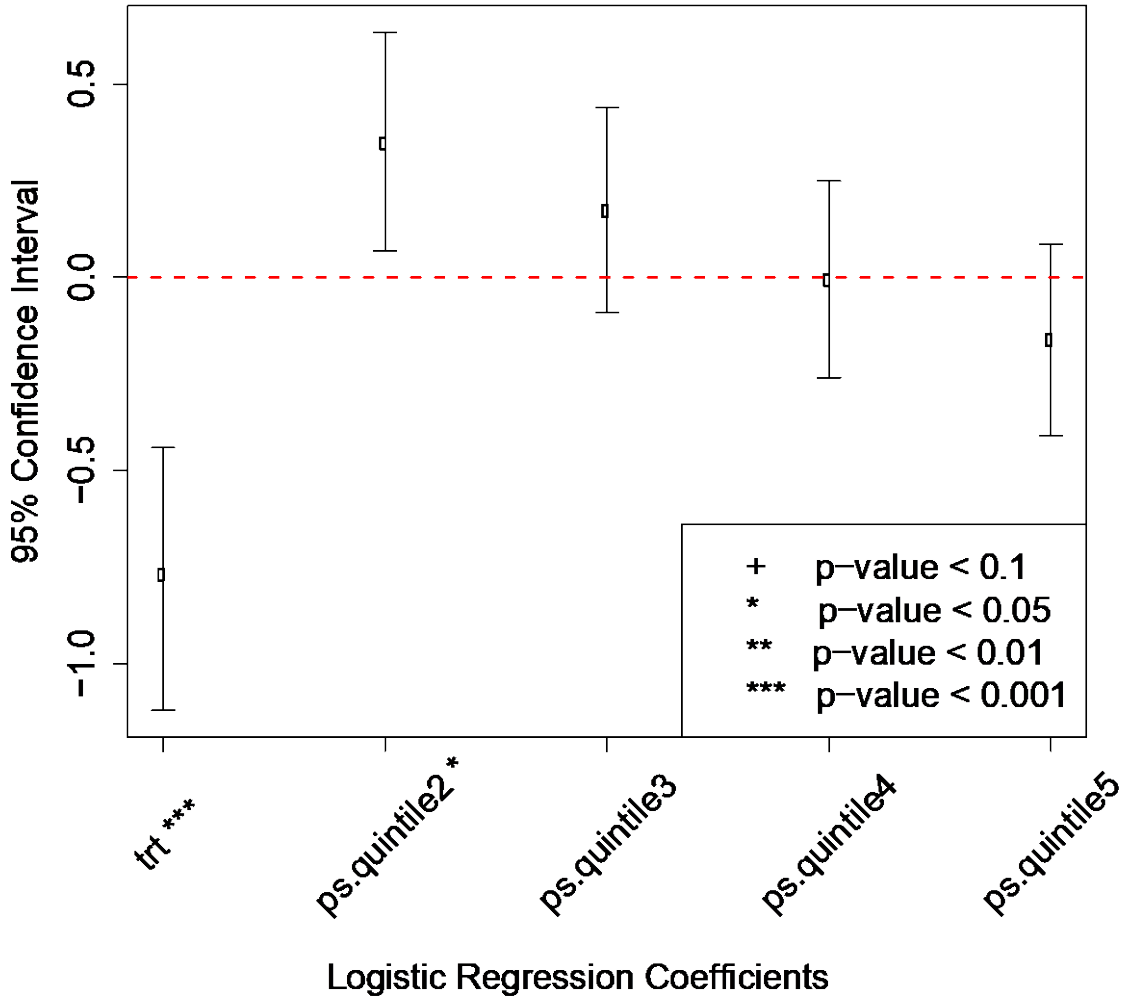
Appendix 11. Density estimate of propensity scores (on logit scale) for simultaneous (sim) and interval (int).



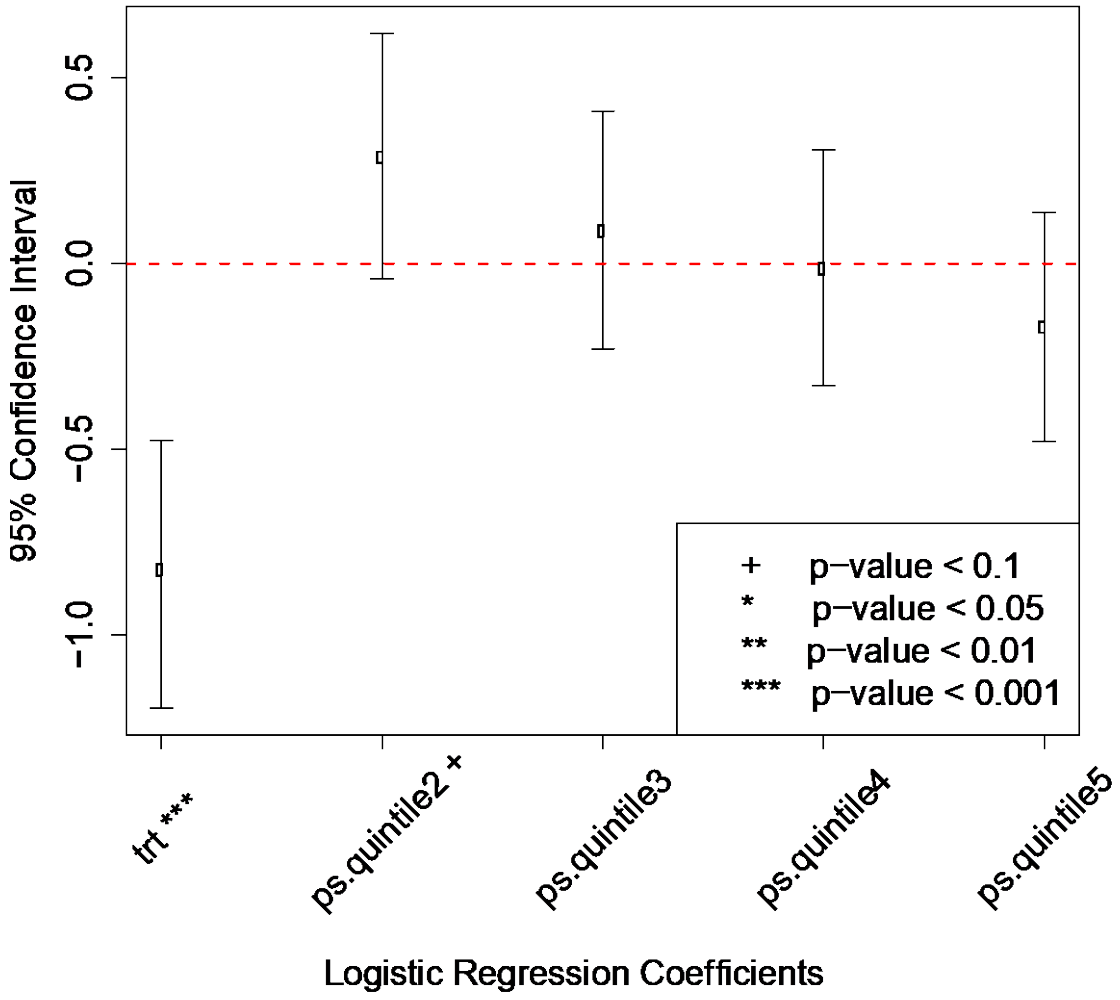
Appendix 12. Logistic regression log odds estimates for overall model.



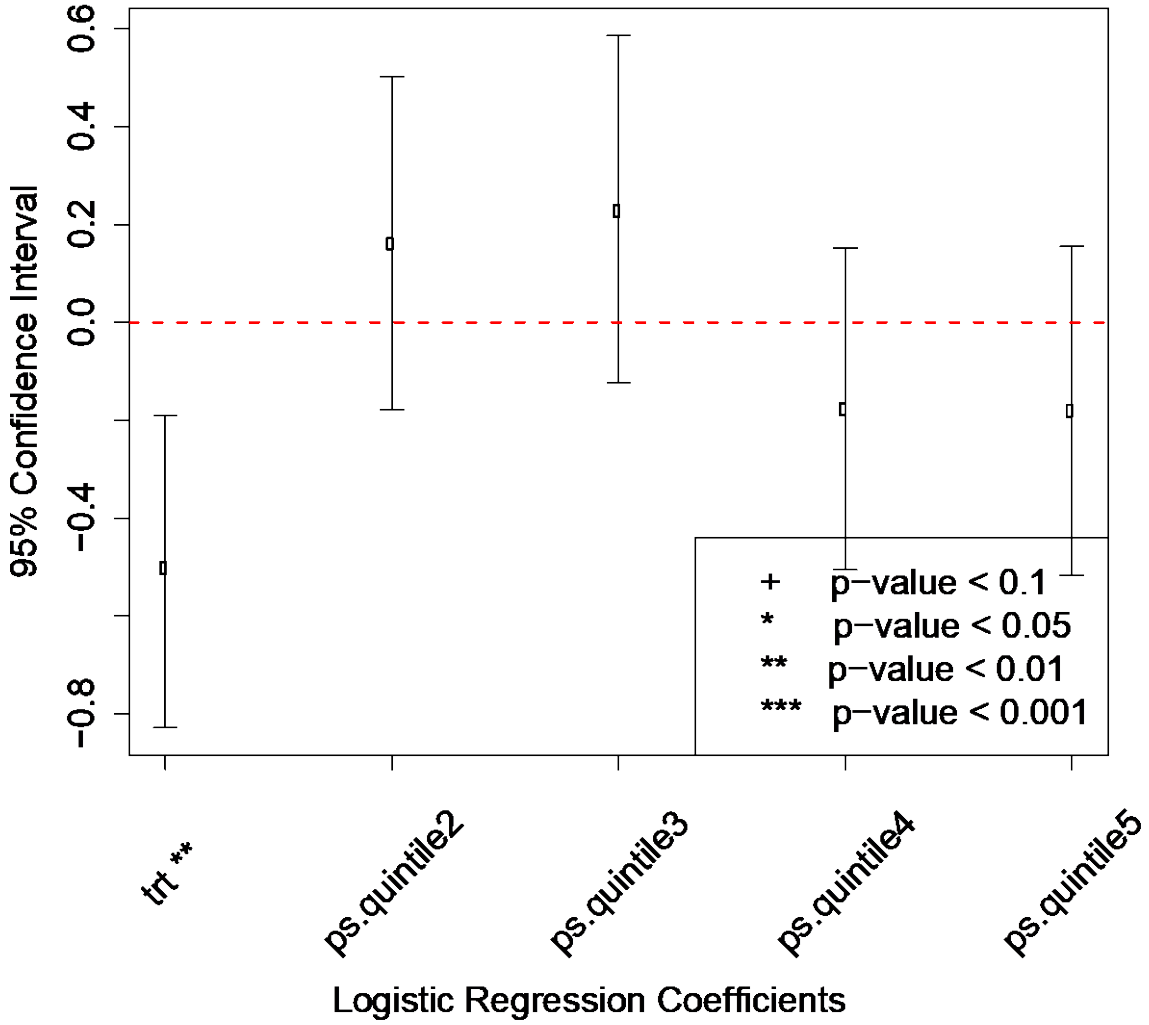
Appendix 13. Logistic regression log-odds estimates for gestational age ≤49 days (ga0).



Appendix 14. Logistic regression log-odds estimates for gestational age 50–56 days (ga1).



Appendix 15. Logistic regression log-odds estimates for gestational age 57–63 days (ga2).



References

1. Lee, B. K., Lessler, J. and Stuart, E. A. 2010. Improving propensity score weighting using machine learning. Statistics in Medicine, 29: 337346.

2. Austin, Peter C. Using Ensemble-Based Methods for Directly Estimating Causal Ef- fects: An Investigation of Tree-Based G-Computation. Multivariate behavioral research 47.1 (2012): 115135. PMC. Web. 27 Oct. 2017.

3. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011;46:399- 424.

4. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. Stat Med 2006; 25: 20842106.

5. Rosenbaum P.R. Model-based direct adjustment. The Journal of the American Statis- tician. 1987a;82:387394.

6. Rosenbaum P.R., Rubin D.B. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983a;70:4155.

7. Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Associa- tion, 79, 516524.

8. Greenland, S. (2004). Model-based Estimation of Relative Risks and Other Epidemio- logic Measures in Studies of Common Outcomes and in Case-Control Studies. American Journal of Epidemiology 160(4): 3015.

9. Greenland, S. (1991), Estimating standardized parameters from generalized linear models. Statist. Med., 10: 10691074.

