Targeted Smooth Bayesian Causal Forests: An analysis of heterogeneous treatment effects for simultaneous versus interval medical abortion regimens over gestation

Jennifer E. Starling†* Jared S. Murray†‡ Patricia A. Lohr§
Abigail R.A. Aiken‖ Carlos M. Carvalho†‡ James G. Scott†‡

†Department of Statistics and Data Sciences, ‡McCombs School of Business, and ‖Lyndon B. Johnson School of Public Affairs, The University of Texas at Austin, Austin, TX
§British Pregnancy Advisory Service, Stratford-upon-Avon, UK

Abstract

We introduce Targeted Smooth Bayesian Causal Forests (tsBCF), a nonparametric Bayesian approach for estimating heterogeneous treatment effects which vary smoothly over a single covariate in the observational data setting. The tsBCF method induces smoothness by parameterizing terminal tree nodes with smooth functions, and allows for separate regularization of treatment effects versus prognostic effect of control covariates. Smoothing parameters for prognostic and treatment effects can be chosen to reflect prior knowledge or tuned in a data-dependent way.

We use tsBCF to analyze a new clinical protocol for early medical abortion. Our aim is to assess the relative effectiveness of simultaneous versus interval administration of mifepristone and misoprostol over the first nine weeks of gestation. Our analysis yields important clinical insights into how to best counsel patients seeking early medical abortion, where understanding even small differences in relative effectiveness can yield dramatic returns to public health. The model reflects our expectation that the treatment effect varies smoothly over gestation, but not necessarily over other covariates. We demonstrate the performance of the tsBCF method on benchmarking experiments. Software for tsBCF is available at https://github.com/jestarling/tsbcf/.

Keywords and phrases: Bayesian additive regression tree, causal inference, regularization, Gaussian process, heterogeneous treatment effects

*Corresponding author: jstarling@utexas.edu
1 Introduction

The goal of this paper is to characterize the effectiveness across gestational age of a new clinical protocol for early medical abortion (EMA), using data from British Pregnancy Advisory Service (BPAS), a non-profit abortion provider in the United Kingdom. Early medical abortion involves a combination of mifepristone and misoprostol and is available through 63 days of gestation in Britain. The current standard of EMA care is an “interval” dosing protocol, where 200 mg oral mifepristone is administered in a clinic setting, followed by administration of 800 micrograms vaginal misoprostol during a second clinic visit 24–48 hours later [Royal College Of Obstetricians And Gynaecologists, 2011]. BPAS, however, recently began offering the option of a “simultaneous” protocol, where patients receive both medications in a single clinic visit, along with instructions about what to do at home to ensure safety and efficacy.

Past research has shown that UK patients overwhelmingly prefer the simultaneous protocol, and that the simultaneous protocol is nearly (97%) as effective as the interval protocol, on average across gestational age [Li et al.; Goel et al.; Lohr et al., 2018]. But these past findings also raise a follow-up question of critical interest to abortion-care providers: whether the simultaneous protocol has a heterogeneous treatment effect across gestational age. In particular, the concern among providers is whether the small decrease in efficacy of the simultaneous protocol might be magnified into a much larger gap at later gestational ages (in particular, 7–9 weeks).

The answer to this question has important public-health consequences even under normal circumstances. But in 2020—as abortion providers across the world confront a radically changed service-delivery landscape in the wake of the global Covid-19 pandemic—the potential consequences are all the more dramatic. In response to the pandemic, UK abortion regulations were modified in March 2020 to allow at-home provision of early medical abortion services with support via telemedicine [Heffer, 2020]. In such an unprecedented context—one where direct patient access to medical support staff is limited at best—the question of how to counsel EMA patients effectively, using the best possible data, is in urgent need of answers. Having a better understanding of even small differences in relative effectiveness, when scaled up to the 200,000+ abortions that take place in the UK annually, and the many millions more worldwide, can yield dramatic returns to public health.

Answering these questions using BPAS’s data, however, poses many statistical complications. Most importantly, patients self-select into the simultaneous or interval protocol. In an earlier paper, Lohr et al. (2018) aimed to control for this potential selection bias using the standard technique of logistic regression with a propensity score adjustment. They confirmed the 97% average efficacy gap found in previous research. But there are three issues with the approach taken in Lohr et al. (2018), which together make it less than ideal for finding heterogeneous treatment effects. First, assuming a simple parametric form for the probability of successful EMA, given treatment and covariates, opens the possibility of bias due to model mis-specification. Second, their analysis treated gestational age—the dominant predictor of a successful EMA—as a categorical variable, discretizing into three gestational age ranges. Yet there are strong reasons to believe that the probability of a successful EMA varies smoothly over gestation, with no sharp “jumps” implied by discretization. Third, their approach estimates average relative effectiveness of the simultaneous protocol over gestation, but does not address subgroups. A key clinical question is whether there are subgroups of patients for whom the efficacy gap is much wider than average, especially at higher gestational ages. Identifying these subgroups would have profound implications for the way that BPAS and other family-planning clinics counsel their patients.

Our approach for causal inference, which builds upon the Bayesian Causal Forests (bcf) model of Hahn et al. (2020) and the BART with Targeted Smoothing model of Starling et al. (2020), is
capable of addressing all of these shortcomings. Our model, called Targeted Smooth Bayesian Causal Forests (tsBCF), is nonparametric, and so it avoids the potential for biases that arise in assuming specific functional forms for the causal estimand. It is capable of detecting heterogeneous treatment effects, should they exist. Finally, it allows for smoothness over a single “target” covariate, addressing a key statistical issue that arises in virtually all pregnancy-related research: that most outcomes vary smoothly with gestational age.

We validate previous findings (Lohr et al., 2018) that on average, we do not see evidence supporting decrease in relative efficacy of the simultaneous protocol with increasing gestational age. We note some decrease in the 7–9 week range (from 0.972 at 6 weeks, to 0.955 at 7 weeks, to 0.948 at 9 weeks), which may be attributed to small sample sizes at later gestations. We identify a drop in efficacy at later gestations among patients age 29 and older, particularly for those who have given birth previously. While the relative effectiveness of the simultaneous protocol for these patients is still over 90% in the 7–9 week range, clinicians may wish to counsel their patients accordingly.

The rest of the paper proceeds as follows. In Section 2, we provide clinical background and an overview of the dataset. Section 3 presents the tsBCF model and reviews relevant literature. Section 4 gives results for the early medical abortion analysis. Section 5 details a simulation study showing that tsBCF maintains accuracy and nominal coverage in recovering heterogeneous relative effectiveness for several clinically relevant scenarios. Section 6 concludes with discussion. The Appendix provides additional detail on fitting the tsBCF model.

## 2 Early Medical Abortion Data

Our dataset consists of de-identified records of early medical abortions provided at BPAS clinics from May 1, 2015 to April 30, 2016. Data was collected from BPAS’s electronic booking and invoicing system, which contained records of services provided to clients, including selected demographic and clinical characteristics. These data were entered by telephone operators at British Pregnancy Advisory Service’s telephone contact center. Details were then validated by clinicians at both in-person consultations and treatment appointments. When possible, hospital discharge summaries or documents from general practitioners were obtained to confirm outcomes. Staff cross-checked the booking and invoicing system for any appointments with British Pregnancy Advisory Services after the date of treatment, and hand-checked medical records if a continuing pregnancy or incomplete abortion was recorded in the complications database. Complications and adverse outcomes were identified either during post-treatment follow-up visits, by patients themselves, or by notifications from other providers. This study was approved and exempted from full human subjects review by British Pregnancy Advisory Services since all data were pre-existing and were provided in a fully de-identified format.

The final dataset for analysis consists of 28,895 patient records. Each record is that of a pregnancy between 4.5 weeks (32 days) and 9 weeks (63 days) gestation, as determined by abdominal or vaginal ultrasonography, where the patient sought a medical abortion and had no contraindications. Patients are typically not aware of pregnancy prior to 4 weeks, and only 12 patients obtained early medical abortions below 4.5 weeks; thus our sample for analysis begins at 4.5 weeks. Early medical abortion is unavailable in the UK after 9 weeks.

Patients were offered the choice between two regimens:

- the interval regimen, with 200mg oral mifepristone administered in a first clinic visit, followed by 800 micrograms vaginal misoprostol in a second clinic visit 24–72 hours later; or
- the simultaneous regimen, where the two medications were administered in a single clinic visit no more than 15 minutes apart.
Table 1: Descriptive characteristics of patients choosing simultaneous or interval administration of mifepristone and misoprostol for early medical abortion. Patients obtained early medical abortion from May 1, 2015 to April 30, 2016 at British Pregnancy Advisory Service clinics. P-values test for differences in distribution of patient characteristics between the simultaneous and interval groups using chi-squared tests.

| Characteristic                        | (N = 28,895) | Interval: (N = 4,354) | Simultaneous: (N = 24,541) | P-value |
|--------------------------------------|--------------|-----------------------|----------------------------|---------|
| Gestational age in weeks, n (%)      |              |                       |                            | < 0.0001|
| 4.5                                  | 407 (1.41)   | 20 (0.46)             | 387 (1.58)                 |         |
| 5-5.5                                | 4,917 (17.02)| 531 (12.20)           | 4,386 (17.87)              |         |
| 6-6.5                                | 9,453 (32.72)| 1,244 (28.57)         | 8,209 (33.45)              |         |
| 7-7.5                                | 7,875 (27.25)| 1,368 (31.42)         | 6,507 (26.51)              |         |
| 8-8.5                                | 5,577 (19.30)| 1,019 (23.40)         | 4,558 (18.57)              |         |
| 9                                    | 666 (2.30)   | 172 (3.95)            | 494 (2.01)                 |         |
| Maternal age in years, n (%)         |              |                       |                            | < 0.0001|
| (11,20]                              | 5,312 (18.38)| 888 (20.40)           | 4,424 (18.03)              |         |
| (20,30]                              | 15,010 (51.95)| 2,178 (50.02)        | 12,832 (52.29)             |         |
| (30,40]                              | 7,695 (26.63)| 1,181 (27.12)         | 6,514 (26.54)              |         |
| Maternal ethnicity, n (%)            |              |                       |                            | < 0.0001|
| Asian                                | 2,610 (9.03) | 510 (11.71)           | 2,100 (8.56)               |         |
| Black                                | 1,839 (6.36) | 310 (7.12)            | 1,529 (6.23)               |         |
| Not Reported                         | 537 (1.86)   | 64 (1.47)             | 473 (1.93)                 |         |
| Other                                | 1,568 (5.43) | 247 (5.67)            | 1,321 (5.38)               |         |
| White                                | 22,341 (77.32)| 3,223 (74.02)        | 19,118 (77.90)             |         |
| BMI, n (%)                           |              |                       |                            | = 0.5364|
| Underweight (<18.5)                  | 2,149 (7.44) | 329 (7.56)            | 1,820 (7.42)               |         |
| Normal (18.5-24.9)                   | 14,667 (50.76)| 2,209 (50.73)        | 12,458 (50.76)             |         |
| Overweight (25.0-29.9)               | 7,367 (25.50)| 1,080 (24.80)         | 6,287 (25.62)              |         |
| Obese (>30.0)                        | 4,712 (16.31)| 736 (16.90)           | 3,976 (16.20)              |         |
| Previous abortions, n (%)            |              |                       |                            | < 0.0001|
| 0                                    | 18,354 (63.52)| 2,949 (67.73)        | 15,405 (62.77)             |         |
| 1-2                                  | 9,876 (34.18)| 1,301 (29.88)         | 8,575 (34.94)              |         |
| 3+                                   | 665 (2.30)   | 104 (2.39)            | 561 (2.29)                 |         |
| Previous births, n (%)               |              |                       |                            | < 0.0001|
| 0                                    | 18,354 (63.52)| 2,949 (67.73)        | 15,405 (62.77)             |         |
| 1-2                                  | 5,070 (17.55)| 609 (13.99)           | 4,461 (18.18)              |         |
| 3+                                   | 5,471 (18.93)| 796 (18.28)           | 4,675 (19.05)              |         |
| Previous Cesarean sections, n (%)    |              |                       |                            | < 0.0001|
| 0                                    | 18,354 (63.52)| 2,949 (67.73)        | 15,405 (62.77)             |         |
| 1-2                                  | 1,346 (4.66) | 169 (3.88)            | 1,177 (4.80)               |         |
| 3+                                   | 9,195 (31.82)| 1,236 (28.39)         | 7,959 (32.43)              |         |
| Previous miscarriages:               |              |                       |                            | < 0.0001|
| 0                                    | 18,354 (63.52)| 2,949 (67.73)        | 15,405 (62.77)             |         |
| 1-2                                  | 2,046 (7.08) | 264 (6.06)            | 1,782 (7.26)               |         |
| 3+                                   | 8,495 (29.40)| 1,141 (26.21)         | 7,354 (29.97)              |         |
in Table[1] including estimated gestational age; maternal age; maternal ethnicity; maternal body-mass index (BMI); and number of previous abortions, births, Cesarean sections, and miscarriages. The question of clinical interest is whether the relative effectiveness of the simultaneous versus interval protocol changes over gestational age or other maternal covariates.

3 Targeted Smooth Bayesian Causal Forests

This section introduces Targeted Smooth Bayesian Causal Forests (tsBCF), which allows us to estimate heterogeneous but smooth treatment effects over gestation. In this section we fix notation, briefly describe the tsBCF model, and discuss the assumptions necessary for causal identification. We then review relevant literature. Many details of the tsBCF model formulation and implementation are deferred to the appendix.

Let \( y \) denote a binary response, \( z \) a binary treatment indicator, \( t \) a scalar “target” covariate, and \( x \) a \( p \)-length vector of additional covariates. From the standpoint of causal inference, there is no difference between \( t \) and \( x \) in our framework; both are covariates that are potentially confounded with, and modify the effect of, the treatment variable. There is, however, a statistical difference between \( x \) and \( t \): we assume that the treatment effect is smoothly varying in \( t \), but not necessarily in \( x \). In our application, \( y_i \) indicates success of the early medical abortion, and \( z_i \) indicates the patient’s selection of simultaneous (\( z = 1 \)) versus interval (\( z = 0 \)) protocols. Our target covariate \( t_i \) is gestational age, measured in half-weeks ranging from 4.5 (32 days) to 9 (63 days). The other covariates \( x_i \) are maternal age (years), Body Mass Index (kg/m\(^2\)), maternal ethnicity (Asian, Black, Other, Not Reported, White), and numbers of previous abortions, births, Cesarean sections, and miscarriages. We assume that our sample consists of independent observations \((y_i, t_i, z_i, x_i)\) for \( i \in \{1, \ldots, n\} \).

Our framework for causal inference assumes that

\[
P(y_i = 1 \mid t_i, x_i, z_i) \equiv r(t_i, x_i, z_i) = \Phi(f(t_i, x_i, z_i))
\]

where \( \Phi \) is the probit link and \( f \) is a real-valued regression function. Our formulation is motivated by two key scientific concerns. First, we want to induce smoothness over \( t \) while remaining agnostic to smoothness over \( x \). Second, we want to capture three-way interactions between \( t, x, \) and \( z \), without specification of any restrictive parametric forms. The probit link function is used for computational convenience, since it allows Bayesian inference via the latent-variable method of Albert and Chib (1993):

\[
\begin{align*}
y_i &= \begin{cases} 1 & \text{if } \tilde{y}_i \geq 0 \\ 0 & \text{if } \tilde{y}_i < 0 \end{cases} \\
\tilde{y}_i &= f(t_i, x_i, z_i) + \epsilon_i \\
\epsilon_i &\sim N(0, 1)
\end{align*}
\]

While a log link would allow for simpler estimation of relative risk, preserving homogeneity or heterogeneity in the treatment effect, domain restriction and computational tractability in the MCMC are limitations. Moreover, since the estimated probabilities are compositions of the link function and a flexible nonparametric regression, the log and probit links recover similar probabilities.

Our causal estimand is the conditional relative effectiveness, given \( x \) and \( t \), of the simultaneous
versus interval protocol. This takes the mathematical form of a relative risk:

\[
RR(t, x) = \frac{P(y_i = 1 \mid t_i, x_i, z_i = 1)}{P(y_i = 1 \mid t_i, x_i, z_i = 0)} = \frac{\Phi \left[ f(t, x, z = 1) \right]}{\Phi \left[ f(t, x, z = 0) \right]}.
\]  

The relative-risk scale is not the most convenient for the purposes of causal inference; in particular, trying to estimate a relative risk in our framework creates a subtle phenomenon that we refer to as “structural heterogeneity,” discussed in detail below. However, relative efficacy is a standard estimand in the literature on abortion medications—and in particular, it the estimand studied by Lohr et al. (2018), who analyzed this same data set (though we allow RR to depend on \(x\) and \(t\)).

We formulate assumptions using the counterfactual outcomes framework of Imbens and Rubin (2015), where \(y_i(0)\) and \(y_i(1)\) denote potential outcomes under interval and simultaneous protocols, respectively. Observations are assumed to correspond to realized potential outcomes, such that \(y_i = z_i y_i(1) + (1 - z_i) y_i(0)\). We make three other standard assumptions: non-interference, no unmeasured confounders, and sufficient overlap to estimate treatment effects everywhere in covariate space. We now list these more formally.

**Assumption 1.** We make the stable unit treatment value assumption (SUTVA), which excludes interference between units and multiple versions of treatment (Imbens and Rubin, 2015).

**Assumption 2.** We assume strong ignorability, which specifies that

\[
(y_i(0), y_i(1)) \perp z_i \mid t_i, x_i
\]

and

\[
0 < P(z = 1 \mid t_i, x_i) < 1
\]

hold for all \(i \in 1, \ldots, n\) observations.

Our statistical model represents \(f(t, x, z)\) as follows:

\[
f(t_i, x_i, z_i) = \mu(t_i, x_i, \hat{\pi}_i) + z_i \tau(t_i, x_i),
\]  

where \(\hat{\pi}\) is an estimate of the propensity score, \(P(z_i = 1 \mid x_i, t_i)\). Under our causal-identification assumptions, we can express the relative effectiveness of the simultaneous protocol (3) as

\[
RR(t_i, x_i) = \frac{\Phi \left[ \mu(t_i, x_i, \hat{\pi}_i) + \tau(t_i, x_i) \right]}{\Phi \left[ \mu(t_i, x_i, \hat{\pi}_i) \right]}.
\]  

We also calculate a secondary outcome, number needed to treat, which will aid in interpreting our results. We invert the absolute risk reduction to obtain

\[
NNT(t_i, x_i) = 1/ \left[ \Phi \left( \mu(t_i, x_i, \hat{\pi}_i) + \tau(t_i, x_i) \right) - \Phi \left( \mu(t_i, x_i, \hat{\pi}_i) \right) \right].
\]  

This quantity represents the number of patients who would need to select the simultaneous protocol before one patient is impacted by EMA failure due to selecting the simultaneous instead of interval regimen.

**Statistical modeling of prognostic and treatment effects.** We assume that both \(\mu(t_i, x_i, \hat{\pi}_i)\) and \(\tau(t_i, x_i)\) have independent tsBART priors (Starling et al. 2020). The tsBART prior is a recent proposal for nonparametric regression designed to induce smoothness in the target covariate \(t\).
It is analogous to the BART model of Chipman et al. (2010), except that it replaces scalar terminal node parameters with smooth Gaussian process priors over the target covariate \( t \). The resulting model is a sum of function-valued treed Gaussian processes in \( t \) (c.f. Gramacy and Lee, 2008), in the same way that BART is a sum of scalar-valued regression trees. Our adaptation of this method for causal inference also uses Gaussian Process priors to induce smoothness in gestational age \( t \), but it allows for separate regularization of the degree of smoothness for the prognostic (\( \mu \)) and treatment-effect (\( \tau \)) components of the estimate.

Our model formulation is directly based on the Bayesian Causal Forests, or BCF, framework of Hahn et al. (2020), and our parameterization in (4) closely follows the recommendations of their paper. BCF estimates heterogeneous treatment effects using BART priors, allowing for careful regularization of treatment-effect estimates. However, while BCF has several desirable properties and represents a gold standard for estimating heterogeneous treatment effects, it has one fundamental disadvantage for analyzing our data on early medical abortion: it lacks a mechanism to include a priori knowledge of smoothness in gestational age, as we have here. The tsBCF model is designed specifically to address this shortcoming. The primary advantage of this modification is superior interpretability: our goal is to produce estimates of gestational-age-specific effectiveness that can be interpreted easily by clinicians and patients, without putting them in a position whether they might over-interpret small wiggles in the fit that are simply model artifacts due to the lack of smoothing across gestational age. Incorporating smoothing directly into the model, as opposed to smoothing nonparametric fits post hoc, is a statistically rigorous way to generate more interpretable model summaries.

For details on the tsBART model and its parameterization, we refer the reader to Starling et al. (2020), since we use that paper’s notation in describing our hyperparameter choices. Our model does not modify the tsBART priors, except to select different numbers of trees and hyperparameters for the prognostic versus treatment priors. To model \( \mu \), we use a tsBART prior with 200 trees, depth penalty \( \beta = 2 \), splitting probability \( \eta = 0.95 \), and default smoothing parameter \( \kappa_\mu = 1 \), with a half-Cauchy prior on the scale of leaf parameters (Gelman, 2006). We model \( \tau \) using 50 trees instead of 200, reflecting our belief that patterns of heterogeneity in the treatment are generally simpler than heterogeneity in the prognostic effect of the covariates (Hahn et al., 2020). We also use stronger regularization in \( \tau \) than in \( \mu \), setting \( \beta = 3 \), \( \eta = 0.25 \), and a half-Normal prior on the scale of the tree leaves, with the scale of the half-Normal prior adjusted via a fixed scale in the leaf prior for \( \tau \), as in Hahn et al. (2020). We set \( \kappa = 1 \); Starling et al. (2020) offer approaches for tuning the tsBART smoothing parameter, and similar tuning may be undertaken here in conjunction with prior knowledge about the likely form of heterogeneity. The model is fit via Bayesian backfitting, described in Appendix A.2.

Our model makes it possible for different covariate vectors to be used in estimation of \( \mu \) and \( \tau \), and we make use of this flexibility by including the estimated propensity score in the model for \( \mu \). We refer readers to Rosenbaum and Rubin (1983), Hahn et al. (2020) for a discussion in support of including both covariates and the propensity score. Briefly, propensity score inclusion is an effective dimension-reduction technique yielding a prior that flexibly adapts to complex patterns of confounding, while inclusion of control covariates is necessary when the response does not depend on covariates strictly through the estimated propensity score. We estimate propensity scores using BART with a probit link function.

We refer interested readers to Chipman et al. (2010) for a detailed description of BART. BART has been successful in a variety of contexts including prediction and classification (Chipman et al., 2010; Murray, 2017; Linero, 2018; Hernández et al., 2018), survival analysis (Sparapani et al., 2016; Starling et al., 2020), and causal inference (Hill, 2011; Hahn et al., 2020; Logan et al., 2017; Sivaganesan et al., 2017). Linero and Yang (2017) propose smoothing a regression tree ensemble by
randomizing the decision rules at internal nodes of the tree. This model induces smoothness over all covariates by effectively replacing the step function induced by the binary trees with sigmoids, instead of smoothing over one targeted covariate. Probit versions of BART and tsBART are defined in (Chipman et al., 2010) and (Starling et al., 2020); we draw on these formulations in specifying the model for our binary response.

### Structural heterogeneity

The major difficulty in fitting and interpreting our model, versus the original BCF model, is that the original BCF model is formulated for a continuous response. Our outcome is binary, and our causal estimand is a relative risk: non-linear transformation of the “prognostic” and “treatment” effect parameters $\mu$ and $\tau$. This raises an issue not addressed in the BCF literature, which we call structural heterogeneity. To see this phenomenon, suppose that $\tau$ were actually homogenous across all units: that is $\tau(x_i, t_i) \equiv \tau_0$. In this case, our model would estimate the relative effectiveness of the simultaneous protocol as

$$RR_i = \frac{\Phi(\mu_i + \tau_0)}{\Phi(\mu_i)} \text{, where } \mu_i = \mu(t_i, x_i, \hat{\pi}_i).$$

Thus even if the treatment effect $\tau_0$ is homogeneous across units on the probit scale, our causal estimand is not homogenous across all units: it depends on $\mu_i$, each unit’s baseline probability of success. This is a mathematical artifact of two structural modeling choices:

- parameterizing the treatment effect on the probit scale, which allows efficient inference for a binary outcome in the Bayesian framework; and
- estimating a relative effectiveness, which makes our work consistent with the prior literature on the effectiveness of abortion medication.

Said concisely: shrinking treatment effects $\tau(x, t)$ toward a common homogenous effect $\tau_0$ does not necessarily imply shrinkage towards homogeneous relative risks.

This fact raises an important question in interpreting the results of our model. Suppose we see heterogeneous relative risks a posteriori. How much of this heterogeneity is due to “structural” variation in the prognostic effects $\mu_i$, versus the presumably more interesting “causal” variation in treatment effects $\tau_i$? In Appendix A.3 we undertake a detailed study of this issue. To briefly summarize our results:

- The degree to which structural heterogeneity depends on the average baseline risk, treatment effect size and variability (parametrized by a scale $s_\tau$), and variability in $\mu_i$ (parametrized by a scale $s_\mu$). To account for this, we propose setting $s_\mu$ based on prior elicitation of a plausible range of baseline risks, and then calibrating $s_\tau$ by calculating an average relative risk for a small hold-out sample.
- For our analysis of the BPAS data, we estimate that no more than 48% of the posterior variation in relative risks is structural. The remainder is driven by heterogeneous causal effects, i.e. differences in treatment effects $\tau(x_i, t_i)$. We estimate this percent by comparing the variance of our posterior relative risk estimates to the variance in relative risks calculated using the posterior mean $\tau(x_i, t_i)$ estimate for each MCMC draw.

See Appendix A.3 for details.

### 4 Early Medical Abortion Modeling

We now focus on our scientific problem, estimating relative effectiveness of simultaneous versus interval administration of mifepristone and misoprostol across gestation. We apply tsBCF to
the BPAS data described in Section 2. Our goal is to model relative effectiveness of the simultaneous versus interval protocols across gestational age, and identify whether there are subgroups of patients for whom the drop in efficacy under simultaneous is markedly wider at later gestations (7–9 weeks). We use “relative effectiveness” to refer to relative risk when interpreting results, as discussing risk of successful procedure is less clinically intuitive; a value of 0.95 is interpreted as the simultaneous protocol being 95% as effective as the interval protocol.

For each MCMC iteration (indexed by \( b \)) of the tsBCF backfitting algorithm, we draw posterior relative risk \( \hat{R}R_i^{(b)} \) for patient \( i \). Averaging across patients observed at each gestational age yields draws of posterior relative risk draws for that gestation. Averaging across subgroups of patients at a given gestation yields posterior relative risk draws for that subgroup–gestation combination. We summarize results with posterior means and 95% credible intervals. Formally, for MCMC draws \( b \in \{1, \ldots, B\} \), \( \hat{R}R_i^{(b)} \) is the \( b^{th} \) draw for relative risk for individual \( i \) who is observed at gestational age \( t_i \). We obtain posterior draws of estimated relative effectiveness at gestational age \( t \) as

\[
\hat{R}R_t^{(b)} = \sum_{i|t_i=t} \hat{R}R_i^{(b)}.
\]  

(7)

The posterior mean and credible interval for \( \hat{R}R_t \) are calculated using draws \( \{\hat{R}R_1^{(1)}, \ldots, \hat{R}R_t^{(B)}\} \). Conditioning on \( x \) and \( t \) in the summation (7) gives gestation-specific posterior mean relative risk for that subgroup.

We obtain posterior draws of our secondary outcome, number needed to treat, at gestational age \( t \) using the same posterior draws of \( \hat{\mu}(t_i, x_i, \hat{\pi}_i) \) and \( \tau(t_i, x_i) \). We calculate our secondary outcome, number needed to treat (6,(8). In this setting, we can also interpret NNT as number needed to fail. Large values indicate that the benefits of offering simultaneous administration are not offset by many additional failed procedures (Figure 1, Panel B).

We also investigate the degree to which confounding is likely present due to targeted selection (Hahn et al., 2018). We fit a logistic regression model to our binary outcome using estimated propensity scores as the covariate. The pseudo-\( R^2 \) (McFadden, 1974) is 0.004, indicating lack of evidence for targeted selection. Finally, we compare variability in posterior relative risk under heterogeneous versus homogeneous \( \tau_i \), and find that there is an average of 2.06 times more variability in posterior relative risk under heterogeneous versus homogeneous \( \tau_i \), giving us confidence that the following clinical findings for subgroups and individual patients are not just an artifact of
Figure 1: Average relative effectiveness and number needed to treat for the cohort. (A) Posterior mean relative effectiveness (solid line) and 95% credible interval (shaded) for simultaneous versus interval protocols, averaged over all patients, across gestational age. The dashed line gives the posterior mean from BCF for comparison. (B) The posterior mean number of patients who would need to select the simultaneous protocol before one patient is impacted by EMA failure due to selecting the simultaneous instead of interval regimen; the number needed to treat, or equivalently, number needed to fail. The average relative effectiveness remains high over the course of gestation, as does the number needed to treat, indicating comparable effectiveness of the two regimens.

Subgroup analysis. We now aim to identify subgroups of patients who may have a larger gap in efficacy at later gestational ages (7–9 weeks). As the posterior formally quantifies our uncertainty, we use exploratory techniques to present the complex posterior object in more visually tractable ways while avoiding statistical modeling downstream of the posterior draws. We focus on patients in the 7–9 weeks gestation range, as high effectiveness and narrow credible intervals at earlier gestations support recommending simultaneous administration for all patients. Specifically, we take a “fit-the-fit” approach to subgroup analysis (Hahn et al., 2020), where individual relative effectiveness estimates \( \hat{RR}_i = \sum_{b=1}^{B} \bar{RR}_i^{(b)} \) for patients observed at 7–9 weeks are used as the response in a CART model, with covariates \( x_i \).

The CART tree fits (Figure 2) identify maternal age as the most important subgroup; patients 29 and older have somewhat lower relative effectiveness (Panel a) compared to their younger counterparts. Number of previous births also decreases efficacy, though slightly less so in the younger group of patients. Panel b gives the CART fit on the \( \tau \) scale, illustrating that subgroup findings are not a consequence of transforming the outcomes to relative risk.

The CART fit gives point estimates of relative effectiveness for each node; without uncertainty quantification, it is unclear whether these subgroups are meaningfully different from each other. We query the posterior draws for subgroups at each level of the CART tree. Less overlap in posterior densities indicates meaningful splits, and more dispersion indicates greater heterogeneity within a tree split subgroup. We plot posteriors for the top split and the terminal nodes (Figure 3). The split on age 29 is the most important subgroup, with little overlap in the posteriors. Presence of previous births had a slight negative impact on relative effectiveness; this was more pronounced in the older cohort, subtler in patients ages 22–28, and absent in patients under 22.

To quantify how subgroup differences at later gestations translate to patient impact, we plot the distribution (over MCMC draws) of subgroup average NNT differences between the leftmost
and rightmost leaves of the CART tree (Figure 4). Differences generally range from 10 to 40 patients, with no mass at zero, indicating that the relative effectiveness gap in these subgroups translates to real difference in number of patients in each subgroup who would need to select the simultaneous protocol at 7–9 weeks before seeing one additional failed procedure.

**Individual relative effectiveness estimates.** Inspecting individual relative effectiveness estimates ensures that there are not smaller subgroups of patients undetected by our analysis, where relative effectiveness is markedly lower. Figure 5 plots posterior mean relative effectiveness for each patient, by age (Panel A) and number of previous births (Panel B), showing the full range of individual relative effectiveness estimates. Trends in age and previous births are consistent with our subgroup analysis (Figures 2, 3). Increase in both covariates correspond with decreased relative effectiveness; this trend is stronger in age than previous births, consistent with their respective positions in the CART tree. Table 2 provides detail on cohort characteristics by posterior mean relative effectiveness below 0.90, from 0.90–0.95, and above 0.95.

Posterior projection plots give interpretable lower-dimension model summaries for age and previous births across gestation. The Woody et al. (2019) approach to posterior summarization and interpretation involves two stages: first, fitting a flexible model, and second, projecting posterior draws from the model onto simpler structures such as linear or generalized additive models. As our stage one tsBCF model is Bayesian, summaries come with valid Bayesian credible intervals, as data is only used once to move from prior to posterior (Woody et al., 2019). We project using a generalized additive model across gestation by age groups (Figure 6, Panel A) and presence of previous births (Figure 6, Panel B), with shaded credible intervals projected from the posterior.

**Clinic resource planning.** In addition to counseling patients, clinics must plan staff and resources appropriately when offering the simultaneous protocol. The treatment effect on the treated gives insight here; for patients who experienced a failed procedure under the simultaneous protocol, we plot the distribution of MCMC draws of the differences in the observed number of failures compared to the expected failures had those patients selected the interval protocol (Figure 7). We report these differences on the order of expected additional surgeries per thousand patients, giving clinics a sense of the likely volume of procedures. We find that a clinic can an-
Figure 3: Posterior distributions of relative effectiveness for each CART tree split from Figure 2. Visualizing tree split posteriors gives insight into significance of differences between subgroups of patients. Age is the most important covariate defining subgroups, with a split at age 29. Within each age group, previous births decrease relative effectiveness slightly.

anticipate approximately 40–60 additional surgeries per thousand patients per 1,000 patients treated when offering the simultaneous protocol.

Sensitivity to smoothing parameters. The early medical abortion analysis used our suggested default smoothness parameter settings ($\kappa_\mu = 1$ and $\kappa_\tau = 1$). Here, we perform a sensitivity analysis for robustness of our analysis to smoothness parameter choice. We let $\kappa_\mu = 1$ and fit the tsBCF model to the early medical abortion dataset three times, for $\kappa_\tau \in \{1/3, 1, 3\}$. These choices represent a three-times change in magnitude in each direction from the default, corresponding to varying the length-scale of the treatment trees’ covariance from one to three to nine. While there are small differences in the overall estimated relative effectiveness (Figure 8), we do not see clinically meaningful variation across smoothness parameter settings, indicating that our analysis is robust to choice of smoothing parameter. The small differences in week 9 lend support to our intuition that the slight kick-up is due to small sample size.

5 Simulations

We compare tsBCF to several existing models in a benchmarking study designed to simulate five clinically plausible scenarios with a binary response. We generate latent-scale prognostic effects, treatment effects, and random noise for each scenario, and assess how well the models recover relative risk. Our goal is to verify that tsBCF successfully estimates heterogeneous relative risks with reasonable uncertainty quantification and nominal coverage, while inducing smoothness over the target covariate. We compare the following models.

- **tsBCF1**: The tsBCF method with default smoothing parameters $\kappa_\mu = 1$ and $\kappa_\tau = 1$.
- **tsBCF2**: The tsBCF method with smoothing parameters $\kappa_\mu = 1$ and $\kappa_\tau = 3$.
- **BCF**: The Bayesian Causal Forest model described in [Hahn et al. (2020)](Hahn et al. (2020)). We expect this
Figure 4: Distribution of differences in subgroup average NNT for the leftmost and rightmost CART tree leaves. Differences generally range from 10 to 40 patients, indicating that the relative effectiveness gap in these subgroups at later gestational ages translates to real differences in number of patients in each subgroup who would need to select the simultaneous protocol before seeing one additional failed procedure.

model to perform well but lack smoothness (Figure 7).

- **BART**: Ordinary BART used to model the response surface in the causal inference setting (Hill, 2011) with estimated propensity scores included as a covariate (Rosenbaum and Rubin, 1983).

Simulated data is generated as follows. For independent observations $i \in \{1, \ldots, n\}$, draw a vector of covariates $x_i = \{x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}\} \sim N(0, 1)$ and draw target covariate $t_i \in \{0.1, 0.2, \ldots, 1\}$. We generate observations from

$$P(y_i = 1 \mid t_i, x_i, z_i) = \Phi (\mu(t_i, x_{1i}, x_{2i}) + \tau(t_i, x_{3i})z_i)$$

where $\mu$ is the prognostic function and $\tau$ is the treatment effect function. The prognostic function is the same for all five scenarios:

$$\mu(t_i, x_{1i}, x_{2i}) = \frac{1}{4} t_i^{1.5} + \frac{x_{1i}}{6} + \frac{x_{2i}}{4}$$

Each observation is assigned to treatment ($z_i = 1$) or control ($z_i = 0$) based on a binomial draw with propensity score

$$\pi_i = \Phi \left( \rho \cdot \left[ \frac{x_{1i}}{6} - \frac{x_{2i}}{4} \right] + (1 - \rho) \left[ -1 (x_{4i} > 5 + 1(x_{4i} < 5)) \right] \right)$$

where $\rho \in [0, 1]$ controls the degree to which the propensity score is based on a somewhat accurate prediction of the potential outcome, since $\left[ \frac{x_{1i}}{6} - \frac{x_{2i}}{4} \right]$ is found in the prognostic effect (10), while $x_{4i}$ and $x_{5i}$ are not used in prognostic or treatment effect generation. We estimate propensity scores using the **dbarts** R package (Chipman et al., 2010); any accurate prediction is viable (Hahn et al., 2020). We set $\rho = 0.25$ to introduce a small amount of targeted selection.

We vary $\tau(t_i, x_i)$ by scenario, reflecting different latent-scale treatment effects as follows.

- Scenario A represents a treatment effect that varies smoothly over the target covariate $t$, with
Figure 5: Posterior mean relative effectiveness for each patient, by (A) age and (B) number of previous births. Increase in both covariates correspond with decreased relative effectiveness; this trend is stronger in age than previous births, consistent with their respective positions in the CART tree.

Figure 6: Posterior projection plots for age and previous births. (A) Posterior projection of relative effectiveness by age group over gestation. (B) Posterior projection of previous births over gestation. The marginal effect of both covariates is consistent with our subgroup analysis. Points show marginal posterior means across gestation; shaded regions give credible intervals.

homogeneity in $x$.

$$\tau(t_i, x_i) = 0.1 + 0.2t_i - 0.05 \sin(1.5\pi t_i)$$

- Scenario B represents heterogeneous treatment effects that vary smoothly over $t$ with modest differences in subgroups.

$$\tau(t_i, x_i) = 0.1 + 0.2 \mathbb{1}(x_{3i} > -1/2) + 0.15 \mathbb{1}(x_{3i} > 1/2) + 0.2t_i - 0.05 \sin(1.5\pi t_i)$$

- Scenario C represents heterogeneous treatment effects, similar to Scenario B except that the
Figure 7: Distribution of differences in observed failures under simultaneous versus expected failure under interval, for patients who experienced a failed procedure under the simultaneous protocol. Clinics choosing to offer the simultaneous protocol can anticipate an additional 40-60 surgeries per 1,000 patients treated.

\[ \tau(t_i, x_i) = 0.1 + 0.2 \cdot I(x_{3i} > -1/2) + (0.15 + 0.2t_i) \cdot I(x_{3i} > 1/2) + 0.2t_i - 0.05 \sin(1.5\pi t_i) \]

- Scenario D gives heterogeneous treatment effects with small effects in general, except for a small subgroup with a pronounced effect.

\[ \tau(t_i, x_i) = 0.05 + 0.05 \cdot I(x_{3i} > -1/2) + (0.15 + 0.2t_i) \cdot I(x_{3i} > 1/2) + 0.2t_i - 0.05 \sin(1.5\pi t_i) \]

- Scenario E is a constant treatment effect, requiring shrinking to homogeneity in both \( x \) and \( t \).

\[ \tau(t_i, x_{3i}) = 0.1 \]

For 50 replicates of each scenario, we generate a dataset and fit all models for each scenario. Table 3 gives results averaged across replicates for each scenario. RMSE is the average root mean squared error for estimating heterogeneous relative risk; nominal coverage is 95%, and interval length is for posterior credible interval of the relative risk estimates for each unit. TsBCF and BCF perform comparably in RMSE, coverage, and interval length across scenarios. BART performs nearly as well but over-covers. Differences are small compared to standard error size; these results confirm that in introducing smoothness in a target covariate adds interpretability without compromising performance.

Figure 9 illustrates the relationships of RMSE, coverage, and interval length. Panel A gives coverage versus RMSE, where the tsBCF methods slightly outperform BCF while maintaining similar coverage and interval lengths. Panel B gives coverage versus interval length, where tsBART and BCF are maintaining nominal coverage and have similar interval lengths. Panel C gives RMSE versus interval length, where tsBCF and BCF are performing comparably. Values are averaged over scenarios. Together, these panels demonstrate that tsBCF recovers the heterogeneous relative risks while maintaining coverage and yielding reasonable measures of uncertainty in scenarios.
Figure 8: Posterior mean relative effectiveness for three settings of the tsBCF model’s $\kappa$ smoothness parameter. The solid line gives the posterior mean relative effectiveness over gestation from our analysis (Default), with shaded 95% credible interval. Dashed lines give posterior means for $\kappa = 1/3$ (Smother) and $\kappa = 3$ (Wigglier). We do not see clinically meaningful differences in the three fits across gestation, indicating robustness to choice of smoothness parameter.

when the underlying treatment effect is smooth over the target covariate.

Figure 9: Simulation results for each method. Panel A gives coverage versus RMSE. Panel B gives coverage versus interval length. Panel C gives RMSE versus interval length. The tsBCF models recover the heterogeneous relative risks while maintaining coverage and yielding reasonable measures of uncertainty.

Figure 10 shows the advantage of ensuring that the estimated treatment effect function is smooth in the target covariate. We plot the true treatment effects versus estimates recovered from our four models, using Scenario B from the previous simulations to illustrate. Each panel shows the simulated treatment effect (dashed) versus the estimated treatment effect function (solid) for the three heterogeneous subgroups. While BCF and BART recover the treatment effects reliably, only the tsBCF estimate is smooth over the target covariate. As seen in Figure 9 and Table 3, when the true treatment effect is smooth this results in improved accuracy, in addition to ensuring that treatment effect estimates match a priori knowledge of smoothness.
Table 2: Descriptive characteristics by ranges of posterior mean individual relative effectiveness. P-values assess differences in distribution of patient characteristics between the three ranges of relative effectiveness. Only 26 patients (0.09%) have estimated relative effectiveness less than 0.90; all are over 30, and all but one have at least one previous birth. Patients in the 0.90–0.95 category are also predominantly over 30. Additionally, 21,304 patients (74%) have estimated relative effectiveness greater than 0.95.

6 Discussion

Targeted Smooth Bayesian Causal Forests allows for estimation of heterogeneous treatment effects which evolve smoothly over a target covariate. This addresses a key statistical issue that arises in virtually all pregnancy-related research: that most outcomes vary smoothly with gestational age. Our model is nonparametric, and so avoids potential bias arising from specification of functional forms of the causal estimand. TsBCF enjoys similar advantages as BCF, including hyperparameters which are set efficiently using heuristics similar to those of Hahn et al. (2020) and Chipman et al. (2010). Like tsBART, tsBCF has easily-tuned hyperparameters to control degree of smoothness, with default settings yielding excellent performance (Starling et al., 2020).

Our analysis of the early medical abortion data using the tsBCF model answers key clinical questions which were not addressed satisfactorily in previous research. We validate previous findings (Lohr et al., 2018) that on average, we do not see evidence supporting decrease in efficacy
as gestational age increases. We do note some decrease in the 7–9 week range (from 0.972 at 6 weeks, to 0.955 at 7 weeks, to 0.948 at 9 weeks), which may be attributable to small sample sizes at later gestations. We identify a drop in efficacy at later gestations for patients age 29 and older, particularly for those who have given birth previously. While relative effectiveness for these patients is still over 90% in the 7–9 week range, clinicians may wish to counsel their patients accordingly.

A limitation of our work is the relatively small set of available covariates; it is possible that unobserved confounders exist. We do not know what covariates influenced each patient’s choice in protocol [Lohr et al., 2018]. Patients received counseling on the expected differences in effectiveness and side effects based on a small BPAS pilot study; aside from clinicians’ use of a common comparison chart in a printed patient guide, counseling is not standardized. However, our propensity score estimates reflect the similar safety profiles of both protocols, which share
Table 3: Simulation results averaged across replicates for each scenario and model. RMSE is the average root mean squared error for estimating heterogeneous $\tau$; nominal coverage is 95%, and interval length is for posterior credible interval of the relative risk estimates for each unit. TsBCF and BCF perform comparably in RMSE, coverage, and interval length across scenarios. BART performs nearly as well but over-covers.

| Scenario | Model         | RMSE | SD(RMSE) | Coverage | Interval Length |
|----------|---------------|------|----------|----------|-----------------|
| A        | tsBCF         | 0.129| 0.042    | 0.967    | 0.614           |
|          | BCF           | 0.136| 0.043    | 0.960    | 0.650           |
|          | BART          | 0.139| 0.053    | 0.998    | 0.894           |
|          | Causal Forests| 0.165| 0.030    | 0.517    | 0.256           |
| B        | tsBCF         | 0.186| 0.050    | 0.940    | 0.739           |
|          | BCF           | 0.194| 0.055    | 0.945    | 0.806           |
|          | BART          | 0.186| 0.046    | 0.983    | 0.907           |
|          | Causal Forests| 0.278| 0.034    | 0.242    | 0.257           |
| C        | tsBCF         | 0.198| 0.042    | 0.949    | 0.817           |
|          | BCF           | 0.207| 0.042    | 0.949    | 0.889           |
|          | BART          | 0.204| 0.042    | 0.976    | 0.932           |
|          | Causal Forests| 0.302| 0.031    | 0.262    | 0.264           |
| D        | tsBCF         | 0.185| 0.047    | 0.930    | 0.702           |
|          | BCF           | 0.187| 0.047    | 0.945    | 0.765           |
|          | BART          | 0.187| 0.051    | 0.986    | 0.913           |
|          | Causal Forests| 0.214| 0.032    | 0.442    | 0.258           |
| E        | tsBCF         | 0.089| 0.050    | 0.992    | 0.545           |
|          | BCF           | 0.094| 0.050    | 0.990    | 0.568           |
|          | BART          | 0.133| 0.057    | 0.999    | 0.888           |
|          | Causal Forests| 0.085| 0.030    | 0.843    | 0.260           |

low rates of significant adverse effects, as well as the simultaneous protocol’s reduced burden of care. Nonetheless, our work presents a clearer and more nuanced picture of the relative efficacy of simultaneous versus interval protocols for administration of mifepristone and misoprostol than previously available.
REFERENCES

J. H. Albert and S. Chib. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

S. Cameron, A. Glasier, A. Johnston, H. Dewart, and A. Campbell. Can women determine the success of early medical termination of pregnancy themselves? *Contraception*, 91:6–11, 2015.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 03 2010.

M. Creinin and M. Chen. Medical abortion reporting of efficacy: the mare guidelines. *Contraception*, 94:97–103, 2016.

A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

A. Goel, S. Mittal, B. K. Taneja, N. Singal, and S. Attri. Simultaneous administration of mifepristone and misoprostol for early termination of pregnancy: a randomized controlled trial. 283(6):1409. ISSN 1432-0711. doi: 10.1007/s00404-011-1881-2. URL https://doi.org/10.1007/s00404-011-1881-2

R. B. Gramacy and H. K. Lee. Bayesian treeed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

P. R. Hahn, C. M. Carvalho, D. Puelz, and J. He. Regularization and Confounding in Linear Regression for Treatment Effect Estimation. *Bayesian Analysis*, 13(1):163–182, 03 2018.

P. R. Hahn, J. S. Murray, and C. M. Carvalho. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 2020.

G. Heffer. Coronavirus: Government relaxes abortion rules during COVID-19 crisis. *Sky News*, 2020. URL https://news.sky.com/story/coronavirus-government-relaxes-abortion-rules-during-covid-19-crisis-11965811

B. Hernández, A. Raftery, and e. a. Pennington, S.R. Bayesian Additive Regression Trees using Bayesian model averaging. *Stat Comput*, 28:869, 2018.

J. L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

G. Imbens and D. Rubin. Causal inference in statistics, social, and biomedical sciences. *Cambridge University Press*, 2015.

Y.-T. Li, F.-M. Chen, T.-H. Chen, S.-C. Li, M.-L. Chen, and T.-C. Kuo. Concurrent use of mifepristone and misoprostol for early medical abortion. 45(4):325–328. ISSN 10284559. doi: 10.1016/S1028-4559(09)60252-7. URL https://linkinghub.elsevier.com/retrieve/pii/S1028455909602527

A. Linero and Y. Yang. Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity. *arXiv:1707.09461v1 [stat.ME]*, 2017.

A. R. Linero. Bayesian Regression Trees for High-Dimensional Prediction and Variable Selection. *Journal of the American Statistical Association*, 2018.

B. R. Logan, R. Sparapani, R. E. McCulloch, and P. W. Laud. Subgroup finding via Bayesian additive regression trees. *Statistical Methods in Medical Research*, 0(0):1–15, 2017.

P. A. Lohr, J. E. Starling, J. G. Scott, and A. R. Aiken. Simultaneous compared with interval medical abortion regimens where home use is restricted. *Obstetrics and Gynecology*, 2018.
D. McFadden. Conditional logit analysis of qualitative choice behavior. *P. Zarembka (ed.), Frontiers in Econometrics. Academic Press.*, pages 105–142, 1974.

J. S. Murray. Log-Linear Bayesian Additive Regression Trees for Categorical and Count Responses. *arXiv preprint arXiv:1701.01503*, 2017.

M. T. Pratola, H. A. Chipman, J. R. Gattiker, D. M. Higdon, R. McCulloch, and W. N. Rust. Parallel Bayesian Additive Regression Trees. *Journal of computational and graphical statistics: a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 23(3):830–852, 2014. ISSN 1061-8600.

P. Rosenbaum and D. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983.

Royal College Of Obstetricians And Gynaecologists. The care of women requesting induced abortion. *London (UK): RCOG*, 2011.

S. Sivaganesan, P. Muller, and B. Huang. Subgroup finding via Bayesian additive regression trees. *Statistics in Medicine*, 36:2391–2403, 2017.

Sparapani, Logan, MucCulloch, and Laud. Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*, 35(16):2741–2753, 2016.

J. Starling, J. Murray, C. Carvalho, R. Bukowski, and J. Scott. Bart with targeted smoothing: An analysis of patient-specific stillbirth risk. *Annals of Applied Statistics*, 2020.

S. Woody, C. Carvalho, and J. Murray. Model interpretation through lower-dimensional posterior summarization. *arXiv:1905.07103v3 [stat.ME]*, 2019.
A Appendix

A.1 Fitting the tsBCF model using data augmentation.

Here we provide details on the model parameterization and prior used for fitting tsBCF. We rewrite (4) using a redundant multiplicative parameterization (Gelman, 2006).

\[ f(t_i, x_i, z_i) = \alpha_t + \xi \mu(t_i, x_i, \hat{\pi}_i) + [b_1 z_i + b_0 (1 - z_i)] \tau(t_i, x_i) \]

where the treatment effect on the latent probit scale is now \( (b_1 - b_0) \tau(t_i, x_i) \), and the \( \alpha_t \) act as target-specific offsets which we estimate directly from the data. This expanded model ensures that the treatment effect estimation is invariant to the treatment coding; Hahn et al. (2020) notes that the original treatment effect parameterization can yield different posterior inferences depending on the choice of values for treatment and control.

Priors are as follow, with hyperparameters \( \{\lambda, s_\mu, s_\tau, l_\mu, l_\tau\} \) set by the user. We set \( \nu_\mu = 1 \) to induce the half-Cauchy prior, and \( \nu_\tau \) is set to induce the half-Normal prior.

\[ \xi \sim N(0, 1) \]
\[ b_1 \sim N(.5, .5) \text{ and } b_0 \sim N(-.5, .5) \rightarrow b = (b_1 - b_0) \sim N(0, 1) \]
\[ \mu(t_i, x_i, \hat{\pi}_i) \sim tsBART \text{ with } n_\mu = 200 \text{ trees, terminal nodes } m_\mu, \text{ and covariance } C_\mu: \]
\[ m_\mu \sim Gp(0, C_\mu) \]
\[ C_\mu = \frac{s_\mu^2}{n_\mu \Delta_\mu} \exp \left[ -0.5 \left( \frac{t - t'}{l_\mu} \right)^2 \right] \]
\[ \frac{1}{\Delta_\mu} \sim IG \left( \frac{\nu_\mu}{2}, \frac{\nu_\mu}{2} \right) \]
\[ \tau(t_i, x_i) \sim tsBART \text{ with } n_\tau = 50 \text{ trees, terminal nodes } m_\tau, \text{ and covariance } C_\tau: \]
\[ m_\tau \sim Gp(0, C_\tau) \]
\[ C_\tau = \frac{s_\tau^2}{n_\tau \Delta_\tau} \exp \left[ -0.5 \left( \frac{t - t'}{l_\tau} \right)^2 \right] \]
\[ \frac{1}{\Delta_\tau} \sim IG \left( \frac{\nu_\tau}{2}, \frac{\nu_\tau}{2} \right) \]

We constrain \( \sigma^2 = 1 \) for identifiability. For a continuous response, the prior for \( \sigma^2 \) follows Chipman et al.’s recommendation for a rough over-estimation of \( \hat{\sigma} \). We choose \( \nu = 3 \) and \( q = 0.90 \), and estimate \( \hat{\sigma} \) by regressing \( y \) onto \( x \) (including the target variable as a covariate), then choose \( \lambda \) s.t. the \( q \)th quantile of the prior is located at \( \hat{\sigma} \), i.e. \( P(\sigma \leq \hat{\sigma}) = q \). Length-scale parameters \( (l_\mu, l_\tau) \) can be stated in terms of expected wiggliness of the tsBART fits, as described in Starling et al. (2020).

A.2 Bayesian Backfitting Algorithm

We leverage the Bayesian backfitting MCMC algorithms of tsBART and BCF to design a Bayesian backfitting algorithm for tsBCF. We refer interested readers to Chipman et al. (2010) for a full discussion of the original Bayesian backfitting, and Starling et al. (2020) and Hahn et al. (2020) for tsBART and BCF algorithms respectively. Briefly, Bayesian backfitting involves an MCMC algorithm where each tree, and its parameters are sampled one at a time given the partial residuals.
from the other \( m - 1 \) trees. One iteration of the sampler consists of looping through the trees, sampling each tree \( T_j \) via a Metropolis step, and then sampling its associated leaf parameters \( M_j \), conditional on \( \sigma^2 \) and the remaining trees and leaf parameters. After a pass through all trees, \( \xi, \Delta_\mu, b_1, b_0, \Delta_\tau \), and \( \sigma^2 \) are updated in a Gibbs step.

### A.2.1 Updating trees and leaves

In general, to sample \( \{T_j, M_j\} \) conditioned on the other trees and leaf parameters \( \{T_k, M_k\} \), define the partial residual as

\[
    r_{ij} = y_i - \sum_{k=1, k \neq j}^m g(x_i; T_k, M_k).
\]

More specifically in the tsBCF setting, for \( \mu(t_i, x_i, \hat{\pi}_i) \):

- The “data” is \( y_i - \alpha_t - \left[ b_1 z_i + b_0 (1 - z_i) \right] \tau(t_i, x_i) \)
- The variance is \( \xi^2 \)
- The partial residuals are \( r_{ij} = \text{data}_i - \sum_{k=1, k \neq j}^{200} g(x_k; T_\mu k, M_\mu k) \).

Similarly, for \( \tau(t_i, x_i) \):

- The “data” is \( \frac{y_i - \alpha_t - \xi \mu(t_i, x_i, \hat{\pi}_i)}{[b_1 z_i + b_0 (1 - z_i)]} \)
- The variance is \( \frac{\sigma^2}{[b_1 z_i + b_0 (1 - z_i)]^2} \)
- The partial residuals are \( r_{ij} = \text{data}_i - \sum_{k=1, k \neq j}^{50} g(x_k; T_\tau k, M_\tau k) \).

Using \( r_j \) as the working response vector, at step \( s \) of the MCMC one samples \( T_j^{(s)} \) by proposing one of four local changes to \( T_j^{(s-1)} \), marginalizing analytically over \( M_j \). The local change is selected randomly from the following candidates:

- **grow** randomly selects a terminal node and splits it into two child nodes
- **prune** randomly selects an internal node with two children and no grandchildren, and prunes the children, making the selected node a leaf

In line with common practice, we implement only **prune** and **grow** proposals due to computational cost (Pratola et al., 2014). Once the move in tree space is either accepted or rejected, \( M_j \) is sampled from its full conditional given \( T_j \) and \( \sigma^2 \).

### A.2.2 Full conditionals

The posterior conditional distributions for the Bayesian backfitting algorithm are as follows. For simplicity we assume that target covariate values \( t \) are on a common discrete grid, though this is not a requirement.
For updating \((\sigma^2|\ldots)\),

\[
p(\sigma^2) \sim IG \left( \frac{\nu}{2}, \frac{\nu \lambda}{2} \right)
\]

\[
p(y|\sigma^2) = \prod_{i=1}^{n} N \left( y_i | \hat{y}_i, \sigma^2 \right), \text{ where } \hat{y}_it = \alpha_t + \xi \mu(t_i, x_i, \pi_i) + [b_1 z_i + b_0 (1 - z_i)] \tau(t_i, x_i)
\]

\[
p(\sigma|\ldots) \propto \left( p(\sigma^2) \cdot p(y|\sigma^2) \right) \sim IG \left( \frac{\nu + n}{2}, \frac{RSS + \nu \lambda}{2} \right), \text{ where } RSS = \sum_{i=1}^{n} (y_{it} - \hat{y}_{it})
\]

For updating \((\xi|\ldots)\),

\[
\xi \sim N(0, 1)
\]

\[
p(r_i|\ldots) = \prod_{i=1}^{n} N \left( y_i - \alpha_t - [b_1 z_i + b_0 (1 - z_i)] \tau(t_i, x_i), \frac{\sigma^2}{\mu(t_i, x_i, \pi_i)^2} \right)
\]

\[
p(\xi|\ldots) \propto \left( p(\xi) \cdot p(r_i|\xi) \right) \sim N(\mu^*, v^{2*}), \text{ where }
\]

\[
v^{2*} = \left( 1 + \frac{1}{\sigma^2} \sum_{i=1}^{n} \mu(t_i, x_i, \pi_i)^2 \right)^{-1}
\]

\[
\mu^* = v^{2*} \left( \sum_{i=1}^{n} \mu(t_i, x_i, \pi_i)^2 r_i \right)
\]

For updating \((\Delta \mu|\ldots)\), let \(n_{bots}\) be the number of leaves \(l\) across all \(n_{\mu}\) control trees, and \(T\) be the length of the target covariate mesh. Let \(j\) index trees and \(l\) index leaves. The full conditional only uses control tree fits.

\[
\Delta \mu \sim Ga \left( \frac{\nu_{\mu}}{2}, \frac{\nu_{\mu}}{2} \right)
\]

\[
p(m_{jl}|\ldots) = \prod_{j,l} N_T \left( m_{jl} | 0, \frac{s^2_{\Delta \mu n_{\mu}}}{\Delta \mu n_{\mu}} \right), \text{ where } C_0 = \left[ -\frac{1}{2} \left( \frac{t - t'}{l_{\mu}} \right)^2 \right]
\]

\[
p \left( \frac{1}{\Delta \mu} \big| \ldots \right) \propto \left( p(\Delta \mu) \cdot p(m_{jl} | \Delta \mu) \right) \sim IG \left( \frac{\nu_{\mu} + n_{bots}T}{2}, \frac{\nu_{\mu} + SSQ}{2} \right), \text{ where } SSQ = \sum_{j,l} m_{jl}^T C_{\mu^{-1}} m_{jl}
\]

For updating \((\Delta \tau|\ldots)\), let \(n_{bots}\) be the number of leaves \(l\) across all \(n_{\tau}\) treatment trees, and \(T\) be the length of the target covariate mesh. Let \(j\) index trees, \(l\) index leaves. This full conditional only
uses treatment tree fits.

\[ \Delta_{T} \sim Ga \left( \frac{\nu_{T}}{2}, \frac{\nu_{T}}{2} \right) \]

\[ p(m_{jl} | \ldots) = \prod_{j,l} N_{T} \left( m_{jl} \mid 0, \frac{s_{\tau}^{2}}{\Delta_{T} \lambda_{T} | C_{0}} \right) , \text{where } C_{0} = \left[ -0.5 \left( \frac{t - t'}{l_{T}} \right)^{2} \right] \]

\[ p \left( \frac{1}{\Delta_{T}} \mid \ldots \right) \propto \left( p(\Delta_{T}) \cdot p(m_{jl} \mid \Delta_{T}) \right) \sim IG \left( \frac{\nu_{T} + n_{bots} T}{2}, \frac{\nu_{T} + SSQ}{2} \right) , \text{where } SSQ = \sum_{j,l} m_{jl}^{T} C_{\tau}^{-1} m_{jl} \]

For updating \((b_{1} \mid \ldots)\), let \(n_{z}\) be the number of treatment observations. This full conditional uses only treatment observations.

\[ b_{1} \sim N(\mu_{b_{1}} = .5, \sigma_{b_{1}}^{2} = .5) \]

\[ p(r_{i} \mid \ldots) = \prod_{i=1}^{n_{z}} N \left( y_{i} - \alpha_{i} - \xi \mu(t_{i}, x_{i}, \tilde{\pi}_{i}) \bigg| r_{i}, \frac{\sigma^{2}}{\tau(t_{i}, x_{i})^{2}} \right) \]

\[ p(b_{1} \mid \ldots) \propto \left( p(b_{1}) \cdot p(r_{i} \mid b_{1}) \right) \sim N(\mu^{*}, v^{*}) , \text{where} \]

\[ v^{*} = \left( \frac{1}{\sigma_{b_{1}}^{2}} + \frac{1}{\sigma^{2}} \sum_{i=1}^{n_{z}} \tau(t_{i}, x_{i})^{2} \right)^{-1} \]

\[ \mu^{*} = v^{*} \left( \frac{\mu_{b_{1}}}{\sigma_{b_{1}}^{2}} + \frac{1}{\sigma^{2}} \sum_{i=1}^{n_{z}} \tau(t_{i}, x_{i})^{2} r_{i} \right) \]

For updating \((b_{0} \mid \ldots)\), let \(n_{z}\) is the number of control observations. This full conditional uses only control observations.

\[ b_{0} \sim N(\mu_{b_{0}} = -.5, \sigma_{b_{0}}^{2} = .5) \]

\[ p(r_{i} \mid \ldots) = \prod_{i=1}^{n_{z}} N \left( y_{i} - \alpha_{i} - \xi \mu(t_{i}, x_{i}, \tilde{\pi}_{i}) \bigg| r_{i}, \frac{\sigma^{2}}{\tau(t_{i}, x_{i})^{2}} \right) \]

\[ p(b_{0} \mid \ldots) \propto \left( p(b_{0}) \cdot p(r_{i} \mid b_{0}) \right) \sim N(\mu^{*}, v^{*}) , \text{where} \]

\[ v^{*} = \left( \frac{1}{\sigma_{b_{0}}^{2}} + \frac{1}{\sigma^{2}} \sum_{i=1}^{n_{z}} \tau(t_{i}, x_{i})^{2} \right)^{-1} \]

\[ \mu^{*} = v^{*} \left( \frac{\mu_{b_{0}}}{\sigma_{b_{0}}^{2}} + \frac{1}{\sigma^{2}} \sum_{i=1}^{n_{z}} \tau(t_{i}, x_{i})^{2} r_{i} \right) \]
A.2.3 Marginal likelihood for prognostic tree updates

The marginal likelihood for updating the prognostic tree fits $\mu(t_i, x_i, \hat{\pi}_i)$ is the version from BART with Targeted Smoothing, with homogeneous variances. Here, we derive the marginal log-likelihood here for a single leaf. In the backfitting algorithm, this is calculated for multiple leaves depending on whether a birth move or death move is proposed for the tree. The likelihoods are then used in calculating the acceptance probability for the Metropolis step.

Let $y_i$ represent the length $n_i$ vector of residuals for a given leaf. Let $T_{\mu}$ be the tree structure for the $j^{th}$ tree, and $t_{\text{len}}$ be the length of the grid of unique target values. We integrate out leaf means vector $m_l$ to obtain the marginal log-likelihood as follows.

$$p(y_l|T_j, \sigma^2) = \int_{\mathbb{R}} N_{n_l} (y_l|W_l m_l, \sigma^2 I) \cdot N_{t_{\text{len}}} (m_l|m_0, K^{-1}) \, dm_l$$

where $W_l$ is a $n \times t_{\text{len}}$ matrix, with one row for each observation; all entries are zero, except a 1 in the column corresponding to each observation $i$'s associated time $t_i$. Set $m_0 = 0$. The marginal log-likelihood is then

$$p(y_l|T_{\mu_j}, \sigma^2) = -\frac{n_l}{2} \log (2\pi \sigma^2) + \frac{1}{2} \log (|K|) - \frac{1}{2} \log (|C|) - \frac{1}{2} \left[ \frac{1}{\sigma^2} y_l^T y_l + m_0^T K m_0 - b^T C^{-1} b \right]$$

where $C = \left( \frac{1}{\sigma^2} W_l^T W_l + K \right)$ and $b = \left( \frac{1}{\sigma^2} W_l^T y_l + K m_0 \right)$. For computational purposes, $W_l^T W_l = \begin{bmatrix} n_1 \ldots n_{\text{max}} \end{bmatrix}$, the vector of sample sizes for each time. In addition, $y_l y_l = \sum_{i=1}^{n_l} y_i$, the sum of all $y_i$ in leaf $l$.

A.2.4 Marginal likelihood for treatment tree updates

The marginal likelihood for updating treatment fits is slightly more complex, requiring heterogeneous variances, since the variance for updating $\tau(t_i, x_i)$ is $\left( \frac{\sigma^2}{b_i z_i + b_0 (1-z_i)} \right)^2$. Let $y_l$ represent the length $n_i$ vector of residuals for a given leaf. Let $T_{T_j}$ be the tree structure for the $j^{th}$ tree. We integrate out leaf means vector $m_l$ to obtain the marginal log-likelihood as follows. Instead of $\sigma^2 I = (\omega I)^{-1}$, use the $(n_l \times n_l)$ precision matrix $\Lambda = \text{diag} [\omega_1, \ldots, \omega_{n_l}]$.

We integrate out leaf means vector $m_l$ to obtain the marginal log-likelihood as follows.

$$p(y_l|T_{T_j}, \sigma^2) = \int_{\mathbb{R}} N_{n_l} (y_l|W_l m_l, \Lambda^{-1}) \cdot N_{t_{\text{len}}} (m_l|m_0, K^{-1}) \, dm_l$$

where $W_l$ is gain a $n \times t_{\text{len}}$ matrix, with one row for each observation; all entries are zero, except a 1 in the column corresponding to each observations $i$'s associated time $t_i$. We again let $m_0 = 0$. The marginal log-likelihood is then

$$p(y_l|T_{T_j}, \sigma^2) = -\frac{n_l}{2} \log (2\pi) + \frac{1}{2} \left[ \log (|\Lambda|) + \log (|K|) - \log (|C|) - y_l^T \Lambda y_l - b^T C^{-1} b \right]$$

where $C = (W_l^T \Lambda W_l + K)$ and $b = (W_l^T \Lambda y_l + K m_0)$. For computational purposes, $W_l^T \Lambda W_l$ is the the $t_{\text{len}} \times t_{\text{len}}$ diagonal matrix of sums of precisions for each sample size. $W_l^T \Lambda y_l$ is the vector of $\omega_i y_i$ sums for each time point, and $y_l^T \Lambda y_l = \sum_{i=1}^{n_l} \omega_i y_i^2$. Finally, $\log (|\Lambda|) = \sum_{i=1}^{n_l} \log (\omega_i)$.
A.3 Selecting prior scales for marginal tree variances.

When modeling a binary response, care must be taken in selecting sensible scale hyperparameters $\mu$ and $\tau$. In Section 3, we propose an intuitive method for setting these hyperparameters based on a combination of prior elicitation and data-driven specification. Here, we define structural heterogeneity, provide intuition, make suggestions for setting prior scales, and provide detail to support our finding that approximately 0.02% of the variance we find in relative risk is comprises structural heterogeneity.

Defining structural heterogeneity. In the scalar response setting, shrinking towards a constant $\tau$ across all observations corresponds to a homogeneous treatment effect. In the probit case, our causal estimand of interest is relative risk, a non-linear transformation which includes both $\mu(t_i, x_i, \hat{\pi}_i)$ and $\tau(t_i, x_i)$. Let $b = \Phi(\alpha + \mu)$ represent baseline risk, and $r = \frac{\Phi(\alpha + \mu + \tau)}{\Phi(\alpha + \mu)}$ relative risk. If $\tau = 0$ for all observations, relative risk is of course constant for all observations. A fixed non-zero $\tau$ for all observations does not imply homogeneous relative risk, as relative risk will still vary as $\mu(t_i, x_i, \hat{\pi}_i)$ varies. We call this variability structural heterogeneity, and the degree to which it exists depends on the size of $\mu$ and $\tau$, and the size of the model offset $\alpha$.

Structural heterogeneity intuition. Before considering our early medical abortion setting, we use a very general example to briefly illustrate how structural heterogeneity arises due to the non-linear transformations involved in calculating relative risk, and how the amount of heterogeneity depends on various factors. Figure 11 provides intuition about the behavior of structural heterogeneity. Each row represents a different offset ($\alpha$), corresponding to average baseline risks of 0.80, 0.90, 0.93, and 0.95; columns show three different $\tau_i$ values, $-0.5$, $-0.313$, and $-0.1$. For each of the $(\alpha, \tau_i)$ combinations, we draw 1,000 $\mu_i$ from four Gaussian distributions with increasing variance and calculate the individual relative risks $RR_i = \frac{\Phi(\alpha + \mu_i + \tau_i)}{\Phi(\alpha + \mu_i)}$. Amount of structural heterogeneity depends on the size of the offset, the magnitude of $\tau_i$, and the variability in $\mu_i$; structural heterogeneity generally increases as $\tau_i$ increases in magnitude and as $\mu_i$ is more disperse, and generally decreases as baseline risk increases. The violin plot highlighted in grey, where $\alpha=1.48$, $\tau_i=-0.313$, and $\mu_i \sim N(0, 0.3)$ corresponds most closely to our early medical abortion analysis; this combination corresponds to the average offset, the posterior mean $\tau$ value, and the standard error of the posterior $\mu_i$ estimates from our analysis.

Setting prior scales. We suggest setting prior scales $s_\mu$ based on prior elicitation of a plausible range of baseline risks, and $s_\tau$ using estimation of baseline risk and relative risk from a small amount of held-out data. We select a range of 0.86 to 0.999 as a plausible range of relative risks, and estimate that the range of plausible values covers about 3.3 times the standard deviation, such that

$$s_\mu = \frac{\Phi^{-1}(0.999) - \Phi^{-1}(0.860)}{3.3}.$$ 

We then hold out 500 observations, obtain point estimates for relative risk and baseline risk, and use Nelder Mead to solve for a reasonable $s_\tau$ value.

Quantifying structural heterogeneity using posterior draws. We estimate the degree to which structural heterogeneity is present in our early medical abortion analysis as follows. We set prior scales $s_\mu$, $s_\tau$ as described above. For each posterior draw $b$, we calculate the variance in individual
relative risk as

\[ \hat{\sigma}_{RR}^{(b)} = \text{Var} \left( \hat{R}_R^{(b)} \right) \]  

(13)

where \( R_{R_i}^{(b)} \) is estimated using posterior draws \( \mu_{i}^{(b)}, \tau_{i}^{(b)} \).

Let \( \bar{\tau}^{(b)} = \sum_{i=1}^{n} \tau_{i}^{(b)} \) represent the mean of \( \tau \) across all observations for draw \( b \). We now let \( \hat{\sigma}_{RR,het}^{(b)} \) represent a second estimate of the variance in \( R_{R_i}^{(b)} \) for draw \( b \), calculated as in (13) but replacing posterior draws \( \tau_{i}^{(b)} \) with heterogeneous \( \bar{\tau}^{(b)} \).

We estimate how much more variability is present in relative risk under heterogeneous \( \tau_i \) compared to homogeneous \( \tau \) by computing the ratio of \( \hat{\sigma}_{RR}^{(b)} \) to \( \hat{\sigma}_{RR,het}^{(b)} \) for each draw and averaging across draws. We find that there is an average of 2.06 times more variability in posterior relative risk under heterogeneous versus homogeneous \( \tau_i \), giving us confidence that our clinical findings are not just an artifact of structural heterogeneity.
Figure 11: Illustrates structural heterogeneity in relative risk for six combinations of offset $\alpha$ and treatment effect $\tau_i$, with $\mu_i$ coming from distributions with different levels of variability. Amount of structural heterogeneity depends on the size of the offset, the magnitude of $\tau_i$, and the variability in $\mu_i$; structural heterogeneity generally increases as $\tau_i$ increases in magnitude and as $\mu_i$ is more disperse, and generally decreases as baseline risk increases. The violin plot highlighted in grey ($\alpha=1.48$, $\tau_i=-0.313$, and $\mu_i \sim N(0, 0.3)$) corresponds most closely to our early medical abortion analysis; this combination corresponds to our average offset, posterior mean $\tau$, and standard error of the posterior $\mu_i$ estimates from our analysis.