*The Annals of Applied Statistics*
2020, Vol. 14, No. 1, 28–50
https://doi.org/10.1214/19-AOAS1268


# BART WITH TARGETED SMOOTHING: AN ANALYSIS OF PATIENT-SPECIFIC STILLBIRTH RISK

BY JENNIFER E. STARLING[1,*], JARED S. MURRAY[2,†], CARLOS M. CARVALHO[2,‡], RADEK K. BUKOWSKI[3] AND JAMES G. SCOTT[1,**]

[1]*Department of Statistics and Data Sciences, University of Texas at Austin,* [*]*jstarling@utexas.edu;* [**]*james.scott@mccombs.utexas.edu*

[2]*McCombs School of Business, University of Texas at Austin,* [†]*jared.murray@mccombs.utexas.edu;* [‡]*carlos.carvalho@mccombs.utexas.edu*

[3]*Department of Women's Health, Dell Medical School, University of Texas at Austin, radek.bukowski@austin.utexas.edu*

This article introduces BART with Targeted Smoothing, or tsBART, a new Bayesian tree-based model for nonparametric regression. The goal of tsBART is to introduce smoothness over a single target covariate $t$ while not necessarily requiring smoothness over other covariates $x$. tsBART is based on the Bayesian Additive Regression Trees (BART) model, an ensemble of regression trees. tsBART extends BART by parameterizing each tree's terminal nodes with smooth functions of $t$ rather than independent scalars. Like BART, tsBART captures complex nonlinear relationships and interactions among the predictors. But unlike BART, tsBART guarantees that the response surface will be smooth in the target covariate. This improves interpretability and helps to regularize the estimate.

After introducing and benchmarking the tsBART model, we apply it to our motivating example—pregnancy outcomes data from the National Center for Health Statistics. Our aim is to provide patient-specific estimates of stillbirth risk across gestational age ($t$) and based on maternal and fetal risk factors ($x$). Obstetricians expect stillbirth risk to vary smoothly over gestational age but not necessarily over other covariates, and tsBART has been designed precisely to reflect this structural knowledge. The results of our analysis show the clear superiority of the tsBART model for quantifying stillbirth risk, thereby providing patients and doctors with better information for managing the risk of fetal mortality. All methods described here are implemented in the R package *tsbart*.

**1. Introduction.** An ongoing research challenge in obstetrics is to quantify the risk of stillbirth, defined as fetal death after 20 weeks of gestation. Stillbirth is a major public-health problem with 23,595 reported cases in the U.S. in 2013 alone (MacDorman and Gregory (2015)). Stillbirth is less well understood than other adverse pregnancy outcomes, and stillbirth rates have remained largely unchanged, even as many other serious adverse pregnancy outcomes (e.g., neonatal death) have become rarer. Providing better estimates of stillbirth risk as gestational age advances can yield important insights for obstetricians and patients. If an obstetrician knew, for example, that a patient's stillbirth risk was likely to rise earlier in pregnancy than usual or was likely to rise to higher than normal levels at later gestational ages, then proactive steps could be taken to manage that risk, especially in pregnancy at term. Conservative steps might entail increased monitoring and more frequent prenatal clinic visits, while a more aggressive step might involve an elective Cesarean section or early induction of labor.

Received January 2019; revised May 2019.
*Key words and phrases.* Bayesian additive regression tree, ensemble method, Gaussian process, regression tree, regularization.

Statistically speaking, we can think of stillbirth risk as a regression function $h(t, x)$ representing the conditional probability[1] of stillbirth at gestational age $t$, given that the fetus survived in utero until just before $t$, and given other characteristics $x$ of the maternal-fetal dyad. Thus, the fundamental biomedical problem we address in this paper is to provide better patient-specific estimates of $h(t, x)$. This fills an important knowledge gap, since the current obstetrics literature does not provide an especially nuanced characterization of this function. In particular, the way that $h(t, x)$ depends upon maternal-fetal characteristics is not well understood. Structurally, obstetricians do expect that stillbirth risk evolves smoothly, without sudden jumps or discontinuities, as gestational age ($t$) advances; however, they do not have strong prior knowledge about how it should change with other maternal-fetal characteristics ($x$).

The central argument of our paper is that this situation calls for nonparametric regression with *targeted smoothing* in gestational age $t$, that is, we require that $h(t, x)$ be smooth with respect to $t$ (the target covariate), but we remain agnostic about smoothness with respect to $x$. This approach realizes two complementary advantages when quantifying stillbirth risk. First, from a clinical perspective targeted smoothing reflects prior knowledge, aids interpretability and assists doctors in communicating stillbirth risks to patients as clearly as possible. For example, smoothing helps prevent doctors and patients alike from overinterpreting the small jumps or wiggles in $h(t, x)$ that arise in a completely nonparametric estimate but that are likely just noise. Second, from a statistical perspective targeted smoothing can reduce variance without inflating bias.

To incorporate these benefits into our analysis of stillbirth risk, we propose a Bayesian approach called BART with Targeted Smoothing, or tsBART, which is based on the highly successful Bayesian Additive Regression Trees (BART) model introduced by Chipman, George and McCulloch (2010). The original BART model is a Bayesian ensemble-of-trees approach to nonparametric regression. It predicts a scalar response $y$ using a sum of many binary regression trees, where each tree is encouraged by a prior to be a "weak learner," that is, to have relatively few splits and to use only a small set of the available predictors. BART with Targeted Smoothing is similar in this regard, and we use the same prior over tree space proposed in the original BART paper. Where tsBART differs is in the prior used for the terminal nodes of each tree. BART specifies a Gaussian prior for the scalar mean parameters in each terminal node. tsBART replaces the Gaussian prior with a Gaussian process prior over univariate functions in the "target" covariate $t$, so that each terminal node is parameterized by a smooth function of $t$.

Thus to summarize our contributions:

1. We introduce the tsBART model and demonstrate its advantages for problems where targeted smoothing is desirable.
2. We apply this method to data on birth records from the National Center for Health Statistics in order to produce accurate estimates for $h(t, x)$ and to provide clinicians with more granular knowledge of patient-specific stillbirth risk.

It would certainly be possible to estimate stillbirth risk using existing techniques for modeling time-to-event data (see, e.g., Mandujano, Waters and Myers (2013)). Thus, a major focus of our paper is to demonstrate that the specific features we had in mind when designing tsBART—targeted smoothing in gestational age, while avoiding strong assumptions in other covariates—have some very real advantages for this kind of problem. Available techniques either lack smoothness entirely (and thus tend to have smaller bias) or enforce smoothness globally (and thus tend to have smaller variance). Each approach has its advantages, but

[1] Or, in continuous time, the hazard rate.

tsBART enjoys the best of both worlds for quantifying stillbirth risk. It easily handles complex interactions and nonlinear effects, maintains computational tractability, and offers a full picture of posterior uncertainty, all while maintaining smoothness in $t$.

Moreover, while our motivating example involves estimating a smooth hazard function, the tsBART model is much more general than this. The same approach can work in any nonparametric regression problem where targeted smoothing is desired a priori, regardless of whether the response is continuous, binary, or (as in our case) a time-to-event outcome. Across a series of benchmarking examples, we show that our approach to targeted smoothing can lead to a favorable bias-variance tradeoff vs. both classes of competing methods—those that make global smoothness assumptions, and those that make no smoothness assumptions. Our simulation studies also bear out another considerable advantage. When the targeted smoothing assumption is valid, tsBART tends to yield superior frequentist coverage vs. plausible alternative methods.

The paper proceeds as follows. Section 2 provides an overview of the stillbirth risk-curve modeling problem and dataset. Section 3 details the tsBART model and reviews the relevant literature. Section 4 presents the results of simulation studies showing the advantages of the method. Section 5 then presents our core scientific contribution, an analysis of stillbirth risk using the tsBART model. Section 6 concludes with a brief discussion. Further details, including on computational methods, are in the Appendices.

All methods described in this paper are implemented in the R package *tsbart*; see the Supplementary Material (Starling et al. (2020)).[2]

## 2. Stillbirth risk.

2.1. *Background.* Stillbirth is a significant public health concern that affects tens of thousands of Americans each year. In the U.S. in 2013, a total of 23,595 stillbirths were reported (MacDorman and Gregory (2015)). The National Vital Statistics System notes that stillbirth has been significantly overlooked in public-health research and obstetrics guidance, and its mechanisms are not well understood. Obstetricians do know that the risk of stillbirth typically (but not universally) cumulatively increases with time in utero. But this risk must be balanced against the potential negative consequences of early delivery. Preterm and early term births are associated with increased risk of neonatal mortality and morbidity, adverse neurodevelopmental and cognitive outcomes and increased healthcare costs (e.g., Muraskas and Parsi (2008), Kornhauser and Schneiderman (2010)). Obstetricians can therefore benefit greatly from access to better estimates of stillbirth risk over gestational age, so that they can give clinical advice that minimizes the overall risk of adverse perinatal outcomes. Conservative patient management might entail increased monitoring and more frequent prenatal clinic visits, while more aggressive steps include an early delivery via either elective Cesarean section or early induction of labor. From a statistical perspective, this means that accurate uncertainty quantification is vital for helping doctors understand which cases have a less precisely estimated risk profile.

Previous research on adverse perinatal outcomes has focused more heavily on neonatal death than on stillbirth (e.g., Bailit et al. (2010), Clark, Frye and Myers (2010), Reddy, Bettegowda and Dias (2011)). A more recent line of work attempts to refine these broad conclusions by seeking to model stillbirth risk based on a patient's individual risk factors. In particular, Mandujano, Waters and Myers (2013) model hazard functions for stillbirth by stratifying patients into two broad categories, low risk vs. high risk. Here "high risk" is determined by presence or absence of at least one of several preexisting maternal conditions (e.g., diabetes,

[2]https://github.com/jestarling/tsbart.

chronic hypertension, and others). The model provides two stillbirth risk curves, one each for the high-risk and low-risk groups, for a U.S. cohort. This model does not meaningfully distinguish among the individual risk factors with potentially distinct etiologies, nor does it incorporate recent evidence that many other maternal and fetal characteristics—including maternal race, plurality, birth weight and sex of the fetus—appear to correlate with stillbirth risk (Xu et al. (2013), MacDorman and Gregory (2015)). Finally, it fails to allow for the possibility of statistical interactions between risk factors. Our targeted smoothing approach is specifically designed to address these shortcomings.

2.2. *Data description*. Our analysis uses anonymized birth data from the National Center for Health Statistics from the years 2004 and 2005 (Table 1). Each medical record is associate with a single pregnancy. Each record contains the gestational age in weeks at which the pregnancy was delivered, based on calculation from the woman's last normal menstrual period or a clinical estimate. The outcome of each pregnancy is recorded as either a stillbirth or a live birth. Each record also contains information about the maternal-fetal dyad, including maternal risk factors, such as diabetes, hypertension, sociodemographic variables and fetal characteristics, such as sex or estimated fetal weight.

TABLE 1

*Cohort characteristics for our dataset on stillbirth. The "low-risk" and "high-risk" designations are not used formally in our model, but they are provided for the sake of comparison with Mandujano, Waters and Myers (2013), Table 1. Our cohort is similar in composition to the cohort analyzed there. Numbers in parentheses are percentages with respect to the given cohort*

| Characteristic | Full Cohort ($n = 4{,}553{,}868$) | Low risk ($n = 4{,}137{,}260$) | High risk ($n = 416{,}608$) |
|---|---|---|---|
| Maternal age (Yrs) | | | |
| <20 | 452,060 (9.93) | 418,953 (10.13) | 33,107 (7.95) |
| 20–29 | 2,401,223 (52.73) | 2,204,168 (53.28) | 197,055 (47.30) |
| 30–39 | 1,585,226 (34.81) | 1,415,991 (34.23) | 169,235 (40.62) |
| 40–49 | 115,020 (2.53) | 97,855 (2.37) | 17,165 (4.12) |
| 50+ | 339 (0.01) | 293 (0.01) | 46 (0.01) |
| Maternal race and ethnicity | | | |
| White, non-Hispanic | 2,757,816 (60.56) | 2,520,632 (60.93) | 237,184 (56.93) |
| Black, non-Hispanic | 693,751 (15.23) | 619,761 (14.98) | 73,990 (17.76) |
| Hispanic | 809,086 (17.77) | 736,908 (17.81) | 72,178 (17.33) |
| Other | 293,215 (6.44) | 259,959 (6.28) | 33,256 (7.98) |
| Parity | | | |
| Primiparous | 1,490,501 (32.73) | 1,370,443 (33.12) | 120,058 (28.82) |
| Multiparous | 3,063,367 (67.27) | 2,766,817 (66.88) | 296,550 (71.18) |
| Maternal risk factors | | | |
| Anemia | 115,663 (2.54) | 0 (0.00) | 115,663 (27.76) |
| Cardiac disease | 20,937 (0.46) | 0 (0.00) | 20,937 (5.03) |
| Lung disease | 63,063 (1.38) | 0 (0.00) | 63,063 (15.14) |
| Diabetes mellitus | 159,765 (3.51) | 0 (0.00) | 159,765 (38.35) |
| Hemoglobinopathy | 4260 (0.09) | 0 (0.00) | 4260 (1.02) |
| Chronic hypertension | 43,935 (0.96) | 0 (0.00) | 43,935 (10.55) |
| Renal disease | 14,210 (0.31) | 0 (0.00) | 14,210 (3.41) |
| Rh isoimmunization | 31,317 (0.69) | 0 (0.00) | 31,317 (7.52) |
| Infant sex | | | |
| Male | 2,330,557 (51.18) | 2,117,958 (51.19) | 212,599 (51.03) |
| Female | 2,223,311 (48.82) | 2,019,302 (48.81) | 204,009 (48.97) |

The dataset consists of 8,371,461 pregnancies with 7,940,495 live births, 100,072 stillbirths and 330,894 cases where stillbirth outcome is missing. We restrict our analysis to complete cases with all maternal-fetal information and stillbirth response present. Analysis is also limited to pregnancies delivered from 34 to 42 weeks inclusive, as this is the range where clinicians might plausibly recommend to deliver a baby based on elevated stillbirth risk, barring truly exceptional circumstances. These restrictions yield 4,553,868 pregnancies for analysis, of which 7,175 are stillbirths, for an overall prevalence of 1.58 stillbirths per thousand pregnancies from 34 to 42 weeks' gestation. The prevalence in the high risk category was 2.85 stillbirths per thousand, while the prevalence in the low risk group was 1.45 per thousand. Prevalence is comparable to the dataset analyzed by Mandujano, Waters and Myers (2013), where overall prevalence was 1.45 births per thousand, 2.68 in the high risk group and 1.34 in the low risk group. A full table of summary statistics for our sample is shown in Table 1. In practice, we work with a smaller case-control sample of this full dataset. This is described in Section 5; full details of the data pipeline are also available at github.com/jestarling/tsbart-analysis.

Maternal-fetal characteristics were selected for inclusion in our regression models based on clinical knowledge, availability of data and previous research findings on risk factors for stillbirth (e.g., Mandujano, Waters and Myers (2013), Muraskas and Parsi (2008), Kornhauser and Schneiderman (2010)). Maternal covariates include maternal age, primiparity, whether the labor was induced, ethnicity (White non-Hispanic, Black non-Hispanic, Hispanic, Other), aggregate pregnancy weight gain quantile, presence of diabetes mellitus, presence of chronic hypertension, and an indicator for the presence of any other risk factor. Other risk factors include anemia, cardiac disease, lung disease, hemoglobinopathy and Rh sensitization. Consistent with the analysis of Mandujano, Waters and Myers (2013), pregnancy-related complications, such as gestational diabetes, abruption or preeclampsia, were not included as risk factors. Fetal covariates include infant sex and birth weight quantile.

We did not exclude any variables on statistical grounds. One of the benefits of the BART framework, which also applies to the tsBART method, is that variable selection procedures are not generally required. As discussed in Section 3, the BART prior guides the model to choosing subsets of the most relevant covariates for inclusion in each tree.

Birth weight cannot be observed directly by a doctor contemplating whether to delivery a pregnancy early due to elevated stillbirth risk. However, birth weight quantile acts as a sensible proxy for the information doctors *would* actually have at their disposal in a prenatal visit—fetal weight quantile in utero which is estimated routinely using ultrasound and fetal growth charts. Because fetal weight quantile at later gestational ages correlates very strongly with birthweight quantile, we do not expect that there is substantial error introduced by using birth weight quantile (which we have and a doctor would not) as a proxy for fetal weight quantile in utero (which a doctor would have).

**3. BART with targeted smoothing.** We now introduce the tsBART model, which later in Section 5 we will use to analyze the stillbirth data just described. Throughout the remaining sections, we let $t \in \mathcal{T}$ represent the target covariate, that is, the covariate in which the response surface is assumed to be smooth which in our case is gestational age (discrete time measured in weeks or days). We let $x \in \mathcal{X}$ represent a vector of covariates other than $t$ which in our case are the characteristics of a particular maternal-fetal dyad.

Because tsBART is a general approach for targeted smoothing in nonparametric regression, we first introduce the model in full generality. We then explain how to adapt it more specifically for modeling the hazard function for stillbirth, $h(t, x)$, which represents the conditional probability of stillbirth at gestational age $t$, given that a fetus has survived in utero through gestational age $t - 1$.

3.1. *The BART model.* Before introducing tsBART, we briefly review the original BART framework. BART (for Bayesian Additive Regression Trees) is a fully Bayesian ensemble-of-trees model (Chipman, George and McCulloch (2010)). BART models the mean response for a nonlinear regression function as the sum of a large number of binary trees, each of which is constrained by the BART prior to be shallow (and therefore a weak learner). The model is defined by a likelihood and prior, and inference is performed by sampling from the posterior. Specifically, suppose that $y_i$ is a scalar response and that $x_i$ is a vector of covariates. The BART model assumes that

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \tag{3.1}$$

$$f(x_i) = \sum_{j=1}^{m} g(x_i; T_j, M_j). \tag{3.2}$$

Here, each $T_j$ is a binary tree that induces a step function in $x$ via a partition of the covariate space, while the $M_j = \{\mu_{1j}, \ldots, \mu_{b_j j}\}$ are the $b_j$ terminal node values in tree $j$ (i.e., the levels of the step function). We can think of each $g$ as a basis function parameterized by the binary tree defined by $(T_j, M_j)$.

The BART prior consists of three elements. The first component is the conjugate prior for the error variance, $\sigma^2 \sim \nu\lambda/\chi^2_\nu$. The second component is the specification of independent Gaussians $\mu_{hj} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_0, \tau^2)$ on the terminal node parameters $M_j = \{\mu_{1j}, \ldots, \mu_{b_j j}\}$ of each tree. The third component is the prior over tree space, composed of a set of probabilities governing three things: the choice of splitting covariate, the choice of splitting value for each covariate and whether a node at a given depth is a terminal node. We refer interested readers to Chipman, George and McCulloch (2010), who recommend default hyperparameters that favor shallow trees which both regularizes the estimate and encourages rapid mixing.

BART has been successful in a variety of contexts including prediction and classification (Chipman, George and McCulloch (2010), Linero (2018), Linero and Yang (2018), Murray (2017), Hernández et al. (2018)), survival analysis (Sparapani et al. (2016)) and causal inference (Hahn, Murray and Carvalho (2017), Hill (2011), Logan et al. (2019), Sivaganesan, Müller and Huang (2017)).

3.2. *The tsBART model.* Motivated by the success of BART models, we introduce tsBART, an extension of BART for estimating regression functions that are smooth in a target covariate. Consider a regression problem with scalar response $y_i = f(t_i, x_i) + e_i$, where the underlying mean function $f(t_i, x_i)$ depends both on $t$ (a scalar) and $x$ (a vector), and should be smooth in $t$. To adapt BART for this setting, we replace the scalar node-level parameters $\mu_{hj}$ with univariate functions in $t$, $\mu_{hj}(t)$, and we assume that only $x$ variables (but not the target variable $t$) are used to define tree splits (see Figure 1). These univariate functions in $t$ can in principle be assigned any prior over function space; in the applications considered in this paper, we use Gaussian process priors.

More formally, we express the tsBART model as follows. Suppose that each observation $i$ in our dataset consists of predictor variables $(t_i, x_i)$ together with outcome $y_i$ for $i = 1, \ldots, N$. (Recall that $t_i$ is the target variable for smoothing, while $x_i$ is a vector of all other variables.) We now let

$$\begin{aligned} y_i &= \alpha(t_i) + f(t_i, x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \\ f(t_i, x_i) &= \sum_{j=1}^{m} g(t_i, x_i; T_j, M_j). \end{aligned} \tag{3.3}$$

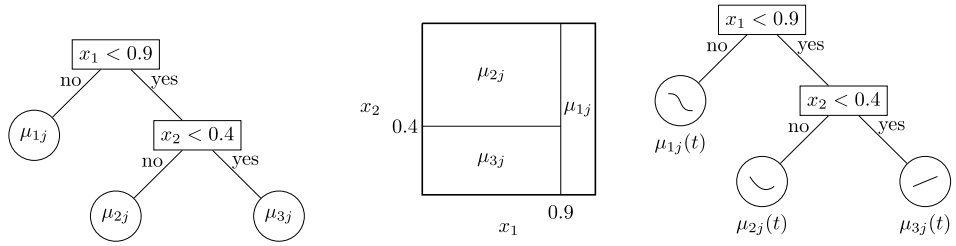


FIG. 1. *Illustrates the extension of BART to tsBART for a single tree, where nodes are parameterized by smooth functions instead of scalar parameters. (Left) A vanilla BART example binary tree $T_j$ where terminal nodes are labeled with the corresponding scalar parameters $\mu_{hj}$. (Middle) The corresponding vanilla BART partition of the sample space and the step function $g(T_j, M_j)$. (Right) Our BART with Targeted Smoothing modification, where the $\mu_{hj}(t)$ parameters associated to terminal nodes are now functions of $t$.*

Here, $T_j$ is a binary tree whose terminal nodes partition the "nontarget" covariate space $\mathcal{X}$ into $b_j$ disjoint regions, just as in the original BART model. But unlike BART, we parametrize the terminal nodes of the tree not by scalars but by a collection of Gaussian processes in $t$: $M_j = \{\mu_{1j}(t), \ldots, \mu_{b_j j}(t)\}$, with each function $\mu_{hj}(t)$ associated with one terminal node. The right panel of Figure 1 illustrates an example with $b_j = 3$ terminal nodes. The overall response is the sum of $m$ such trees, so that at any fixed design point $x = (x_1, \ldots, x_p)$, the response $f(t, x)$ is the sum of $m$ Gaussian processes.[3] We center the model at $\alpha(t)$, a baseline function of $t$, so that the trees parametrize deviations from the baseline that are associated with $x$. We estimate $\alpha(t)$ using the sample mean response for observations at each $t$.

We use the same prior over tree space as in the original BART paper. To model the $\mu_{hj}(t)$'s in each terminal node, we use a zero-centered Gaussian process prior,

$$\mu(t) \sim \mathrm{GP}\big(0, C_\theta(t, t')\big),$$

where $C_\theta(t, t')$ is the covariance function with hyperparameter $\theta$ which can be either chosen based on prior knowledge or tuned using the data. (Zero-centering is appropriate here because we separate out the mean term $\alpha(t)$ in equation (3.3).)

In principle, any covariance function can be used. For all examples in this paper, we use the squared-exponential covariance function with variance parameter $\tau^2/m$ and length scale $l$. That is,

$$C(t, t') = \frac{\tau^2}{m} \exp\left\{-\frac{d(t, t')^2}{2l^2}\right\}, \tag{3.4}$$

where $d(t, t')$ is the Euclidean distance between $t$ and $t'$. Here, $\tau^2$ determines the marginal variance of the $\mu_{hj}$'s, while $l$ governs their "wiggliness." As in the original BART model, we scale the variance parameter $\tau^2$ inversely by the number of trees $m$. Since the mean-response function $f(t, x)$ is the sum of $m$ trees, this implies that the marginal prior variance of $f(t, x)$ at any point $t$ is $\tau^2$. We then assign $\tau$ a folded-Cauchy prior as in Gelman (2006).

The tsBART model also requires specifying $l$, the length scale of the Gaussian process prior, with larger $l$ corresponding to more wiggliness. This length scale can be set using prior knowledge, but in Section 3.3 we provide a method to tune it automatically over a grid of possible values. As we also explain in Section 3.3, a reasonable default choice when using the squared exponential covariance function is $l = T/\pi$, where $T$ is the range of $t_i$ values in the dataset.

[3]This implies that $f(t, x)$ is a Gaussian process in $t$ for fixed $x$, but it is not a Gaussian process in $(t, x)$ jointly.

We make the simplifying assumption of an $i.i.d.$ error structure and complete the model specification by assigning $\sigma^2$ an inverse chi-square distribution $\sigma^2 \sim \nu\lambda/\chi^2_\nu$. For full computational details, including the prior specification and posterior full conditional distributions, see Appendices A.1 and A.2.

3.3. *Tuning the length scale $l$.* We must select $l$, the length-scale parameter of the covariance matrix. To do this, we represent $l$ using a formula by Kratz (2006) for the expected number of times a random function crosses its mean, $\mathrm{E}[N_T(s)]$, on some interval $I = [0, t_{\max}]$. This formula gives us a closed-form solution for the length-scale parameter as a function of the expected number of times that $f(t,x)$ crosses zero. Recall that if $f(t,x) = 0$, then the overall response at predictor $x$ is simply $\alpha(t)$, which we can think of as the baseline response over $t$. The more times that $f(t,x)$ crosses zero, the more sharply the covariate-specific mean response deviates from the overall mean response.

To set $\mathrm{E}[N_T(s)]$, let $r(s)$ be the correlation function between time 0 and time $s$:

$$r(s) = \frac{\mathrm{E}[\{f(0,x) - \mu(0)\}\{f(s,x) - \mu(s)\}]}{\mathrm{sd}(f(0,x)) \cdot \mathrm{sd}(f(s,x))}.$$

Per Kratz,

$$\mathrm{E}\big[N_T(s)\big] = t_{\max} \cdot \exp\left[-\frac{s^2}{2}\right]\left(\frac{\sqrt{r''(0)}}{\pi}\right),$$

and we let $s = 0$ in order to maximize $t_{\max}$. We use the squared exponential covariance kernel, so

$$r(t) = \frac{\mathrm{Cov}(f(0,x), f(t,x))}{\tau^2} = \exp\left[-\frac{t^2}{2l^2}\right].$$

Some algebra yields

$$l = \frac{t_{\max}}{\pi \mathrm{E}[N_T(0)]}. \tag{3.5}$$

This opens up several options for choosing the length scale. The first is by subjective choice. This would entail eliciting a guess for $\kappa \equiv E[N_T(0)]$, the average number of times that $f(t,x)$ will cross zero over all values of the covariates—or equivalently, the average number of times that each response $\alpha(t) + f(t,x)$ will cross the overall mean response $\alpha(t)$. This is a useful basis for elicitation, since the number of crossings is a sensible and intuitive measure for the wiggliness of our response as a function of $t$.

The second option is to choose a default value for $\kappa$. If a default must be chosen, we recommend $\kappa = 1$, or equivalently, $l = t_{\max}/\pi$. This encodes the belief that each response surface in $t$ will cross the overall mean response $\alpha(t)$ once on average across all predictor values. This allows for a substantial amount of heterogeneity in the mean responses over time while still shrinking toward the overall mean.

A final option, which we use in our simulation studies and real-data examples, is to tune $\kappa = \mathrm{E}[N_T(0)]$ over a grid of candidate values. This could be done using cross validation, as in the original BART paper, although we use the Watanabe–Akaike information criterion, or WAIC (Watanabe (2013)). WAIC is calculated as the log pointwise posterior predictive density plus a penalty for effective number of parameters, to avoid overfitting. It provides an estimate of generalization error without requiring that we split the data into multiple subsets; see Appendix A.3 for details. In our simulation we note that values of $\kappa \approx 1$ are frequently chosen by this data-driven approach, lending further credence to the choice of $\kappa = 1$ as a reasonable default.

3.4. *Adapting tsBART for binary and time-to-event outcomes.* In their original paper Chipman, George and McCulloch (2010) provide a probit version of the BART model for binary outcomes $Y \in \{0, 1\}$:

$$p(Y = 1 \mid x) = \Phi\big(G(x)\big), \tag{3.6}$$

$$G(x) = \sum_{j=1}^{m} g(x; T_j, M_j), \tag{3.7}$$

where $\Phi(\cdot)$ is the standard normal CDF, and where $G$ is the standard BART model.

Our tsBART model can be extended in the same way. Suppose that we observe a binary response $c_i$, together with target covariate $t_i$ and nontarget covariates $x_i$. The tsBART-probit model introduces a latent Gaussian variable $z_i$ (Albert and Chib (1993)) and then parameterizes $z_i$ using tsBART in a manner parallel to the original BART probit model:

$$c_i = \begin{cases} 1 & \text{if } z_i \geq 0, \\ 0 & \text{if } z_i < 0, \end{cases} \tag{3.8}$$

$$z_i = \alpha(t_i) + f(t_i, x_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1), \tag{3.9}$$

$$f(t_i, x_i) = \sum_{j=1}^{m} g(t_i, x_i; T_j, M_j). \tag{3.10}$$

Here, $\alpha(t)$ and $f(t, x)$ are defined exactly as in equation (3.3), and each $g_j$ is assigned the same prior outlined the previous subsection. Marginalizing over $z_i$ yields the desired probability under the probit model, $P(c_i = 1 \mid x_i, t_i) = \Phi\{\alpha(t_i) + f(t_i, x_i)\}$.

Crucially for our application, it is also straightforward to extend tsBART probit to discrete right-censored time-to-event outcomes, as noted by Sparapani et al. (2016) in the context of the original BART-probit model. Suppose that $t_i \in \mathcal{T}$ is a discrete time-to-event outcome, and that $c_i$ is a censoring indicator: $c_i = 1$ means that an event occurred at time $t_i$, while $c_i = 0$ means that observation $i$ was right-censored at time $t_i$. In our stillbirth risk-modeling problem, $c_i = 1$ corresponds to a stillbirth at gestational age $t_i$, while $c_i = 0$ corresponds to a live birth at $t_i$ (which is right-censoring with respect to the stillbirth event). The object of interest is the set of conditional probabilities $\mathbf{p} = \{p_{it}\}$, where $p_{it}$ is the conditional probability of an event at time $t$ for observation $i$, given than no event has happened through time $t - 1$. These conditional probabilities define the discrete-time hazard function $h(t, x)$. For ease of exposition, we assume here that the possible event times are $\mathcal{T} = \{1, \ldots, T\}$, but this is not a requirement.

To accommodate this data structure, we use the following standard factorization of the likelihood for a discrete-time hazard model. We introduce binary auxiliary variables $\{\tilde{c}_{is} : s = 1, \ldots t_i\}$ for each observation $i = 1, \ldots, N$, where

$$\tilde{c}_{is} = \begin{cases} 1 & \text{if } c_i = 1 \text{ and } s = t_i, \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood for the hazard function is now

$$L(\mathbf{p}) = \prod_{i=1}^{N} \prod_{s=1}^{t_i} p_{is}^{\tilde{c}_{is}} (1 - p_{is})^{1 - \tilde{c}_{is}}.$$

We note, as do Sparapani et al. (2016), that the product form of this likelihood does not come from the assumption that the binary $\tilde{c}_{is}$ events are independent but rather from the definition of each $p_{is}$ as a conditional probability.

We now construct the tsBART-probit model for $\mathbf{p}$, as follows:

$$\tilde{c}_{is} \sim \begin{cases} 1 & \text{if } z_{is} \geq 0, \\ 0 & \text{if } z_{is} < 0, \end{cases} \tag{3.11}$$

$$z_{is} = \alpha(s) + f(s, x_i) + \epsilon_{is}, \quad \epsilon_{is} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1), \tag{3.12}$$

$$f(s, x_i) = \sum_{j=1}^{m} g(s, x_i; T_j, M_j), \tag{3.13}$$

where $\alpha$ and the $g_j$'s are parametrized just as in the tsBART model described previously, treating time as the target covariate for smoothing.

3.5. *Connection with existing work.* Our paper sits in a long line of other research on extensions to the Bayesian tree-modeling framework. Two papers in particular are especially close in spirit to ours. The first is Sparapani et al. (2016) who introduce a model for nonparametric survival analysis using BART. Their model incorporates dependence on $t$ by simply adding time as an ordinary covariate to a BART-probit for the discrete-time hazard function. This does not impose any continuity or smoothness constraints on $f(t, x)$. In contrast, our approach smooths the hazard function over time while still retaining the benefits of BART. The second paper is the treed Gaussian process (TGP) model of Gramacy and Lee (2008). Their model uses a single deep tree with a Gaussian process in each terminal node; our model, in contrast, is a sum of many trees. Our work therefore generalizes that of Gramacy and Lee (2008) in the same way that the BART model generalizes the single-tree Bayesian CART model of Chipman, George and McCulloch (1998).

Smooth or partially smooth extensions of Bayesian tree models have also been proposed previously by Linero and Yang (2018) who smooth 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. In contrast, our approach smooths over just one target covariate while avoiding the high computational cost associated with the method of Linero and Yang (2018).

**4. Simulations.** We conduct two simulation studies to compare tsBART to existing methods. These simulations are designed to evaluate tsBART along several dimensions—out-of-sample predictive performance, credible interval coverage and interpretability—in settings with varying degrees of complexity in covariate interactions.

Given the importance of uncertainty quantification for modeling stillbirth risk, we do not benchmark against pure machine-learning methods that do not readily produce valid confidence or credible intervals. This excludes neural networks, boosting, CART and many other ensemble methods. We do, however, benchmark against BART which has been shown to enjoy comparable or superior predictive performance to all these pure machine-learning methods across a range of scenarios (see, e.g., Chipman, George and McCulloch (2010) who run these comparisons across 42 benchmark datasets). Thus, very little is lost by excluding methods that perform comparably to BART in terms of pure prediction but that cannot produce confidence/credible sets for those predictions.

One plausible benchmark might be Random Forests, for which recent research (Wager, Hastie and Efron (2014)) has addressed the problem of accurate uncertainty quantification. However, we choose not to include Random Forest in the simulation benchmarks for two reasons. First, Chipman, George and McCulloch (2010) performed extensive benchmarking of ordinary BART versus Random Forests, and they make a persuasive argument that, if the computational resources are available for BART, it tends to perform a bit better on average.

Additionally, we did investigate the performance of Random Forests on the stillbirth dataset that we analyze in Section 5. We found that the stillbirth risk curve estimates provided by Random Forest had many of the same interpretational problems posed by BART—namely, by not imposing adequate smoothness over time, it limits the interpretability for clinicians, encouraging them to over-interpret small wiggles in the fit. This analysis is included in Appendix A.5.

4.1. *Simulation* 1*—direct comparison with BART.* We first conduct a simulation study comparing tsBART to the ordinary BART model. The initial focus on BART is intended to isolate a key feature of our approach, smoothing in $t$ vs. simply including $t$ as another predictor available in the model. BART is also the most relevant practical comparison for our application, since Sparapani et al. (2016) have already shown that ordinary BART probit has cutting-edge performance for discrete-time survival modeling vs. a wide range of competing methods, including many more traditional time-to-event models.

We simulated datasets across three scenarios of modest dimension in the nontarget variables $x$: one with four covariates, one with eight covariates and another with 20 covariates. For all scenarios we used eight discrete time points ($\mathcal{T} = \{1, \ldots, 8\}$) for the target covariate. This mimics the stillbirth data, where information on gestational age is used at a weekly resolution between 34 and 42 weeks. It also reflects many other obstetrics, public health and biomedical applications where data is observed at discrete intervals. We generated each pair of covariates $(x_{ij}, x_{i,j+1})$ for odd $j$ from a bivariate Gaussian with moderate correlation and unit variances. For each case, we simulated data sets with sample sizes $n \in \{100, 500, 1000, 2500\}$, for a total of 12 combinations of sample size ($n$) and dimension of the nontarger covariate ($p$). For each of these 12 combinations, we simulated 100 datasets.

We focus on a ground truth in which the mean response evolves smoothly in $t$, and we seek to answer two key questions: 1) can tsBART adapt to the correct degree of smoothness, and 2) if so, how large are the gains versus an otherwise very similar model that makes no smoothness assumptions? In the $p = 4$ case, we let

$$f(t, x) = g(x_1, x_2) \cdot \cos\big(t + 2\pi h(x_3, x_4)\big),$$

so that the covariates $x$ modify both amplitude and phase shift. We let $g$ and $h$ be simple functions of the covariate pairs; here, we sum each pair of covariates.

In the $p = 8$ and $p = 20$ cases, we continue in a similar fashion, alternating sines and cosines, so that

$$\begin{aligned} f(t, x) = {} & g(x_1, x_2) \cdot \cos\big(t + 2\pi h(x_3, x_4)\big) \\ & + g(x_5, x_6) \cdot \sin\big(t + 2\pi h(x_7, x_8)\big) + \cdots, \end{aligned}$$

where this pattern continues. We again let $g$ and $h$ be sums of each pair of covariates. We generate responses $y(t, x) = f(t, x) + \epsilon$ where $\epsilon \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, 1)$.

We compare BART and tsBART using $m = 200$ trees and 10,000 MCMC draws with a burn-in of 1000 draws. We compare performance by calculating the log loss at each iteration of the algorithm, both in-sample and for a held-out sample, and by taking the mean log loss across all MCMC iterations. Log losses are scaled by sample size. We tune the length scale $l$ using the method described in Appendix A.3. We compare models using log loss since our goal is not to classify patients by whether they will experience stillbirth but to provide well-calibrated probabilities of stillbirth to clinicians. Log loss is a proper scoring rule which measures how effectively each method calibrates its probability estimates.

tsBART consistently outperforms ordinary BART (Table 2) in the out of sample log loss. tsBART has the most significant gains in scenarios with small sample sizes or more predictors. Figure 2 illustrates the out of sample fits and log loss in a single scenario, where $n = 500$ and $p = 4$; tsBART tends to smooth out the long-range periodicities in $f(t, x)$ much less than ordinary BART.

TABLE 2

*In-sample and out-of-sample log loss (scaled by sample size), averaged across 100 replicates, comparing BART with tsBART. tsBART consistently outperforms ordinary BART in the out-of-sample log loss. tsBART has the most significant gains in scenarios with small sample sizes or more predictors, as evidenced by p-values from a paired Wilcoxon test comparing out-of-sample results*

| | | In-sample | | Out-of-sample | | |
|---|---|---|---|---|---|---|
| $p$ | $n$ | BART | tsBART | BART | tsBART | $P$-value |
| 4 | 100 | −1.61 | **−1.49** | −1.97 | **−1.92** | <0.001 |
| 4 | 500 | −1.53 | **−1.47** | −1.80 | **−1.74** | <0.001 |
| 4 | 1000 | −1.47 | **−1.46** | −1.76 | **−1.72** | 0.001 |
| 4 | 2500 | **−1.43** | −1.44 | −1.68 | **−1.67** | 0.367 |
| 8 | 100 | −1.74 | **−1.66** | −2.18 | **−2.07** | <0.001 |
| 8 | 500 | −1.66 | **−1.63** | −2.02 | **−1.92** | <0.001 |
| 8 | 1000 | **−1.55** | −1.58 | −1.95 | **−1.91** | 0.002 |
| 8 | 2500 | **−1.48** | −1.53 | −1.88 | **−1.87** | 0.742 |
| 20 | 100 | −2.04 | **−2.02** | −2.59 | **−2.31** | <0.001 |
| 20 | 500 | **−1.94** | −1.99 | −2.41 | **−2.28** | <0.001 |
| 20 | 1000 | **−1.81** | −1.94 | −2.31 | **−2.27** | <0.001 |
| 20 | 2500 | **−1.66** | −1.84 | **−2.27** | −2.32 | 0.023 |

4.2. *Simulation 2—comparison with BART and splines.* We next compare tsBART to four existing models in a simulation study designed to mimic the basic properties of the hazard functions we expect to see in our stillbirth data. We generate hazard functions and corresponding survival data for three scenarios, where covariates determine shape of the hazard function with increasing degrees of interaction complexity. We compare the following methods:

1. tsBART: The BART with Targeted Smoothing method with smoothing parameter $\kappa$ tuned as described in Appendix A.3.

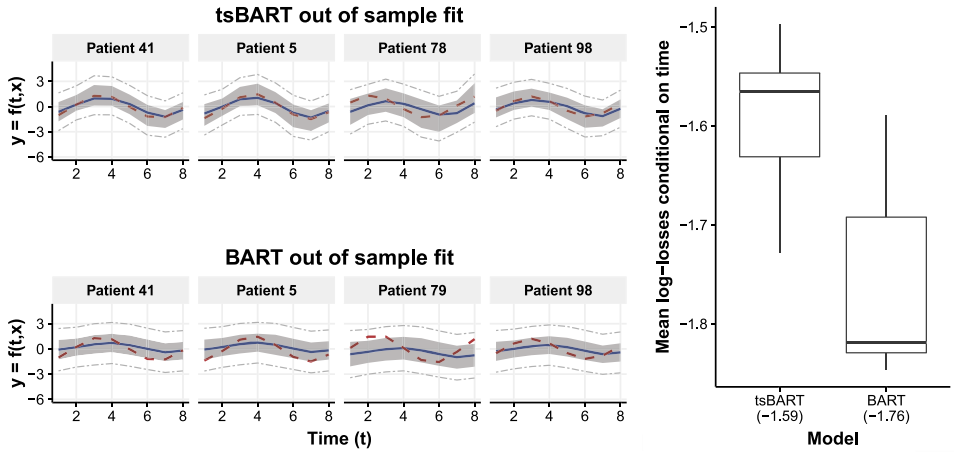


FIG. 2. *Compares a single model fit for tsBART and BART for one scenario ($n = 500$, $p = 4$), to illustrate the difference in fit when the ground truth is smooth in a single target covariate $t$. (Left) The bold dashed line represents the true function value, while the solid line and shading give posterior means and 95% credible intervals. Lighter dashed lines give the prediction intervals. (Right) tsBART outperforms BART across $t$ and on average. The boxplot gives the distribution of average log loss at each $t$ for both methods with overall average log loss in parenthesis on the $x$-axis label.*

2. tsBART (default): The BART with Targeted Smoothing method with our suggested $\kappa = 1$ default smoothing parameter.

3. BART: an ordinary BART-probit model which also sets hyperparameters $(\nu, \lambda)$, as recommended in Chipman, George and McCulloch (2010), and includes time $t$ as a covariate

4. Splines 1: a logistic regression model using cubic B-splines with seven degrees of freedom with main effects for all covariates included in $x_i$. Use of the spline basis induces targeted smoothness in $t$ by ensuring that, for fixed $x$, $f(t, x)$ is piecewise polynomial with continuous first and second derivatives.

5. Splines 2: another logistic regression model using cubic B-splines and seven degrees of freedom with the addition of interactions between each basis element in $t$ and each covariate in $x_i$.

6. P-Splines: a penalized spline model including the same covariates and a penalized spline basis with 9 spline basis elements and a second-order smoothing penalty (Eilers and Marx (1996)). (The maximum possible number of basis elements is determined by the fact that there are only nine distinct values for gestational age, 34–42 weeks.) By allowing for all possible basis elements to enter the model while penalizing deviations from smoothness, penalized splines provide flexibility while still regularizing the stillbirth risk curve estimates.

We evaluate the performance of tsBART for three scenarios, representing increasing degrees of difficulty in how $x$ parametrizes the hazard function.

We simulate data as follows. Let $t$ be a grid of times on the unit interval, spaced in increments of 0.1. Generate $n = 1000$ ten-dimensional covariates $x_i = \{x_{i1}, \ldots, x_{i10}\}$ where $x_{ij} \overset{\text{i.i.d.}}{\sim} \text{U}(0, 1)$. The first five covariates in each $x_i$ impact the response; the rest are noise. In each case we define the hazard function as the weighted combination of two "template" hazard functions $f_1(t)$ and $f_2(t)$, where weights $w(x_i)$ depend on covariates $x_i$:

$$h(t, x) = 0.25x_{i5} + w(x_{i1}, \ldots, x_{i4}) f_2(t) + \big[1 - w(x_{i1}, \ldots, x_{i4})\big] f_1(t).$$

The differences between the three scenarios are in how the weight depends on the covariates: linearly, linearly with interactions or nonlinearly with interactions. Figure 3 illustrates resulting hazard functions for each scenario. There are four general hazard function shapes, dictated by high versus low baseline risk and with or without a sharp increase in hazard beginning at $t = 0.75$. Appendix A.4 provides further detail, and code is available at https://github.com/jestarling/tsbart-analysis.

For each of the three scenarios, we simulate 500 datasets to compare point-wise coverage of tsBART compared to the methods detailed in Section 5. The mean-squared error of the estimates are small and comparable across all methods. Most striking, however, is that tsBART gives far better coverage than other methods, both with the smoothing parameter tuned and set to the default value of 1 (Table 3). No other method consistently produces credible/confidence sets that are close to the nominal value of 95%. We conclude that tsBART

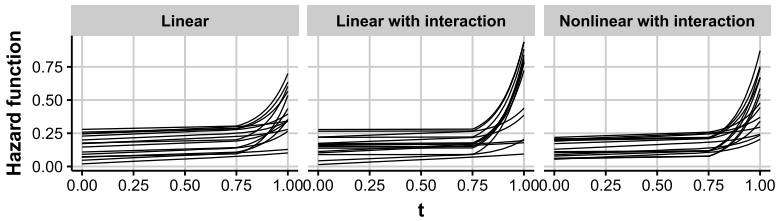


FIG. 3. *Illustrates the different ground truth hazards used in the three simulation scenarios. The same basic shapes of hazard functions are present in all three scenarios; the difference is in how covariates x influence which shape arises. There is a mixture of gently-rising hazards and hockey stick hazards; linearity determines the straightness of the rise, and presence of an interaction increases strength of the hockey stick.*

TABLE 3

*Average coverage rates (nominal is 95%) and mean squared error across 500 simulated datasets for each weighting scenario and model combination. For tsBART and BART, coverage is for posterior credible intervals and mean squared error uses the posterior mean. For the spline-based methods, coverage is for prediction intervals. tsBART has better coverage, even with the default smoothing parameter, and MSE for all methods is small and comparable.*

| Weighting scenario | Method | Coverage | MSE |
|---|---|---|---|
| Linear | tsBART (tuned) | 0.9310 | 0.0014 |
| | tsBART (default) | 0.9092 | 0.0017 |
| | BART | 0.7642 | 0.0011 |
| | Splines 1 (Linear) | 0.7925 | 0.0007 |
| | Splines 2 (Interaction) | 0.7788 | 0.0014 |
| | P-Splines | 0.7720 | 0.0007 |
| Linear (with interaction) | tsBART (tuned) | 0.9571 | 0.0019 |
| | tsBART (default) | 0.9443 | 0.0022 |
| | BART | 0.7907 | 0.0022 |
| | Splines 1 (Linear) | 0.8874 | 0.0039 |
| | Splines 2 (Interaction) | 0.7213 | 0.0391 |
| | P-Splines | 0.8718 | 0.0036 |
| Nonlinear (with interaction) | tsBART (tuned) | 0.9539 | 0.0013 |
| | tsBART (default) | 0.9354 | 0.0016 |
| | BART | 0.7408 | 0.0012 |
| | Splines 1 (Linear) | 0.8918 | 0.0006 |
| | Splines 2 (Interaction) | 0.8392 | 0.0013 |
| | P-Splines | 0.8747 | 0.0006 |

is capable of matching or exceeding other methods in terms of mean-squared error while producing error bars that are statistically trustworthy and scientifically sensible.

We acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the results reported within this paper. URL: http://www.tacc.utexas.edu

**5. Results for modeling stillbirth risk.** We now turn to our motivating application, by applying the tsBART method to estimate patient-specific stillbirth risk, using the data described in Section 2. To model the discrete-time hazard function for stillbirth, $h(t, x)$, we use the extension of tsBART-probit formulation described in Section 3.4. Our target covariate for smoothing is gestational age in weeks: $t_i \in \{34, 35, \ldots, 42\}$. We let $y_i$ be an indicator of whether stillbirth has occurred for each pregnancy, and $x_i$ be the vector of maternal-fetal covariates for each patient, including maternal age, primiparity, ethnicity, infant sex, presence of diabetes mellitus, presence of chronic hypertension, presence of other risk factors, whether the pregnancy was induced and birth weight and weight gain quantiles.

We first focus on the question of whether tsBART does, indeed, yield better-calibrated risk estimates over existing methods for our dataset. For the purpose of evaluating all models while maintaining computational tractability, we created five balanced case-control samples of $n = 1000$ pregnancies each. (Since stillbirth is a rare event, using a balanced case-control sample also more clearly highlights differences among methods.) We then split each balanced case-control sample into training and testing sets. We used the training set to fit tsBART in addition to each of the four models discussed in Section 4: vanilla BART, the two B-spline models and P-splines. We tune the length-scale parameter of tsBART using the method described in Appendix A.3, and we set tree-prior hyperparameters $(\nu, \lambda)$ as recommended in Chipman, George and McCulloch (2010). We then used the fitted model to predict the hazard

TABLE 4
*Overall out-of-sample log loss for each method, averaged over five evenly balanced case-control samples. tsBART outperforms other methods, with the tuned smoothness parameter only slightly outperforming the default.*

| Method | Log loss |
|---|---|
| tsBART (tuned) | **−1.711** |
| tsBART (default) | −1.713 |
| BART | −1.810 |
| Splines 1 | −1.725 |
| Splines 2 | −1.919 |
| P-splines | −1.724 |

functions for all held-out points, and we computed held-out log losses. We repeat this process over five balanced case-control datasets and average the results (Table 4).

tsBART outperforms other methods, with the tuned smoothness parameter setting only slightly outperforming the default (untuned) setting. To provide some intuition for these results, Figure 4 also shows relative out-of-sample log losses of all methods as a function of gestational age with tuned tsBART normalized to 1. The figure shows that tsBART's gains are especially apparent at higher and lower gestational ages, where fewer observations are available. Most methods are comparable at gestational ages across the middle of the available range (37–39 weeks).

We next turn to the question of how obstetricians might use the results of tsBART to understand stillbirth risk and to communicate that risk to their patients. To do so, we construct a set of hypothetical "test" patients representing various configurations of maternal-fetal characteristics:

- Patient 1 is a young, primiparous, white patient in her early 20's with no medical history, normal weight gain and normal birth weight for a female infant.
- Patient 2 is otherwise similar to Patient 1 but has hypertension.

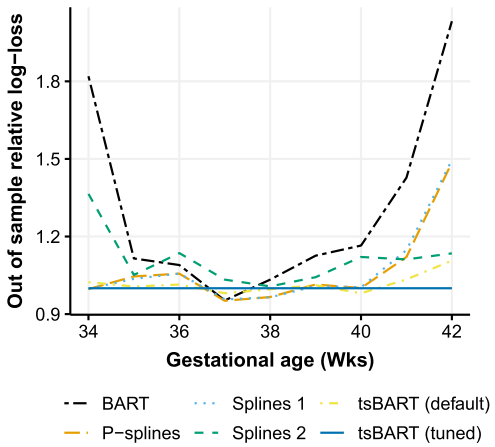


FIG. 4. *Illustrates performance of each method relative to tsBART with tuned smoothing parameter κ. Shows weekly out-of-sample log loss for each method, averaged over five evenly balanced case-control samples. tsBART's gains are especially apparent in higher and lower gestational ages; other methods have small gains in the* 37 *to* 39 *week range, at the expense of inflation at extreme gestational ages where sample sizes are small.*

- Patient 3 is otherwise similar to Patient 1 but has both hypertension and diabetes.
- Patient 4 is also a young white patient in her early 20's but is multiparous with birth weight less than the 10th quantile.
- Patient 5 is a white patient in her early 40's with diabetes, hypertension and other risk factors present; her labor is induced, and her infant is male.

To maintain computational tractability, we again select a case-control sample of the overall data set. We include all stillbirths in the case-control sample. Then, for each gestational age, we sample 2% of the live births at that age. As a result, stillbirths are 50 times more prevalent in our sample than they are in the full data set, both overall and at each gestational age. This approach yields a dataset that is still reasonably large with 91,078 pregnancies, all 7,175 stillbirth cases and 83,903 live-birth controls. While we would prefer to fit the model to all 4.55 million data points, we are not yet able to do so, owing to computational constraints. Scalable Bayesian ensemble methods are an active area of research, and we are currently drawing on this work to develop methods for scaling tsBART to use the entire dataset.

We use this large case-control sample to fit all methods from Section 4. We use the results to produce estimates of the stillbirth hazard function for each of our hypothetical test patients. We then rescale the estimated hazard functions to account for the 50-fold downsampling of live births in our case-control sample, and we express the resulting hazard functions as a stillbirth rate per 1000 live births.

The results are shown in Figure 5. Each column represents one test patient, while each row shows a particular method. In each panel we show the estimated conditional probability of stillbirth risk at gestational age $t$, given survival through time $t-1$, along with 95% uncertainty intervals. Estimated probabilities for all other methods are also visible in grey within each panel and for easier comparison across panels. For tsBART and BART the estimates are posterior-mean predicted probabilities and (Bayesian) credible intervals; for spline methods the estimates are predicted probabilities and (frequentist) prediction intervals.

These plots have several features of interest (we focus on the tsBART results in the top row). First, there is considerable heterogeneity in the estimated stillbirth risk curves, in their shape, level and degree of interaction between maternal-fetal covariates and gestational age. Patient 1, for example, has a lower overall risk with a relatively small increase in risk at very late gestational ages (41–42). Patients 2–4 have slightly higher overall risk at earlier gestational ages but more much pronounced "spikes" in risk at late gestational ages, when the inherent stillbirth risk at an advanced stage of pregnancy is exacerbated by these patients' covariates (hypertension, diabetes + hypertension and low fetal weight, respectively). Patient 5, on the other hand, has a higher overall risk at all gestational ages but a much more linear risk trajectory across gestational age compared with Patients 1–4, without the pronounced spike.

This striking heterogeneity across the patients illustrates the shortcomings of collapsing patients into two risk groups, as in Mandujano, Waters and Myers (2013). Our method, in contrast, can produce individualized estimates of risk for any patient and across all gestational ages.

We note that the estimates from the BART model are generally similar in shape to the tsBART estimates but lack smoothness over gestational age. This results in increased variance and poorer overall out-of-sample performance, as evident from Table 4. It also invites clinicians and patients to over-interpret small wiggles in the risk curves that are a result of estimation noise rather than clinically meaningful differences. The spline models, meanwhile, tend to result in estimates that are either over smoothed (Splines 1, P-splines) or under smoothed and erratic (Splines 2). We attribute this to the fact that Splines 1 and P-splines are underparametrized. They fail to include clinically meaningful interactions (e.g., between hypertension and diabetes). This results in higher bias, poorer estimation performance and

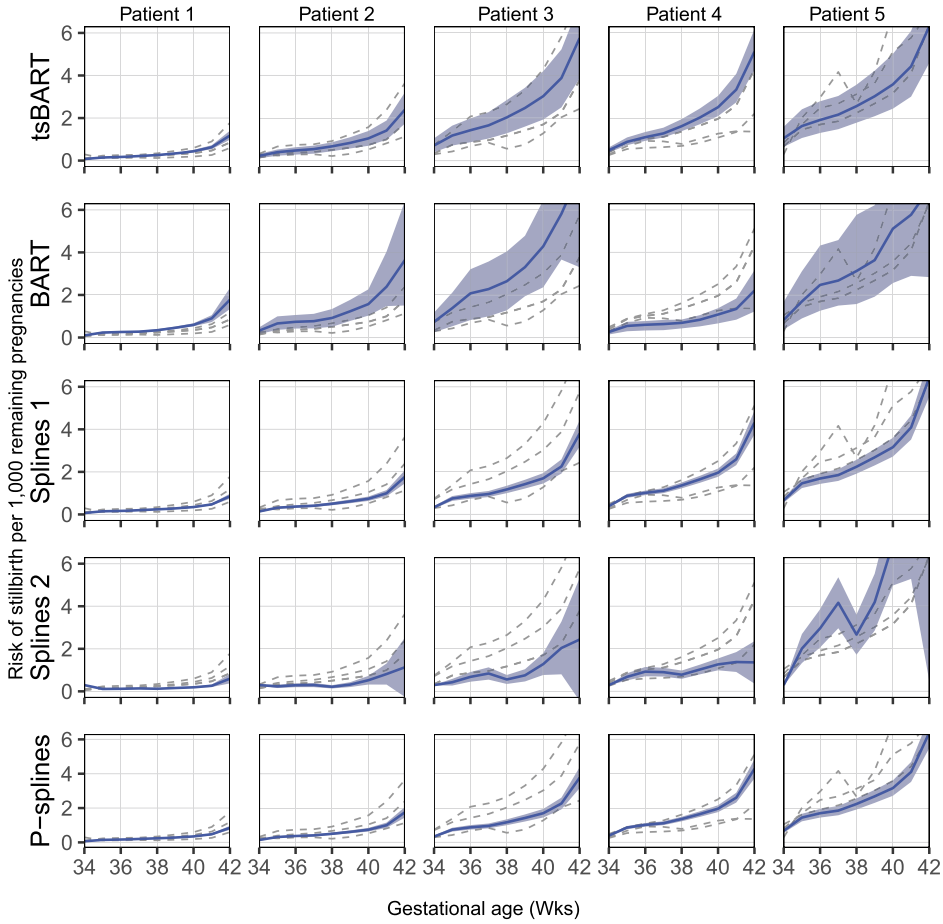


FIG. 5. *Estimated stillbirth risk curves for five hypothetical patients with different combinations of maternal-fetal covariates, using full case-control sample. Each row is a method, and each column is a hypothetical patient. In each row the posterior mean and credible interval are highlighted (dark line and shading), while the other methods' posterior means are in dashed lines for comparison. Patient* 1 *is a low-risk patient (young, primiparous, no medical history, normal weight gain and birth weight). Patient* 2 *introduces hypertension; Patient* 3 *introduces both diabetes and hypertension. Patient* 4 *is multiparous with very low birth weight, and Patient* 5 *has a combination of risk factors (older, diabetes, hypertension, medical history, induced labor). tsBART gives a smooth fit with sensible credible intervals, consistent with clinical intuition about the way stillbirth risk evolves with gestational age.*

infeasibly narrow confidence intervals that, in light of our simulation studies (Section 4), are likely to be anticonservative. Splines 2, meanwhile, is likely overparametrized. It allows for the possibility of all pairwise interactions between maternal-fetal covariates and gestational age, needlessly inflating variance for the sake of finding a small handful of clinically important interactions. This suggests that the spline models, in order to yield good performance for stillbirth prediction, would require more nuanced model selection and attention to functional form since including more flexible interactions was not a fruitful approach.

TsBART, in contrast, produces the best out-of-sample performance, smooth estimates and wider more clinically sensible error bars. It also finds the important interactions out of the box without the need to specify them by hand or to conduct a specification search for the right form of the model. In addition, the posterior credible intervals from tsBART are noticeably wider for patients with unusual combinations of characteristics—an intuitive result which reflects a higher degree of uncertainty about rarer, more medically complex cases.

**6. Discussion.** Our tsBART model is a novel extension of BART which allows for targeted smoothing over a selected covariate. tsBART enjoys the same advantages as BART—excellent predictive performance, easily tunable hyperparameters and avoiding specification of interactions. Hyperparameters are set efficiently via data-driven approaches using recommendations from Chipman, George and McCulloch (2010) and our suggested method for tuning the length-scale of the covariance function. tsBART provides regularization in the form of constraining trees to be shallow learners in the prior which is a well studied and highly successful approach to regularization in regression.

The kind of stillbirth risk analysis made possible by tsBART represents a substantial advancement on previous work in obstetrics (Mandujano, Waters and Myers (2013)) in terms of capturing heterogeneity of risk curves by patient and quantifying levels of certainty around each risk curve. Further investigation into nuanced approaches for stillbirth risk modeling is warranted; maternal-fetal covariates such as age, weight gain and birth weight may play a role in risk of stillbirth and may interact with other covariates in complex ways. Our fully Bayesian approach naturally allows the model to capture rich and complex interactions and quantify uncertainty about stillbirth risk which appropriately varies by patient.

We recognize the potential limitation of confounding between the decision to induce labor, risk of stillbirth and maternal-fetal covariates. We currently consider the decision to induce to be a proxy for other maternal-fetal covariates which may increase stillbirth risk but are not included in the model; future work may include modeling this covariate in a causal framework. A second limitation is the inability to link birth records to the same mother, potentially violating the independence assumption (the deidentified nature of the data prevents this linking). However, because our data set spans only two years, it is unlikely that a large fraction of the overall births are multiple births to the same mother. Moreover, the concern about nonindependence is mitigated because we have included many of the known risk factors for stillbirth in our model. While it is not plausible that two stillbirth events for the same mother are marginally independent, it is much more plausible that they are conditionally independent, or nearly so, given these risk factors.

Future areas of methodological work may include extension of tsBART to a causal inference framework for observational data as well as extension to other priors with other types of structure. tsBART may be adapted in the accelerated framework of He, Saar and Hahn (2018) to speed computation time. It would also be interesting to explore more nuanced characterizations of partial dependence of stillbirth risk on individual covariates. For example, plots of individual conditional expectation (ICE) may be used to assess partial relationships between response and specific covariates, using the techniques described in Goldstein et al. (2015). ICE plots go beyond the simple partial dependence plot by showing the functional relationship between response and feature at the level of individual observations (rather than averaging across the sample). This could potentially give insight into the extent of potential heterogeneity in the conditional expectation function. ICE plots can be created using the ICEbox R library (Goldstein et al. (2015)).

## APPENDIX

**A.1. Review of the Bayesian backfitting MCMC.** The original BART model is typically fit using an algorithm called Bayesian backfitting (Chipman, George and McCulloch (2010), Hastie and Tibshirani (2000)). We review this algorithm, then we describe the modifications necessary to fit the BART with Targeted Smoothing model.

Bayesian backfitting involves sampling each tree and its parameters one a time, given the partial residuals from all other $m - 1$ trees. One iteration of the sampler consists of looping through the $m$ 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 $m$ trees, $\sigma^2$ is updated in a Gibbs step.

To sample $\{T_j, M_j\}$ conditioned on the other trees and leaf parameters $\{T_{-j}, M_{-j}\}$, define the partial residual as

$$R_{ij} = y_i - \sum_{k=1, k\neq j}^{m} g(x_i; T_k, M_k). \tag{A.1}$$

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;
- *change* randomly selects an internal node and draws a new splitting rule;
- *swap* randomly selects a parent-child pair of internal nodes and swaps their decision rules.

The *change* and *swap* moves are computationally expensive; in practice, BART is often implemented with only *prune* and *grow* proposals (Pratola et al. (2014)). Once the move in tree space is either accepted or rejected, $M_j$ is sampled from its Gaussian full conditional, given $T_j$ and $\sigma^2$.

**A.2. Fitting the tsBART model with Bayesian backfitting.** Our approach to fitting tsBART retains the form of the Bayesian backfitting MCMC algorithm, as detailed by Chipman, George and McCulloch (2010). The primary modification is that all conjugate updates are modified to their multivariate forms. We assume an i.i.d. error structure, although this is easily modified, and we also use a multiplicative parameterization of the scale parameter to facilitate faster MCMC mixing (Gelman (2006), Hahn, Murray and Carvalho (2017)). Thus, our model is

$$\begin{aligned}
y_i &= \alpha(t_i) + \eta f(t_i, x_i) + \epsilon_i, \quad \epsilon_i(t) \overset{\text{i.i.d.}}{\sim} \mathrm{N}\big(0, \sigma^2\big), \\
f(t_i, x_i) &= \sum_{j=1}^{m} g(t_i, x_i; T_j, M_j), \qquad M_j = \big\{\mu_{1j}(t), \ldots, \mu_{b_j j}(t)\big\}, \\
\mu_{hj}(t) &\sim \mathrm{GP}\big(0, C\big(t, t'\big)\big), \\
\eta &\sim \mathrm{N}\big(\tau_0, \gamma^2\big), \\
\gamma^2 &\sim \mathrm{IG}\left(\frac{1}{2}, \frac{1}{2}\right), \\
\sigma^2 &\sim \nu\lambda/\chi^2_\nu.
\end{aligned}$$

Recall that $\mu_{hj}(t)$ is the function at terminal node $l$ of tree $j$. As described previously, this function has a Gaussian process prior with squared exponential covariance function with length scale $l$. Because we have already introduced $\eta$ as a leading multiplicative scale parameter, we set the variance parameter of the covariance function to be $1/m$ and calibrate the prior folded-Cauchy location $\tau_0$ to the marginal standard deviation of $y$.

We use the same prior for over trees $T_j$, as in Chipman, George and McCulloch (2010) and Hahn, Murray and Carvalho (2017), and so we omit many details here and refer the interested reader there. Specifically, these papers parametrize tree depth in terms of the pair $(\alpha, \beta)$; we set $(\alpha = 0.95, \beta = 2)$ which puts high probability on trees of depth 2 and 3 and minimizes probability on trees with depth 1 or greater than 4. For $\sigma^2$, we follow Chipman

et al.'s recommendation for a rough overestimation of $\hat{\sigma}$. We choose $\nu = 3$ and $q = 0.90$ and estimate $\hat{\sigma}$ by regressing $y$ onto $x$ (including the index variable as a covariate), then we choose $\lambda$ s.t.; the $q$th quantile of the prior is located at $\hat{\sigma}$, that is, $P(\sigma \leq \hat{\sigma}) = q$.

The posterior conditional distributions are as follows. For simplicity of notation, we assume times $t$ are on a common discrete grid, where $T$ is again the range of $t$ values in the data set (although this is not a requirement of the method). We update $\sigma^2$ as

$$\sigma^2 \mid \bullet \sim \frac{\nu\lambda + \mathrm{RSS}}{\chi^2_{\nu+N+1}} \quad \text{where } \mathrm{RSS} = \sum_{i,t} \big(y_i(t) - \eta f(t_i, x_i)\big)^2,$$

where $N$ is the count of observations across all time points, $N = \sum_{i=1}^{n} N_i$ where $N_i$ is the number of time points for observation $i$ and $\chi^2_{\nu+N+1}$ is a draw from a chi-squared random variable.

The update for each $\mu_h = [\mu_h^{(1)}, \ldots, \mu_h^{(T)}]$ is

$$\mu_h \mid \bullet \sim \mathrm{N}(\tilde{m}, \tilde{\Sigma}) \quad \text{where } \tilde{\Sigma} = (\Lambda + K)^{-1} \text{ and } \tilde{m} = \tilde{\Sigma}(\Lambda \bar{y}_l + K\mu_0),$$

where $\Lambda = N_l^{-1}$ is the inverse of the diagonal matrix of sample sizes for each time point for observations in leaf $l$, $K = \Sigma_0^{-1}$, and $\bar{y}_l$ is the vector of sample means for observations in leaf $l$ at each time point.

The update for $\eta$ is Gaussian,

$$\begin{aligned}
\mu_h \mid \bullet &\sim \mathrm{N}\big(\tilde{m}, \tilde{v}^2\big) \quad \text{where} \\
\tilde{v}^2 &= \left(\frac{1}{\gamma^2} + \frac{1}{\sigma^2}\sum_{i,t} f(t_i, x_i)^2\right)^{-1} \\
\tilde{m} &= \tilde{v}^2\left(\frac{\tau_0}{\gamma^2} + \frac{1}{\sigma^2}\sum_{i,t} y_i f(t_i, x_i)\right).
\end{aligned}$$

Finally, the update for $\gamma^2$ is

$$\gamma^2 \mid \bullet \sim \mathrm{IG}\left(1, \frac{\eta^2 + 1}{2}\right).$$

For updating the trees $T_j$, the marginal likelihood is the corresponding multivariate extension to the marginal likelihood in regular BART. We again let $R_{ij}$ represent the partial residuals, as defined in equation (A.1), and let $R_l$ denote the vector containing residuals for data points in leaf $l$. We then obtain the marginal likelihood for the $b$ terminal nodes as

$$p\big(R_h \mid T_j, M_j, \sigma^2\big) = \int_{\mu_h} \prod_{l \in 1:b} \mathrm{N}\big(R_h \mid W_l \mu_h, \sigma^2 I\big) \cdot \mathrm{N}(\mu_h \mid \mu_0, \Sigma_0) \partial \mu_h,$$

where $W_l$ is a $(t_{\max} \times n)$ matrix where elements indicate times at which each $y_i$ is observed. This Gaussian integral is easily computed in closed form.

**A.3. Additional detail on hyperparameter tuning for length scale.** Here, we provide additional detail regarding tuning the expected number of crossings $\mathrm{E}[N_T(0)]$ for calculating the covariance's length-scale parameter. We select the optimal $\mathrm{E}[N_T(0)]$ by beginning with a grid of candidate values $e_c \in \{e_1, \ldots, e_C\}$. For each candidate $e_c$, we fit the BART with Targeted Smoothing model and calculate WAIC (Watanabe (2013)), yielding a grid of WAIC values $\Omega = \{\omega_1, \ldots, \omega_C\}$.

The WAIC values contain Monte Carlos variation; to overcome this, we fit a cubic spline model to $\Omega$. Let $\zeta$ be the standard deviation of the residuals from this model fit. We select

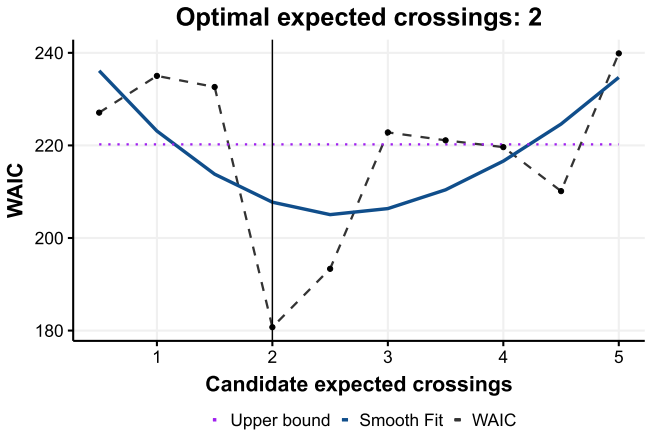


FIG. 6. *Example of tuning the expected number of crossings. The jagged dashed line illustrates the Monte Carlo variation present in WAIC estimates. The solid line shows the spline fit. The horizontal dotted line shows the minimum WAIC value plus one standard deviation from the spline fit. The solid vertical line gives the minimum candidate expected number of crossings value where there is a WAIC value less than one plus the standard deviation.*

the smallest number of expected crossings $e_c$, where the corresponding $\omega_c$ is within $\zeta$ of $\min(\Omega)$. This approach encourages smoothing while maintaining performance. Figure 6 gives a visualization of this tuning. Other methods, such as cross validation, could easily be used for tuning the expected number of crossings; we find this data-driven approach to be efficient while still yielding good results.

**A.4. Simulation details.** Here, we provide more detail for the second simulation described in Section 4. We simulate data as follows. Let $t$ be a grid of times on the unit interval and spaced in increments of 0.1. We generated $n = 1000$ ten-dimensional covariates $x_i = \{x_{i1}, \ldots, x_{i10}\}$ where $x_{ij} \overset{\text{i.i.d.}}{\sim} U(0, 1)$. The first five covariates in each $x_i$ impact the response; the rest are noise.

We generate data using a weighted combination of two risk functions, where $f_1(t) = 0.075t$ is the baseline risk function and $f_2(t) = 0.75 \cdot \max(0.75, t)^\rho$ is a second risk function which controls a large "kick" at $t = 0.75$. We let $\rho = 1 + \log(0.1)/\log(0.75)$, so that the $f_2(t)$ risk at $t = 1$ is five times the baseline risk.

The weights $w(x_i)$ for combining $f_1(t)$ and $f_2(t)$ are dependent on the covariates $x_i$. We generate data for three scenarios–letting weights $w(x_i)$ depend on covariates in either linearly, linearly with interactions or nonlinearly with interactions. These scenarios represent increasing degrees of difficulty in learning the underlying function:

- Linear:

$$w(x_i) = \text{sigmoid}\big[5(x_{i1} - x_{i2} + x_{i3} - x_{i4})\big].$$

- Linear with interaction:

$$w(x_i) = \text{sigmoid}\big[5(x_{i1} - x_{i2}) + 5(x_{i1} - 0.5)(x_{i2} - 0.5) + 5(x_{i3} - x_{i4}) + 5(x_{i3} - 0.5)(x_{i4} - 0.5)\big].$$

- Nonlinear with interaction:

$$w(x_i) = \text{sigmoid}\big[5\big(\max(x_{i1}, x_{i2})\big) - 5\big(\max(x_{i3}, x_{i4})\big)\big].$$

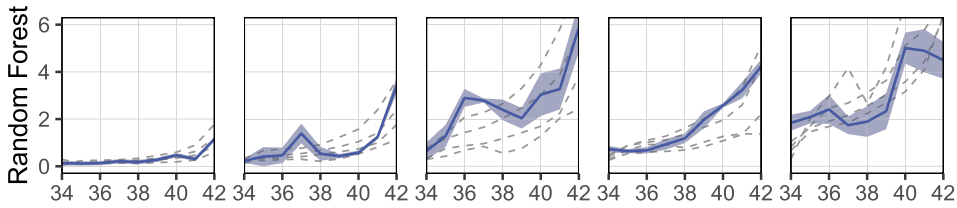


FIG. 7. *Estimated stillbirth risk curves from the Random Forest model, using the same five hypothetical patients and full case-control sample as Figure* 5. *The posterior mean and credible intervals are highlighted* (*dark line and shading*), *while the other methods' posterior means are in dashed lines for comparison. Each column is a hypothetical patient. Patient* 1 *is a low-risk patient* (*young*, *primiparous*, *no medical history*, *normal weight gain and birth weight*). *Patient* 2 *introduces hypertension*; *Patient* 3 *introduces both diabetes and hypertension. Patient* 4 *is multiparous with very low birth weight and Patient* 5 *has a combination of risk factors* (*older*, *diabetes*, *hypertension*, *medical history*, *induced labor*). *Random Forest gives curves generally similar in shape to BART, and does not induce smoothness.*

We then generate the simulated hazard function data according to $h(t)$, rescale responses so that the overall survival probability is roughly 0.5 and simulate event times for each observation:

$$h(t) = 0.25x_{i5} + w(x_i)\,f_2(t) + \big(1 - w(x_i)\big)\,f_1(t).$$

**A.5. Stillbirth results using random forest.** Here we illustrate the Random Forest fit for the stillbirth dataset. We do not include Random Forest in the set of models for stillbirth analysis; while Wager, Hastie and Efron (2014) provide variance estimation for Random Forest, Chipman, George and McCulloch (2010) demonstrated that BART tends to outperform Random Forest. In addition, Random Forest does not induce smoothness, as we see in Figure 7.

## SUPPLEMENTARY MATERIAL

**R package to implement methods** (DOI: 10.1214/19-AOAS1268SUPPA; .zip). The R package *tsbart* implements the BART with Targeted Smoothing method.

**R package to implement methods** (DOI: 10.1214/19-AOAS1268SUPPB; .zip). The package *tsbart-analysis* contains code to replicate figures and tables in the paper.

## REFERENCES

ALBERT, J. H. and CHIB, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88** 669–679. MR1224394

BAILIT, J. L., GREGORY, K. D., REDDY, U. M., GONZALEZ-QUINTERO, V. H., HIBBARD, J. U., RAMIREZ, M. M., BRANCH, D. W., BURKMAN, R., HABERMAN, S. et al. (2010). Maternal and neonatal outcomes by labor onset type and gestational age. *Am. J. Obstet. Gynecol.* **202** 245.e1–245.e12. https://doi.org/10.1016/j.ajog.2010.01.051

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (1998). Bayesian CART model search. *J. Amer. Statist. Assoc.* **93** 935–948.

CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172 https://doi.org/10.1214/09-AOAS285

CLARK, S. L., FRYE, D. R. and MYERS, J. A. (2010). Reduction in elective delivery at $\leq$ 39 weeks of gestation: Comparative effectiveness of 3 approaches to change and the impact on neonatal intensive care admission and stillbirth. *Am. J. Obstet. Gynecol.* **203** 449.e1–449.e6.

EILERS, P. H. C. and MARX, B. D. (1996). Flexible smoothing with $B$-splines and penalties. *Statist. Sci.* **11** 89–121. MR1435485 https://doi.org/10.1214/ss/1038425655

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

GOLDSTEIN, A., KAPELNER, A., BLEICH, J. and PITKIN, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Statist.* **24** 44–65. MR3328247 https://doi.org/10.1080/10618600.2014.907095

GRAMACY, R. B. and LEE, H. K. H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *J. Amer. Statist. Assoc.* **103** 1119–1130. MR2528830 https://doi.org/10.1198/016214508000000689

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2017). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects. Preprint. Available at arXiv:1706.09523v2.

HASTIE, T. and TIBSHIRANI, R. (2000). Bayesian backfitting. *Statist. Sci.* **15** 196–223. MR1820768 https://doi.org/10.1214/ss/1009212815

HE, J., SAAR, Y. and HAHN, P. R. (2018). Accelerated bayesian additive regression trees. Preprint. Available at arXiv:1810.02215.

HERNÁNDEZ, B., RAFTERY, A. E., PENNINGTON, S. R. and PARNELL, A. C. (2018). Bayesian additive regression trees using Bayesian model averaging. *Stat. Comput.* **28** 869–890. MR3766048 https://doi.org/10.1007/s11222-017-9767-1

HILL, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Statist.* **20** 217–240. MR2816546 https://doi.org/10.1198/jcgs.2010.08162

KORNHAUSER, M. and SCHNEIDERMAN, R. (2010). How plans can improve outcomes and cut costs for preterm infant care. *Manag Care* **19** 28–30.

KRATZ, M. F. (2006). Level crossings and other level functionals of stationary Gaussian processes. *Probab. Surv.* **3** 230–288. MR2264709 https://doi.org/10.1214/154957806000000087

LINERO, A. R. (2018). Bayesian regression trees for high-dimensional prediction and variable selection. *J. Amer. Statist. Assoc.* **113** 626–636. MR3832214 https://doi.org/10.1080/01621459.2016.1264957

LINERO, A. R. and YANG, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 1087–1110. MR3874311 https://doi.org/10.1111/rssb.12293

LOGAN, B. R., SPARAPANI, R., MCCULLOCH, R. E. and LAUD, P. W. (2019). Decision making and uncertainty quantification for individualized treatments using Bayesian additive regression trees. *Stat. Methods Med. Res.* **28** 1079–1093. MR3934636 https://doi.org/10.1177/0962280217746191

MACDORMAN, M. F. and GREGORY, E. C. W. (2015). Fetal and perinatal mortality: United States, 2013. *Natl. Vital Stat. Rep.* **66** 1–24.

MANDUJANO, A., WATERS, T. P. and MYERS, S. A. (2013). The risk of fetal death: Current concepts of best gestational age for delivery. *Am. J. Obstet. Gynecol.* **208** 207.e1–207.e8. https://doi.org/10.1016/j.ajog.2012.12.005

MURASKAS, J. and PARSI, K. (2008). The cost of saving the tiniest lives: NICUs versus prevention. *J. of Ethics* **10** 655–658.

MURRAY, J. S. (2017). Log-linear Bayesian additive regression trees for categorical and count responses. Preprint. Available at arXiv:1701.01503.

PRATOLA, M. T., CHIPMAN, H. A., GATTIKER, J. R., HIGDON, D. M., MCCULLOCH, R. and RUST, W. N. (2014). Parallel Bayesian additive regression trees. *J. Comput. Graph. Statist.* **23** 830–852. MR3224658 https://doi.org/10.1080/10618600.2013.841584

REDDY, U., BETTEGOWDA, V. R. and DIAS, T. (2011). Term pregnancy: A period of heterogeneous risk for infant mortality. *Obstet. Gynecol.* **117** 1279–1287.

SIVAGANESAN, S., MÜLLER, P. and HUANG, B. (2017). Subgroup finding via Bayesian additive regression trees. *Stat. Med.* **36** 2391–2403. MR3660139 https://doi.org/10.1002/sim.7276

SPARAPANI, R. A., LOGAN, B. R., MCCULLOCH, R. E. and LAUD, P. W. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Stat. Med.* **35** 2741–2753. MR3513715 https://doi.org/10.1002/sim.6893

STARLING, J. E., MURRAY, J. S., CARVALHO, C. M., BUKOWSKI, R. K. and SCOTT, J. G. (2020). Supplement to "BART with targeted smoothing: An analysis of patient-specific stillbirth risk." https://doi.org/10.1214/19-AOAS1268SUPPA, https://doi.org/10.1214/19-AOAS1268SUPPB

WAGER, S., HASTIE, T. and EFRON, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* **15** 1625–1651. MR3225243

WATANABE, S. (2013). A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14** 867–897. MR3049492

XU, J., MURPHY, S. L., KOCHANEK, K. D. and BASTIAN, B. A. (2013). Deaths: Final data for 2013. *Natl. Vit. Stats. Rpt.* **64**.