A BAYESIAN MACHINE LEARNING APPROACH FOR ESTIMATING HETEROGENEOUS SURVIVOR CAUSAL EFFECTS: APPLICATIONS TO A CRITICAL CARE TRIAL

BY XINYUAN CHEN[1,a], MICHAEL O. HARHAY[2,b], GUANGYU TONG[3,c] AND FAN LI[3,d]

[1]*Department of Mathematics and Statistics, Mississippi State University,* [a]*xchen@math.msstate.edu*

[2]*Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania,* [b]*mharhay@pennmedicine.upenn.edu*

[3]*Department of Biostatistics, Yale School of Public Health,* [c]*guangyu.tong@yale.edu,* [d]*fan.f.li@yale.edu*

Assessing heterogeneity in the effects of treatments has become increasingly popular in the field of causal inference and carries important implications for clinical decision-making. While extensive literature exists for studying treatment effect heterogeneity when outcomes are fully observed, there has been limited development in tools for estimating heterogeneous causal effects when patient-centered outcomes are truncated by a terminal event, such as death. Due to mortality occurring during study follow-up, the outcomes of interest are unobservable, undefined, or not fully observed for many participants in which case principal stratification is an appealing framework to draw valid causal conclusions. Motivated by the Acute Respiratory Distress Syndrome Network (ARDSNetwork) ARDS respiratory management (ARMA) trial, we developed a flexible Bayesian machine learning approach to estimate the average causal effect and heterogeneous causal effects among the always-survivors stratum when clinical outcomes are subject to truncation. We adopted Bayesian additive regression trees (BART) to flexibly specify separate mean models for the potential outcomes and latent stratum membership. In the analysis of the ARMA trial, we found that the low tidal volume treatment had an overall benefit for participants sustaining acute lung injuries on the outcome of time to returning home but substantial heterogeneity in treatment effects among the always-survivors, driven most strongly by biologic sex and the alveolar-arterial oxygen gradient at baseline (a physiologic measure of lung function and degree of hypoxemia). These findings illustrate how the proposed methodology could guide the prognostic enrichment of future trials in the field.

1. Introduction. Personalized medicine, whereby healthcare is tailored for each individual patient, is the pursuit of contemporary clinical research and practice. For healthcare practitioners and clinicians, achieving this goal hinges upon the successful detection and understanding of the heterogeneity in participants' response to treatment strategies based on their individual characteristics. Capturing factors prognostic of a stronger or weaker response to a trial intervention is especially important in critical care, where conditions such as cardiogenic shock, sepsis, and acute respiratory failure are defined by syndromic criteria such that individuals with the same condition can vary in their biologic and clinical presentation, and thus optimal treatment strategies can vary among clinical populations.

While examination of treatment effect heterogeneity for short-term mortality is difficult, due to the small sample sizes common in critical care trials (Harhay et al. (2014)), this outcome is at least available for all individuals, and recent innovations in statistical learning increasingly permit such examinations (Hahn, Murray and Carvalho (2020), Hill (2011)). In

contrast, the estimation of average treatment effects and conditional average treatment effects for clinically important nonmortality outcomes, such as duration of organ support (e.g., ventilation) or need for intensive care unit or hospital-level care (i.e., length of stay) are more intractable because they are not fully observed or, more generally, said to be "truncated" by the event of death. This is because some participants do not survive to the time point when the nonmortality outcome, such as quality of life, can be measured, or for duration-based outcomes such as length of stay, these outcomes are truncated by the intercurrent event of mortality such that the actual time to hospital discharge cannot be assessed. As a result, for those who do not survive until the end of the study, their nonmortality outcome measure is ambiguous. Though not uncommon, the direct survivors-only analysis can produce selection biases because the truncation by death occurs postrandomization and is often informative (Harhay et al. (2019)).

Our motivating application is the Acute Respiratory Distress Syndrome Network (ARD-SNET) ARMA trial, which was an individually-randomized clinical trial that compared respiratory management during mechanical ventilation with a lower tidal volume ventilator strategy (six mL/kg) vs. a higher tidal volume ventilator strategy (12 mL/kg) for participants suffering from acute lung injury (Brower et al. (2000)). The first primary outcome of the ARMA trial was death before a participant was discharged home and was breathing without assistance, and the second primary outcome was the number of days without ventilator use from day 1 to day 28. As interest in critical care is increasingly focused on longer-term and patient-centered outcomes, we focus our analysis on a slightly longer time horizon (but a highly correlated measure to the second primary outcome) by using the outcome of days to returning home (DTRH). As is the case in other critical care intervention studies, a substantial proportion of the randomized trial participants (34.3%) died before being discharged from the hospital, leading to undefined DTRH outcomes in a third of the trial sample.

In the ARMA trial, one of the few critical care trials that successfully identified a statistically significant treatment effect in the past three decades (Matthay, McAuley and Ware (2017), Tonelli et al. (2014)), the survival status of participants is observed after treatment assignment and is regarded as an intermediate variable. Under the potential outcome framework, Frangakis and Rubin (2002) proposed the principal stratification approach as a framework to define causal effects in the presence of an intermediate variable. In the context of the ARMA trial, the joint potential values of the survival status allow us to classify participants into distinct strata, and the potential outcomes of the nonmortality endpoint are only well defined among the always-survivors (those who are likely in healthier or more treatment-responsive conditions at the time of randomization). Therefore, the survivor average causal effect (SACE) can be considered as an interpretable principal causal effect for nonmortality outcomes (Zhang and Rubin (2003)).

The existing literature for SACE can be largely categorized into two streams. The first stream involves deriving nonparametric large-sample bounds to interval identify the SACE under minimal assumptions, for example, Zhang and Rubin (2003), Imai (2008), Ding et al. (2011), Long and Hudgens (2013), and Yang and Small (2016). However, these bounds are often too wide to be informative for real (i.e., clinical or policy) applications. Beyond interval identification the second stream of literature invokes additional structural and parametric modeling assumptions to identify SACE, for example, Hayden, Pauler and Schoenfeld (2005), Egleston et al. (2006), Zhang, Rubin and Mealli (2009), Chiba and VanderWeele (2011), Frumento et al. (2012), Wang, Zhou and Richardson (2017), and Bia, Mattei and Mercatanti (2022). While convenient to implement, fully parametric modeling necessitates restrictive assumptions that are often challenging to verify. In addition, the bulk of this literature has focused on the average causal effect among the always-survivors and has not branched into understanding how the always-survivors may be deferentially affected by treatment due to their individual characteristics.

In this article we address the goal of estimating the heterogeneous treatment effects among the always-survivors stratum in the ARMA tidal volumes trial using the patient-centered and health-systems relevant DTRH outcome that was informatively truncated by in-hospital death. The target estimand for our new approach is the conditional survivor average causal effect (CSACE), which is defined as the average causal effect for an always-survivor with certain baseline characteristics. Proceeding under the Bayesian principal stratification framework, we relax the parametric modeling assumptions by leveraging the Bayesian additive regression trees (BART) ensemble algorithm (Chipman, George and McCulloch (2010)) for estimating both the stratum membership model as well as the stratum-specific potential outcome models. While several Bayesian nonparametric prior models are successfully adapted for the purpose of causal inference with and without an intermediate variable, for example, Dirichlet process mixture models (Kim et al. (2017, 2019)) and dependent Dirichlet process-Gaussian process prior models (Xu et al. (2016, 2022), Roy, Lum and Daniels (2017)), the BART prior model has gained substantial traction for causal inference due to its computational efficiency and flexibility in modeling complex nonlinear interactions with minimum tuning; see, for example, a comprehensive tutorial by Tan and Roy (2019) and empirical evidence supporting the use of the BART approach for estimating heterogeneous causal effects in different contexts (Bargagli-Stoffi, De Witte and Gnecco (2022), Dorie et al. (2019), Hahn, Murray and Carvalho (2020), Henderson et al. (2020), Hu, Ji and Li (2021), Wendling et al. (2018)). We, therefore, propose to integrate the BART priors into the mixture model framework as a computationally convenient and yet effective approach for principal stratification analysis and use the proposed approach to reanalyze a high-profile critical care trial—the ARMA trial—to quantify treatment effect heterogeneity and identify key effect moderators among the always-survivors in a data-driven fashion. A unique feature of the ARMA trial is that there are baseline covariates that are predictive of the survival status and hence the principal strata membership; this important feature provides a strong basis for investigating treatment effect heterogeneity among the always-survivors.

The remainder of this article is organized as follows. Section 2 provides a concise overview of the principal stratification framework. Section 3 introduces the Bayesian machine learning approach for principal stratification analyses and describes the details of drawing posterior samples for estimation and inference. Section 4 provides a reanalysis of the ARMA trial using the proposed Bayesian machine learning method and identifies key effect moderators. Section 5 offers concluding remarks and discusses future extensions. Additional details are given in the Web Appendix (Chen et al. (2024)).

2. Notation and set up. We consider a two-arm randomized trial with N participants in the setting of nonmortality outcome truncated by death. Let T_i represent the binary treatment for participant i, where $T_i = 1$ if participant i is randomized to treatment and $T_i = 0$ otherwise. Under the potential outcomes framework, we let $Y_i(t)$ represent the nonmortality outcome that would have been observed under treatment assignment $T_i = t \in \{0, 1\}$ and $\{Y_i(1), Y_i(0)\}$ be a pair of potential outcomes for each participant corresponding to the treatment and control conditions. We further define $D_i(t)$ as the potential survival status of participant i at the time that the measurement of the nonmortality outcome (e.g., quality-of-life or DTRH) was taken, with 0 indicating death and 1 indicating being alive. For example, in the analysis of the ARMA trial, we define the survival status at 180 days, which is considered as the maximum DTRH. Similarly, $\{D_i(1), D_i(0)\}$ are a pair of potential survival statuses. In what ensues, we use D_i and Y_i to denote, respectively, the observed survival status and observed nonmortality outcome for participant i. Of note, an alternative set up is to view both DTRH and time-to-death as two time-to-event outcomes under the semicompeting risks framework and consider potential survival status as a function of follow-up time (Comment

et al. (2019), Nevo and Gorfine (2022), Xu et al. (2022)). Although we primarily focus on defining the potential survival status at a specific time point due to clinical relevance and to simplify the exploration of heterogeneity of treatment effect on the nonmortality outcome, we discuss the applicability of the semicompeting risks framework to the ARMA trial at the end of Section 5.

We first make the Stable Unit Treatment Value Assumption (SUTVA). The SUTVA implies that there is one version of the treatment and that there is no interference between participants so that each participant's outcome only depends on the participant's own treatment. Under the SUTVA $D_i = T_i D_i(1) + (1 - T_i) D_i(0)$, and $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ for those who survived until the time that the nonmortality outcome is measured. The nonmortality outcome for those who did not survive ($D_i = 0$) is undefined, and we supplementarily augment the definition of outcome such that $Y_i = *$ (Zhang, Rubin and Mealli (2009)). Using the principal stratification framework (Frangakis and Rubin (2002)), each participant can be classified into one of the distinct principal strata according to the joint values of the potential survival status. Specifically, we have the following four possible stratum memberships:

(a) $S_i = 11$, $\{i | D_i(1) = 1, D_i(0) = 1\}$, always-survivors: participants who would survive to the time of outcome measurement under either treatment status;

(b) $S_i = 10$, $\{i | D_i(1) = 1, D_i(0) = 0\}$, protected: participants who would survive to the time of outcome measurement under treatment but would die before then under control;

(c) $S_i = 01$, $\{i | D_i(1) = 0, D_i(0) = 1\}$, harmed: participants who would die before the time of outcome measurement under treatment but would survive under control;

(d) $S_i = 00$, $\{i | D_i(1) = 0, D_i(0) = 0\}$, never-survivors: participants who would die before the time of outcome measurement under either treatment status.

Since the pair of nonmortality potential outcomes is only well defined among the always-survivors, a common causal estimand of interest is the SACE, defined as

$$\Delta_{\text{SACE}} = \mathbb{E}[Y_i(1) - Y_i(0) | S_i = 11].$$

This principal causal effect is derived by averaging the individual potential outcomes contrasts over the population of always-survivors and serves as the basis for concluding effectiveness regarding the treatment without ambiguity in defining the potential outcomes. Assuming $X_i = x$ is the baseline characteristics of individual i, we are additionally interested in the CSACE, defined as

$$\Delta_{\text{CSACE}}(x) = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = x, S_i = 11], \tag{1}$$

which quantifies the conditional treatment effect, given certain baseline characteristics of an always-survivor who would live to the time of outcome measurement regardless of treatment assignment. Variations in $\Delta_{\text{CSACE}}(x)$ measure the degree of treatment effect heterogeneity among the always-survivors and may provide useful evidence for tailoring treatment rules for future participants. Deng et al. (2021) discussed identification strategies for CSACE under truncation by death but under more restrictive conditions, such as principal ignorability (Ding and Lu (2017)). In this article we provide an estimation approach that does not invoke principal ignorability and only requires the following two standard structural assumptions.

ASSUMPTION 1 (Randomization). The assignment T_i is independent of all potential outcomes $\{D_i(1), D_i(0), Y_i(1), Y_i(0)\}$, given baseline characteristics X_i.

ASSUMPTION 2 (Monotonicity). $\mathbb{P}(D_i(1) \geq D_i(0) | X_i = x) = 1$, $\forall x \in \mathcal{X}$, where $\mathcal{X}$ is the support of X.

Assumption 1 is essentially an ignorability assumption and holds by design in a randomized trial. However, it is more general and can be satisfied in stratified randomized studies as well as observational studies as long as X_i captures a sufficient set of control variables. Assumption 2 states that the treatment does not lead to poor survival and rules out the harmed stratum. This assumption is often considered plausible in studies where a treatment is designed to improve the general well-being of participants, as in our motivating application. Under Assumption 2, trial participants belong to one of the three strata of always-survivors, protected, or never-survivors, and depending on the observed treatment status, only a fraction of participants have unobserved stratum membership. In other words, survivors in the treatment arm can be either always-survivors or protected; nonsurvivors in the control arm can be either never-survivors or protected. Assumption 2 may be violated when, for example, in a comparative effectiveness trial where two active treatments with unknown relative benefits are studied. In that case it is of interest to extend our approach along the lines of Zhang, Rubin and Mealli (2009) by incorporating the harmed stratum at the expense of reduced precision and algorithm stability. We return to a discussion of this point in Section 5.

3. A Bayesian machine learning approach for estimating CSACE.

3.1. *Bayesian principal stratification.*

We consider the Bayesian principal stratification framework (Hirano et al. (2000), Mattei, Li and Mealli (2013), Mattei and Mealli (2007)) in which one is required to specify two sets of models: the distribution of potential outcomes $Y(0)$ and $Y(1)$, conditional on the principal strata and covariates (the Y-model), and the distribution of principal strata conditional on the covariates (the S-model). Let $\boldsymbol{\theta}$ generically denote the global parameters, and for participant i, we use $X_{Y,i}$ and $X_{S,i}$ to denote respective vectors of covariates for the Y- and S-model, with $X_i = (X'_{Y,i}, X'_{S,i})'$. According to their treatment assignments and survival status at the time of final outcome measurement, we can reclassify each participant into the following categories:

(a) $\mathcal{O}(1, 1) = \{i | T_i = 1, D_i = 1\}$, participants assigned to the treatment arm and survived;
(b) $\mathcal{O}(1, 0) = \{i | T_i = 1, D_i = 0\}$, participants assigned to the treatment arm and died;
(c) $\mathcal{O}(0, 1) = \{i | T_i = 0, D_i = 1\}$, participants assigned to the control arm and survived;
(d) $\mathcal{O}(0, 0) = \{i | T_i = 0, D_i = 0\}$, participants assigned to the control arm and died.

Stratum memberships for participants in $\mathcal{O}(1, 0)$ and $\mathcal{O}(0, 1)$ are then fully inferred under the monotonicity assumption, which are denoted by S_i^{obs}. We use S^{obs} to denote the collection of S_i^{obs}'s. On the other hand, for participants in $\mathcal{O}(1, 1)$ and $\mathcal{O}(0, 0)$, their stratum memberships cannot be determined directly and are thus labeled as S_i^{mis}. Denote $\pi_{s,i} = \mathbb{P}(S_i = s | X_{S,i}, \boldsymbol{\theta})$ and $f_{st,i} = \mathbb{P}(Y_i(t) | S_i = s, X_{Y,i}, \boldsymbol{\theta})$, for $s = 00, 10, 11$ and $t = 0, 1$, and assume a prior distribution $\mathbb{P}(\boldsymbol{\theta})$ for the parameters $\boldsymbol{\theta}$. The posterior distribution of $\boldsymbol{\theta}$ can be generically written as

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{\theta}|Y, S^{\mathrm{obs}}, T, X) \propto \mathbb{P}(\boldsymbol{\theta}) \times & \prod_{i \in \mathcal{O}(1,1)} (\pi_{11,i} f_{111,i} + \pi_{10,i} f_{101,i}) \times \prod_{i \in \mathcal{O}(1,0)} \pi_{00,i} \\
\times & \prod_{i \in \mathcal{O}(0,1)} \pi_{11,i} f_{110,i} \times \prod_{i \in \mathcal{O}(0,0)} (\pi_{10,i} + \pi_{00,i}).
\end{aligned}
$$

(2)

3.2. *Model specification.*

Posterior inference on $\boldsymbol{\theta}$ from (2) is achieved using data augmentation to impute missing stratum membership S_i^{mis}, which can be performed via a nested Probit modeling approach. We introduce two additional latent variables Z and W to be aug-

mented for each participant, where

$$\{Z_i | m_Z(\bullet), X_{S,i}\} \sim \mathcal{N}(m_Z(X_{Z,i}), 1) \quad \text{and} \quad \begin{cases} S_i = 00 & \text{if } Z_i > 0, \\ S_i = 10 \text{ or } 11 & \text{if } Z_i \leq 0, \end{cases}$$

(3)

$$\{W_i | m_W(\bullet), X_{S,i}\} \sim \mathcal{N}(m_W(X_{W,i}), 1) \quad \text{and} \quad \begin{cases} S_i = 10 & \text{if } W_i > 0, \\ S_i = 11 & \text{if } W_i \leq 0. \end{cases}$$

Here $m_Z(\bullet)$ and $m_W(\bullet)$ are conditional mean functions for Z_i and W_i that can be fully specified by corresponding parameters, and $X_{Z,i}$ and $X_{W,i}$ are vectors of covariates that are subsets of $X_{S,i}$ with possible overlapping elements. Based on (3), the conditional probability of stratum membership for each participant can be expressed as

$$\mathbb{P}(S_i = 00 | m_Z(\bullet), X_{S,i}) = 1 - \Phi(m_Z(X_{Z,i})),$$

$$\mathbb{P}(S_i = 10 | m_Z(\bullet), m_W(\bullet), X_{S,i}) = \Phi(m_Z(X_{Z,i}))\{1 - \Phi(m_W(X_{W,i}))\},$$

$$\mathbb{P}(S_i = 11 | m_Z(\bullet), m_W(\bullet), X_{S,i}) = \Phi(m_Z(X_{Z,i}))\Phi(m_W(X_{W,i})),$$

where $\Phi(\bullet)$ is the cumulative distribution function of a standard normal random variable. Connecting with the notation in (2), we have $\pi_{00,i} = 1 - \Phi(m_Z(X_{Z,i}))$, $\pi_{10,i} = \Phi(m_Z(X_{Z,i}))\{1 - \Phi(m_W(X_{W,i}))\}$, and $\pi_{11,i} = \Phi(m_Z(X_{Z,i}))\Phi(m_W(X_{W,i}))$.

For the Y-models, we specify the three sets of potential outcome models as

$$\{Y_i(t) | S_i = s, m_{st}(\bullet), X_{Y,i}\} \sim \mathcal{N}(m_{st}(X_{st,i}), \sigma_{st}^2),$$

where $t = 0, 1$ for $s = 11$, and $t = 1$ for $s = 10$; $m_{st}(\bullet)$ are conditional mean functions for $Y_i(t)$ with $X_{st,i}$ being vectors of covariates that are subsets of $X_{Y,i}$ with possible overlapping elements, and σ_{st}^2 is the variance parameter that depends on the principal strata and the treatment status. Similar to the conditional mean functions in the S-model, $m_{st}(\bullet)$ are also fully specified by corresponding parameters. To summarize, we have the following:

$$\{Y_i(1) | S_i = 11, m_{111}(\bullet), X_{Y,i}\} \sim \mathcal{N}(m_{111}(X_{111,i}), \sigma_{111}^2),$$

(4)

$$\{Y_i(0) | S_i = 11, m_{110}(\bullet), X_{Y,i}\} \sim \mathcal{N}(m_{110}(X_{110,i}), \sigma_{110}^2),$$

$$\{Y_i(1) | S_i = 10, m_{101}(\bullet), X_{Y,i}\} \sim \mathcal{N}(m_{101}(X_{101,i}), \sigma_{101}^2).$$

Based on the specification in (4), the SACE can be estimated as

$$(5) \qquad \widehat{\Delta}_{\text{SACE}} = \iint \left[\int_{\mathcal{X}} \{m_{111}(X_{111}) - m_{110}(X_{110})\} f(X | S = 11) \mu(dX) \right] p_m p_S,$$

where the outer double integration is taken with respect to p_m—the posterior distributions of parameters in $m_{111}(\bullet)$ and $m_{110}(\bullet)$, and p_S—the posterior distribution of the principal strata membership. In addition, the CSACE evaluated at $X = x$ can be estimated by

$$(6) \qquad \widehat{\Delta}_{\text{CSACE}}(x) = \int \{m_{111}(x) - m_{110}(x)\} p_m.$$

Implicitly in the notation of (5) and (6), we consider super-population inference (rather than finite-sample inference), where the causal estimand is represented by model parameters governing the joint distribution of potential outcomes. Under this framework the posterior distribution of $m_{111}(\bullet)$ and $m_{110}(\bullet)$ does not involve the correlation parameter between $Y_i(1)$ and $Y_i(0)$ (as the observed data likelihood is free of this correlation parameter), and, therefore, it is sufficient to specify the marginal distributions of potential outcomes; also, see Section 3 of Ding and Li (2018) for a detailed discussion on this issue. In Web Appendix A4, we also conduct an additional analysis of the ARMA trial to infer the finite-sample SACE estimand,

varying the correlation between $Y_i(1)$ and $Y_i(0)$ as a sensitivity parameter. The results are almost identical to the super-population analysis.

According to (5) and (6), a central task in the estimation of the SACE and CSACE is to specify the conditional mean functions in the models. Typically, we assume that the parameters in these models are a priori independent and proceed with conjugate diffuse priors. For example, a straightforward specification for the conditional mean functions can be achieved via parametric linear models such that $m_Z(X_{Z,i}) = X'_{Z,i}\beta_Z$, $m_W(X_{W,i}) = X'_{W,i}\beta_W$, and $m_{st}(X_{st,i}) = X'_{st,i}\alpha_{st}$. Then a closed-form Gibbs sampler can be derived with multivariate Gaussian prior assumed for linear coefficients, β_Z, β_W, and α_{st}. A detailed derivation of this Gibbs sampler is provided in Web Appendix A1. This fully parametric specification, however, can result in potential biases for estimating the SACE and CSACE when the true mean functions are nonlinear and with possibly unknown functional forms. An illustration of the bias resulting from model misspecification is provided in Web Appendix A2 using simulated data sets.

3.3. *Integrating Bayesian additive regression trees into principal stratification.* To address the potential limitations of fully parametric models, we propose to use the Bayesian additive regression trees (BART) to estimate the mean functions nonparametrically. BART is an ensemble method in which the mean function of a regression is approximated by the sum of individual trees, with prior distributions imposed to regularize the fit by keeping the individual tree effects relatively small (Chipman, George and McCulloch (2010)). Specifically, let $\mathcal{T}$ denote a binary tree consisting of a set of interior node decision rules and a set of terminal nodes, and let $\mathcal{M} = \{\mu_1, \mu_2, \ldots, \mu_b\}$ denote a set of parameter values associated with each of the b terminal nodes of $\mathcal{T}$. The BART formulation of the mean function relies on a collection of J binary trees $\{\mathcal{T}_1, \ldots, \mathcal{T}_J\}$ and their, respectively, associated set of terminal node values $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ for each binary tree, where $\mathcal{M}_j = \{\mu_{j1}, \mu_{j2}, \ldots, \mu_{jb_j}\}$. Each tree $\mathcal{T}_j$ consists of a sequence of decision rules through which any covariate vector can be assigned to one terminal node of $\mathcal{T}_j$ by following the decision rules prescribed at each of the interior nodes. The decision rules at the interior nodes of $\mathcal{T}_j$ are of the form $\{X_k \leq c\}$ vs. $\{X_k > c\}$, where X_k denotes the kth element of X. A covariate X that corresponds to the lth terminal node of $\mathcal{T}_j$ is assigned the value μ_{jl} and $h(X; \mathcal{T}_j, \mathcal{M}_j)$ is used to denote the function returning $\mu_{jl} \in \mathcal{M}_j$ whenever X is assigned to the lth terminal node of $\mathcal{T}_j$. The mean function of a generic regression model, $m(X)$, can thus be represented as a sum of individual trees

$$m(X) = \sum_{j=1}^{J} h(X; \mathcal{T}_j, \mathcal{M}_j).$$

Under the BART formulation, the trees $\mathcal{T}_j$ and node values $\mathcal{M}_j$ can be thought of as model parameters. The prior distribution on these parameters induces a prior on $h(X; \mathcal{T}_j, \mathcal{M}_j)$ and hence induces a prior on the mean function $m(X)$. To proceed, one needs to specify the following to complete the description of the prior on $(\mathcal{T}_1, \mathcal{M}_1), \ldots, (\mathcal{T}_J, \mathcal{M}_J)$: (i) the distribution on the choice of splitting variable at each internal node, (ii) the distribution of the splitting value c used at each internal node, (iii) the probability that a node at a given node-depth δ splits, which is assumed to be equal to $\tau(1 + \delta)^{-\gamma}$, and (iv) the distribution of the terminal node values μ_{jl}. Regarding (i)–(iii), we defer to defaults suggested in Chipman, George and McCulloch (2010), where, for (i), the splitting variable is chosen uniformly from the set of available splitting variables at each interior node; for (ii), a uniform prior on the discrete set of available splitting values is adopted; for (iii), the depth-related hyperparameters are chosen as $\tau = 0.95$ and $\gamma = 2$. For (iv) the distribution of the terminal node values μ_{jl} is assumed to be $\mu_{jl} \sim \mathcal{N}\{0, (4w^2 J)^{-1}\}$, where w and J are determined via cross-validation,

as we further elaborate in Section 4. To denote the distribution on the regression function $m(\bullet)$ induced by the prior distribution on $\mathcal{T}_j$ and $\mathcal{M}_j$ with parameter values (τ, γ, w) and J total trees, we use the notation $m(\bullet) \sim \text{BART}(\tau, \gamma, w, J)$. Using BART, the mean functions under the Bayesian principal stratification framework can be expressed as

$$m_Z(X_{Z,i}) = \sum_{j=1}^{J_Z} h_Z(X_{Z,i}; \mathcal{T}_{Z,j}, \mathcal{M}_{Z,j}),$$

(7) $$m_W(X_{W,i}) = \sum_{j=1}^{J_W} h_W(X_{W,i}; \mathcal{T}_{W,j}, \mathcal{M}_{W,j}),$$

$$m_{st}(X_{st,i}) = \sum_{j=1}^{J_{st}} h_{st}(X_{st,i}; \mathcal{T}_{st,j}, \mathcal{M}_{st,j}), \quad st \in \{111, 110, 101\},$$

each with the common prior distribution $\text{BART}(\tau, \gamma, w, J)$ that is assumed to be a priori independent of each other; here $m_Z(\bullet)$, $m_W(\bullet)$ stand for the mean functions of the stratum membership model; $m_{st}(\bullet)$ stands for the mean function of the potential outcome model. Essentially, our semiparametric model is a mixture of BART, with the mixture weights represented by a nested Probit BART model.

It is worth mentioning that with only Assumptions 1 and 2 alone, the SACE and CSACE estimands are only partially or set identified (Kadane (1975)); see, for example, the large sample bounds developed in Table 6 of Zhang and Rubin (2003) and discussions of a similar issue arising from the noncompliance context in the Appendix of Hahn, Murray and Manolopoulou (2016). Resembling the approach of Hirano et al. (2000) for studying noncompliance, the parametric mixture approach in Section 3.2 resolves the partial identification issue through prior probability modeling so that the inferential procedure is assisted by a Gaussian mixture model. The proposed approach considers the same prior probability modeling idea to resolve partial identification issue but incorporates the BART priors for mean functions (still within a Gaussian mixture model) to more flexibly study effect moderation. Although not pursued in the current work, an alternative and powerful approach to address partial identification can follow Hahn, Murray and Manolopoulou (2016) to transparently separate identified and unidentified model components with careful prior specifications for each component. Finally, we note that the proposed approach for imposing monotonicity differs from the recent work of Papakostas et al. (2023). Whereas Papakostas et al. (2023) adopted BART priors for components of a compositional representation to enforce a stochastic monotonicity constraint (in the absence of intermediate variables), we adopted BART priors for mean functions in a Gaussian mixture model where the mixture itself already incorporates a structural monotonicity constraint. Finally, we emphasize that our mixture modeling approach is particularly suitable for studying CSACE when there exist baseline covariates that are predictive of the principal strata membership. This is the case in the ARMA trial and allows us to meaningfully describe the subset of always-survivors, as we exemplify in Section 4.3.1.

3.4. *Posterior computation.* For posterior inference we develop a Gibbs sampling procedure based upon the original Metropolis-within-Gibbs sampler proposed in Chipman, George and McCulloch (2010), which works by sequentially updating each tree while holding all other $J - 1$ trees fixed. As a result, each iteration of the Gibbs sampler consists of $2J + 1$ steps where the first $2J$ steps involve updating either one of the trees $\mathcal{T}_j$ or terminal node parameters $\mathcal{M}_j$ and the last step involves updating the residual variance parameter. In each iteration we first update values of Y-model parameters by sampling from their respective full

conditional posterior distributions and follow up by updating S-model parameters via sampling from related full conditional posterior distributions. Unobserved stratum memberships and additional latent variables, Z and W, in the S-model are handled through additional data augmentation steps (Albert and Chib (1993)). In essence each component BART model can be updated separately because: (i) an independent BART prior is assumed for each component model, (ii) the full conditional distributions for each component Y-model only depend on the observed potential outcomes, augmented stratum memberships, covariate vector, and residual variances as well as relevant prior distributions, and (iii) the full conditional distributions for each component S-model only depend on the augmented latent variable, covariate vector, and relevant prior distributions. A detailed outline of the sampling procedure is as follows:

1. Update trees $\{\mathcal{T}_{st,1}, \ldots, \mathcal{T}_{st,J_{st}}\}$ and node parameters $\{\mathcal{M}_{st,1}, \ldots, \mathcal{M}_{st,J_{st}}\}$ in the Y-model via the Bayesian backfitting approach of Chipman, George and McCulloch (2010), using $Y_i(T_i)$ with $T_i = t$ in strata $S_i = s$ as responses. Full conditional distributions of trees and node parameters in the Y-model depend on the observed potential outcomes, $Y_i(T_i)$, latent stratum memberships, S_i, covariate vector $X_{st,i}$, residual variances, σ_{st}^2, and prior distributions of trees and node parameters. Update values of the mean functions, $m_{st}(X_{st,i})$ for $i = 1, \ldots, n$, using the updated $\{\mathcal{T}_{st,1}, \ldots, \mathcal{T}_{st,J_{st}}\}$ and $\{\mathcal{M}_{st,1}, \ldots, \mathcal{M}_{st,J_{st}}\}$, where $t = 0, 1$ for $s = 11$, and $t = 1$ for $s = 10$. These mean function values are needed for the update of residual variances.

2. Update variance parameters in each Y-model. Assuming a conjugate inverse Gamma prior distribution $\text{IG}(a_0, b_0)$ for residual variance, σ_{st}^2 with $t = 0, 1$ for $s = 11$ and $t = 1$ for $s = 10$, we update σ_{st}^2 from its posterior inverse Gamma distribution, $\text{IG}(a_{st}^*, b_{st}^*)$, with $a_{st}^* = a_0 + (1/2) \sum_{i:S_i=s, T_i=t} D_i$, $b_{st}^* = b_0 + (1/2) \sum_{i:S_i=s, T_i=t} \mathbb{I}(T_i = t)\{Y_i - m_{st}(X_{st,i})\}^2$. Here (a_0, b_0) and (a^*, b^*) are shape and rate parameters for the prior and full conditional posterior distribution of σ_{st}^2, respectively. The full conditional of σ_{st}^2 only depends on the stratum-treatment group size, potential outcomes in that group, and the updated mean functions.

3. Update trees $\{\mathcal{T}_{Z,1}, \ldots, \mathcal{T}_{Z,J_Z}\}$ and node parameters $\{\mathcal{M}_{Z,1}, \ldots, \mathcal{M}_{Z,J_Z}\}$ for the S-model, using Z_i as responses. Full conditional distributions of trees and node parameters in the Z-submodel depend on the latent variable, Z_i, covariate vector $X_{Z,i}$, and prior distributions of trees and node parameters. Update mean functions $m_Z(X_{Z,i})$ for $i = 1, \ldots, n$, using the updated $\{\mathcal{T}_{Z,1}, \ldots, \mathcal{T}_{Z,J_Z}\}$ and $\{\mathcal{M}_{Z,1}, \ldots, \mathcal{M}_{Z,J_Z}\}$. These mean function values are needed for the sampling of Z_i during data augmentation.

4. Update trees $\{\mathcal{T}_{W,1}, \ldots, \mathcal{T}_{W,J_W}\}$ and node parameters $\{\mathcal{M}_{W,1}, \ldots, \mathcal{M}_{W,J_W}\}$ for the S-model, using W_i as responses with $S_i = 10, 11$. Similar to the previous step, full conditional distributions of trees and node parameters in the W-submodel depend on the latent variable, W_i, latent stratum memberships, S_i, covariate vector $X_{W,i}$, and prior distributions of trees and node parameters. Update $m_W(X_{W,i})$ for $i = 1, \ldots, n$, using the updated $\{\mathcal{T}_{W,1}, \ldots, \mathcal{T}_{W,J_W}\}$ and $\{\mathcal{M}_{W,1}, \ldots, \mathcal{M}_{W,J_W}\}$. These mean function values are needed for the sampling of W_i during data augmentation.

5. Update the stratum membership, S_i, for each participant. By Assumption 2, latent stratum memberships of certain subgroups of participants are directly ascertained, while the remaining participants require an update through data augmentation. Specifically:

    (a). If $T_i = 1$ and $D_i(1) = 0$, then $S_i = 00$.
    (b). If $T_i = 0$ and $D_i(0) = 1$, then $S_i = 11$.
    (c). If $T_i = 0$ and $D_i(0) = 0$, then

$$p_{00,i} = \mathbb{P}(S_i = 00|\bullet) = \Phi(m_Z(X_{Z,i})),$$
$$p_{10,i} = \mathbb{P}(S_i = 10|\bullet) = \{1 - \Phi(m_Z(X_{Z,i}))\}\Phi(m_W(X_{W,i})).$$

Generate $\epsilon_i \sim \text{Bern}(p_{00,i}/(p_{00,i} + p_{10,i}))$. If $\epsilon_i = 1$, set $S_i = 00$; if $\epsilon_i = 0$, set $S_i = 10$. Here full conditional probabilities, $p_{00,i}$ and $p_{10,i}$, do not involve potential outcomes, because the potential outcomes are not well defined for participants assigned to the control arm ($T_i = 0$) that died ($D_i(0) = 0$).

(d). If $T_i = 1$ and $D_i(1) = 1$, then

$$p_{10,i} = \mathbb{P}(S_i = 10|\bullet) = \Phi(m_W(X_{W,i}))\phi(Y_i; m_{101,i}(X_{101,i}), \sigma_{101}^2),$$

$$p_{11,i} = \mathbb{P}(S_i = 11|\bullet) = \{1 - \Phi(m_W(X_{W,i}))\}\phi(Y_i; m_{111,i}(X_{111,i}), \sigma_{111}^2),$$

where $\phi(y; m, \sigma^2)$ denotes the normal density with response y, mean m, and variance σ^2. Generate $\epsilon_i \sim \text{Bern}(p_{10,i}/(p_{10,i} + p_{11,i}))$. If $\epsilon_i = 1$, set $S_i = 10$; if $\epsilon_i = 0$, set $S_i = 11$. Here potential outcomes are observable for participants assigned to the treatment arm ($T_i = 1$) that survived ($D_i(1) = 1$) such that full conditional probabilities, $p_{10,i}$ and $p_{11,i}$, depend on Y_i.

6. Update latent variable Z_i for each participant by sampling from their truncated normal full conditional distributions,

$$\{Z_i|m_Z(\bullet), X_{Z,i}, S_i\} \sim \mathcal{N}(m_Z(X_{Z,i}), 1)\mathbb{I}(Z_i \geq 0) \quad \text{if } S_i = 00,$$

$$\{Z_i|m_Z(\bullet), X_{Z,i}, S_i\} \sim \mathcal{N}(m_Z(X_{Z,i}), 1)\mathbb{I}(Z_i < 0) \quad \text{if } S_i = 10 \text{ or } 11.$$

This is a data augmentation step that identifies participants in the never-survivor stratum in a specific iteration.

7. Update latent variable W_i for each participant in stratum 10 and 11 by sampling from their truncated normal full conditional distributions

$$\{W_i|m_W(\bullet), X_{W,i}, S_i\} \sim \mathcal{N}(m_W(X_{W,i}), 1)\mathbb{I}(W_i \geq 0) \quad \text{if } S_i = 10,$$

$$\{W_i|m_W(\bullet), X_{W,i}, S_i\} \sim \mathcal{N}(m_W(X_{W,i}), 1)\mathbb{I}(W_i < 0) \quad \text{if } S_i = 11.$$

This is an additional data augmentation step that further identifies participants in the always-survivor stratum in a specific iteration.

We initialize the proposed Gibbs sampler by first assigning participants to the three strata, where participants with directly identifiable stratum memberships are assigned directly, and those with stratum memberships that are not directly identifiable are randomly assigned to one of the possible strata according to their received treatments and survival status. For mean functions in the Y-model, $m_{st}(\bullet)$, initial estimates can be obtained using the BART model, for example, using the *bart* function from *R* package *dbarts*, given that all stratum memberships are fixed and the initial estimate for σ_{st}^2 can be simultaneously obtained from the fitting this initial BART model. For mean functions in the S-model, $m_Z(\bullet)$ and $m_W(\bullet)$, we fit parametric logistic regression models using indicators converted from initial stratum membership assignments and associated covariates; the resulting linear components are used as initial values for $m_Z(X_{Z,i})$ and $m_W(X_{W,i})$. Initial values for Z_i's and W_i's are generated from truncated normal distributions, conditional on initial stratum membership assignments as well as initial estimates of $m_Z(X_{Z,i})$ and $m_W(X_{W,i})$ for all i. Hyperparameters of the inverse Gamma prior for σ_{st}^2 are chosen as $a_0 = b_0 = 0.001$. The SACE and CSACE (for fixed value $X = x$) can then be calculated at each iteration of the Gibbs sampler, and the respective posterior distributions can be obtained after the sampling procedure is terminated. *R* code for implementing this Bayesian approach can be found in the Supplementary Material (Chen et al. (2024)) and at https://github.com/erxc/BART-SACE-HTE.

To compare the proposed Bayesian machine learning approach with its parametric counterparts in estimating the SACE and CSACE, we conducted simulations under two data generating processes and with two levels of total sample sizes. We find that: (i) the proposed

approach, where all mean functions in the Y- and S-model were specified nonparametrically using BART, outperformed parametric approaches in the estimation of both the SACE and CSACE, as measured by relative bias, root mean squared error and the overall precision in the estimation of heterogeneous effects (Hill (2011), Hu, Ji and Li (2021)) and that (ii) for estimating CSACE, the Y-model appears to play a more important role because using BART only in Y-model dominates using BART only in S-model in terms of both bias and efficiency. The details of the simulation study are presented in Web Appendix A2.

4. Application to the ARDSNetwork ARMA study.

4.1. *Data.* The ARMA trial involved 861 participants with acute lung injury and acute respiratory distress syndrome who were randomized to receive mechanical ventilation with a volume of 12 mL per kilogram of predicted body weight ($T_i = 0$) or a lower tidal volume ventilator strategy of six mL per kilogram of predicted body weight ($T_i = 1$). We focus our analysis on the patient-centered nonmortality outcome variable DTRH with 180 days as the maximum for those who survived; correspondingly, the principal strata based on potential survival status are defined at 180 days. The DTRH captures important information to payer and health system stakeholders as a measure of health care utilization and to patients and their caregivers, as it is associated with patients' long-term prognosis and health-related quality of life. We use Y_i to denote the nonmortality outcome of participant i. In-hospital death events occurred in a substantial proportion (34.3%) of enrolled trial participants, and more deaths were observed in the usual care group ($173/429 = 40.3\%$) than those in the treatment group ($146/473 = 30.9\%$) resulting in an absolute risk difference of -9.4%. The study was motivated based on concerns that mechanical ventilation treatment using traditional tidal volumes of 10 to 15 mL per kilogram of body weight may cause stretch-induced lung injury in those with acute lung injury and acute respiratory distress syndrome (Brower et al. (2000)). We assume monotonicity such that the lower tidal volume does not lead to worse survival and hence excludes the harmed stratum. We also exclude four participants who had one or more missing covariates. Summary statistics (means) of the nonmortality DTRH outcome and baseline covariates for the total sample of 857 enrolled participants by treatment arm T_i and survival status D_i are presented in Table 1.

Based on discussions with clinical colleagues, we pre-selected 18 covariates that were measured at baseline, which can be broadly divided into three groups: (i) demographic information, including age, sex, and race/ethnicity, (ii) respiratory measures including tidal volume in milliliter, positive end-expiratory pressure (PEEP) in centimeter water, fraction of inspired oxygen (FiO$_2$) in millimeter Hg, partial pressure of arterial carbon dioxide (PaCO$_2$) in millimeter Hg, partial pressure of arterial oxygen (PaO$_2$) in millimeter Hg, the ratio of PaO$_2$ to FiO$_2$ (PtoF), the first alveolar-arterial oxygen gradient (AaDO$_2$), arterial pH, and (iii) physiological measures including the score of acute physiology, age, and chronic health evaluation (APACHE III), in addition to the Glasgow coma scale score (Glasgow) as a measure of central nervous system failure, platelet count per nanoliter as a measure of coagulation, serum creatine in milligram per deciliter as a measure of renal function, bilirubin in milligram per deciliter as a measure of hepatic function, the use of vasopressors (indicating the need for blood pressure support), and systolic blood pressure (systolic BP) in millimeter Hg. Due to randomization, the baseline characteristics are comparable across different treatment groups. A descriptive comparison of outcomes among survivors suggests that participants receiving low tidal volume treatment appear to have shorter DTRH (average 44.80 days under low tidal volume treatment vs. 47.94 days under the higher tidal volume ventilator strategy). To analyze DTRH outcomes subject to death truncation, we apply our proposed Bayesian machine learning method to estimate both the SACE and CSACE.

TABLE 1
Summary statistics of the key variables (means for numerical variables and proportions for binary indicators) in the ARMA trial

	All	$T_i = 1, D_i = 1$	$T_i = 1, D_i = 0$	$T_i = 0, D_i = 1$	$T_i = 0, D_i = 0$
Sample size	857	303	127	260	167
DTRH (Y_i)	–	44.80	–	47.94	–
Age	51.45	49.80	51.35	51.70	55.08
Sex (female) (%)	0.41	0.36	0.43	0.43	0.44
Race/ethnicity(%)					
White	0.73	0.75	0.70	0.70	0.77
Non-White	0.27	0.25	0.30	0.30	0.23
Tidal volume (ml)	679.95	683.91	684.05	679.07	673.81
PEEP (cm water)	8.40	8.33	9.05	8.11	8.50
PaO_2 (mm Hg)	84.83	85.21	81.39	84.54	87.19
FiO_2 (mm Hg)	0.63	0.61	0.66	0.61	0.68
$PaCO_2$ (mm Hg)	36.32	36.43	36.08	37.18	34.99
PaO_2/FiO_2 (PtoF)	149.00	155.46	135.50	153.24	140.98
$AaDO_2$	325.17	307.06	352.28	309.31	362.13
Arterial pH	7.40	7.40	7.39	7.41	7.39
APACHE III	82.52	76.04	92.87	76.95	95.11
Systolic BP	97.88	100.55	94.89	99.80	92.34
Glasgow coma scale	11.10	11.37	10.74	11.14	10.80
Platelet (count/nl)	109.94	118.05	89.31	115.57	102.16
Creatine (mg/dl)	1.18	1.10	1.15	1.14	1.40
Bilirubin (mg/dl)	0.88	0.99	0.70	0.77	0.98
Vasopressors (%)	0.65	0.74	0.57	0.71	0.47

4.2. *Implementation.* With the set of baseline covariates, we considered model (4) for the potential outcomes. We standardized all continuous covariates to have zero mean and unit variance to improve the numerical stability. We implemented the Gibbs sampling procedure, described in Section 3, by specifying the following priors. For the BART priors, the distribution on the choice of the splitting value at each internal node, the distribution of splitting value used at each internal node, and the probability that a node at given node-depth splits remained the same, as previously described. For the distribution of terminal node values, we considered a five-fold cross-validation based on the set of $w \in \{1, 2, 3, 4\}$ and $J \in \{50, 75, 100, 200\}$ (recall that w and J control the variance of the prior for the node values and the total number of trees) and found that $w = 4$ and $J = 50$ are associated with the best predictive performance of the outcome and thus were adopted to generate our analytical results. Cross-validation results are shown in Web Figure A1 in Web Appendix A3. We set $a_0 = b_0 = 0.001$ for the Gamma priors of the variances. We ran the Markov chain Monte Carlo procedure for 10,000 iterations and used the first 5000 as burn-in. We obtained point estimates along with corresponding 95% credible intervals of the SACE and CSACE based on draws from their respective posterior samples. The SACE here captures the average DTRH reduced under the low tidal volume treatment, compared to the traditional volume treatment for the trial participants who classify as always-survivors.

4.3. *Survivor average causal effect and its conditional counterparts.*

4.3.1. *Who would likely be the always-survivors?.* To characterize the CSACE, we first capture a trial subpopulation that is mostly likely being always-survivors. Because the stratum membership is not fully observed for all participants, capturing the subset of likely

always-survivors is a practical decision that makes it feasible to study effect moderation on the nonmortality outcome. Under monotonicity, we differentiate the observed trial sample as follows:

1. $\mathbb{S}_1$: Subset of participants who are observed to be always-survivors; this set of participants are precisely those assigned to the treatment of traditional tidal volume and survived until the end of study.

2. $\mathbb{S}_0$: Subset of participants who are observed to be never-survivors; this set of participants are precisely those assigned to the treatment of low tidal volume but died;

3. $\mathbb{S}_{(p,1)}$: Subset of participants who are not in $\mathbb{S}_1$ or $\mathbb{S}_0$ but have a posterior probability of at least p ($0 < p < 1$) to be always-survivors; this set of participants will be among those assigned to the treatment of traditional tidal volume but died prior to the end of study or those assigned to the treatment of low tidal volume but survived until the end of study.

In the ARMA trial, 30.3%($= 260/857$) of participants assigned to the treatment of traditional tidal volume survived until the day 180, while the posterior mean of the marginal proportion of always-survivors is estimated be to 60.9%. This motivated us to consider using the set $\mathbb{S}_1 \cup \mathbb{S}_{(p,1)}$ with $p = 0.8$ to approximate the set of always-survivors. This choice of p is given such that the proportion of this set matches the posterior mean of the marginal proportion of always-survivors returned by our chain. We then primarily focused on interpreting CSACE for participants who have at least 80% posterior probability to belong to the always-survivor stratum. We further compared the posterior mean of $\bar{\Delta}_{\text{CSACE}} = N_{11}^{-1} \sum_{i=1}^{N_{11}} \Delta_{\text{CSACE}}(X_i)$, where $N_{11} = |\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}|$, with the posterior mean of SACE, and found that they were identical. In particular, the posterior mean of SACE is -23.87 days and $\bar{\Delta}_{\text{CSACE}} = -23.87$ days. This post hoc check ensures that $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$ is a reasonable approximation to the latent, always-survivor stratum and confirms that the low tidal volume treatment leads to, on average, 24 days (95% credible interval, 16.7–30.9 days) in reductions on DTRH among the always-survivors. That is, low tidal volume treatment led to substantial benefits regarding DTRH over the higher tidal volume treatment among the always-survivors subpopulation who are at a generally lower risk of death. This finding echoes the overall average treatment effect reported in the original trial analysis (Brower et al. (2000)).

To further assess the adequacy of using the likely always-survivor subpopulation for approximating the latent always-survivor subpopulation, we compare in Table 2 the estimated means of covariates from these two subpopulations; the covariate means for the likely always-survivors are computed directly from $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$, whereas the covariate means for the latent always-survivor subpopulation are obtained from the posterior samples. In each iteration of our chain, we further estimate the absolute standardized difference (ASD) for each covariate and present its posterior mean in Table 2. The two subpopulations have almost identical covariate means and the posterior mean of ASD is below 1% for each covariate (much lower than the usual 10% cutoff suggested in Austin and Stuart (2015) for observational studies). This finding reassures the plausibility of subsequent analyses based on the set of likely always-survivors.

Before moving on to studying effect heterogeneity among the always-survivors, we provide some additional intuitions on who are the always-survivors in the ARMA study. First, we obtain the variable importance plots generated from the BART models fitted for augmented latent variables, Z and W, in Web Figures A2 and A3 of Web Appendix A3. From these plots, we observe that Systolic BP, $AaDO_2$, APACHE III, PtoF, FiO_2, Vasopressors, and Platelet are seven key variables of higher importance than the remaining variables. Second, we present in Web Table A2 the posterior mean of each covariate by latent stratum, along with the posterior mean of maximum pairwise ASD used in observational studies (Li and Li (2019), McCaffrey et al. (2013)). The seven key variables identified by the variable importance plots, along

Estimated mean of covariates in two subpopulations and the posterior mean of absolute standardized differences by comparing between these two subpopulations. The second column corresponds to the mean of covariates in the identified likely always-survivors, and the third column corresponds to the posterior mean of the average covariate value in the latent always-survivor stratum

Covariates	$\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$	(Latent) always-Survivors	Posterior mean of ASD
Age	50.34	50.34	0.0019
Sex (female) (%)	0.386	0.385	0.0019
Race (White) (%)	0.271	0.270	0.0016
Tidal volume (ml)	678.74	678.75	0.0012
PEEP (cm water)	8.21	8.21	0.0015
PaO_2 (mm Hg)	85.35	85.32	0.0014
FiO_2 (mm Hg)	0.606	0.606	0.0026
$PaCO_2$ (mm Hg)	36.57	36.58	0.0012
PaO_2/FiO_2 (PtoF)	155.04	154.91	0.0023
$AaDO_2$	306.40	306.75	0.0027
Arterial pH	7.4096	7.4096	0.0016
APACHE III	76.39	76.40	0.0017
Systolic BP	100.15	100.13	0.0012
Glasgow coma scale	11.28	11.28	0.0016
Platelet (count/nl)	117.26	117.43	0.0031
Creatine (mg/dl)	1.117	1.115	0.0013
Bilirubin (mg/dl)	0.868	0.866	0.0014
Vasopressors (%)	0.724	0.724	0.0015

with Arterial pH, correspond to the largest posterior mean ASD values (hence explaining differences between principal strata). From these results, we find that always-survivors are generally associated with having better health profiles, in terms of much lower APACHE III score, $AaDO_2$, FiO_2, higher platelet count, and younger in age. The always-survivors also show the highest percentage of vasopressors use with the highest average level of systolic BP.

4.3.2. *Visualizing conditional survivor average causal effects.* Figure 1 shows the posterior mean and 95% credible intervals of $\Delta_{CSACE}(X_i)$ for the 522 participants identified as likely always-survivors (in both $\mathbb{S}_1$ and $\mathbb{S}_{(0.8,1)}$). The plot indicates an overall benefit in terms of reducing the DTRH among those receiving low tidal volume treatment. But the conditional causal effects clearly differ to some degree, ranging from -46.94 to -8.27 days, suggesting heterogeneity in response to the low tidal volume treatment. A total of 37.8% of identified likely always-survivors correspond to a credible interval excluding zero, which supports a strong, beneficial causal effect due to the low tidal volume treatment.

Henderson et al. (2020) provided several approaches for characterizing the degree of effect heterogeneity without truncation by death, and we apply their approaches to the ARMA trial for the identified likely always-survivors. To begin with, an alternative characterization of treatment effect heterogeneity can be achieved by examining the empirical distribution of the CSACEs over $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$, $H(u) = N_{11}^{-1} \sum_{i \in \mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}} \mathbb{I}\{\Delta_{CSACE}(X_i) \leq u\}$, which could be directly estimated by

$$(8) \qquad \hat{H}(u) = \frac{1}{N_{11}} \sum_{i \in \mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}} \mathbb{P}\{\Delta_{CSACE}(X_i) \leq u | \text{Obs. Data}, \boldsymbol{\theta}\}.$$

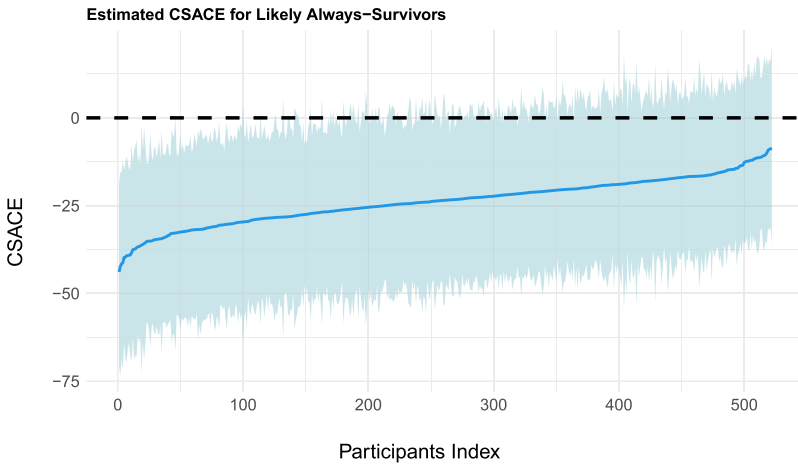**Estimated CSACE for Likely Always−Survivors**

FIG. 1. *Posterior means of CSACE (darker blue) with corresponding 95% credible intervals (lighter blue) for a total of* 522 *participants who are likely always-survivors (in $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$). A negative CSACE value indicates reduced DTRH under the low tidal volume treatment, compared to the traditional tidal volume treatment, which is considered beneficial.*

To better visualize the spread of CSACE over $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$, we estimate the density function associated with (8) by computing the posterior mean of a kernel function $K_\lambda(\bullet)$,

$$(9) \qquad \hat{h}(u) = \frac{1}{N_{11}} \sum_{i \in \mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}} \mathbb{E}\{K_\lambda(u - \Delta_{\mathrm{CSACE}}(X_i))|\text{Obs. Data}, \boldsymbol{\theta}\}.$$

The bandwidth λ is set as $\{0.9 \times \min(\hat{\sigma}_{\mathrm{CSACE}}, \widehat{\mathrm{IQR}}_{\mathrm{CSACE}})\}/(1.34 \times N_{11}^{1/5})$, where $\hat{\sigma}_{\mathrm{CSACE}}$ and $\widehat{\mathrm{IQR}}_{\mathrm{CSACE}}$ are posteriors means of the standard deviation and interquartile range of $\Delta_{\mathrm{CSACE}}(X_i)$. The left panel of Figure 2 presents a histogram of the posterior means of CSACE for each participant in $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$, and the right panel is the estimated posterior mean empirical density (the average of $\hat{h}(u)$ obtained at each MCMC iteration), which refers to the estimate of the entire distribution of the underlying treatment effects among $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$. The variation in the treatment effect suggested by the right panel is larger than that by the
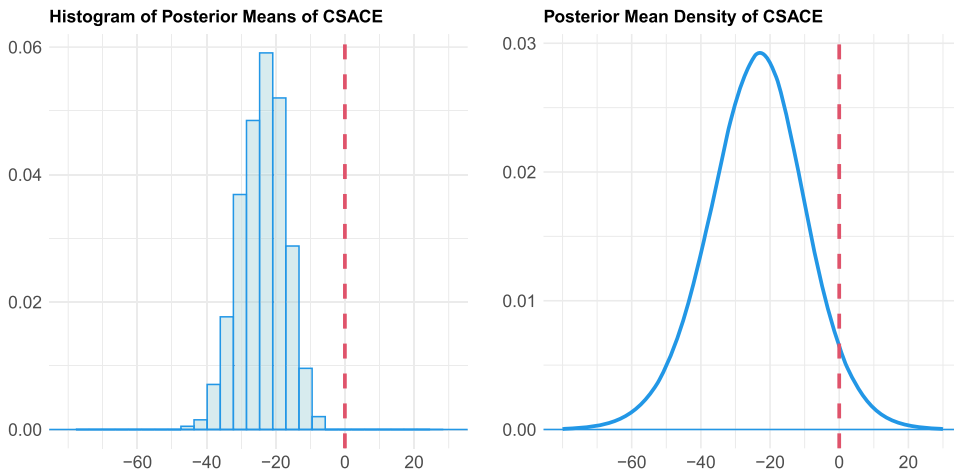

FIG. 2. *Left panel:Histogram of posterior means of CSACE. The histogram is constructed using posterior means of CSACE of each likely always-survivor participants (in $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$). Right panel: Posterior mean density of CSACE. The smooth estimate of the density function was computed, as described in (9).*

left panel, matching the intuition that the variance of conditional means is often smaller than the individual variation. Nonetheless, the estimated CSACEs were primarily negative via either visualization technique, leading to converging evidence. Thus, we conclude that the low tidal volume treatment leads to shorter DTRH, and the greatest reduction according to the posterior means of CSACE reaches over 45 days.

4.3.3. *Quantifying heterogeneity in conditional survivor average causal effects.* In addition to visualization, we apply two metrics considered in Henderson et al. (2020) to the ARMA trial for quantifying the degree of heterogeneity in the estimated CSACE among the likely always-survivors. First, the existence of heterogeneity can be quantified using the posterior probabilities of the differential survivor causal effect for each participant $i \in \mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$, defined as

$$\mathcal{D}_i = \mathbb{P}\{\Delta_{\text{CSACE}}(X_i) \leq \bar{\Delta}_{\text{CSACE}}|\text{Obs. Data}, \boldsymbol{\theta}\},$$

along with the absolute differential survivor causal effect,

$$\mathcal{D}_i^* = \max\{1 - 2\mathcal{D}_i, 2\mathcal{D}_i - 1\},$$

where $\bar{\Delta}_{\text{CSACE}}$ is the average of the CSACE among $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$. Notice that in the ARMA trial application, we have verified that $\bar{\Delta}_{\text{CSACE}} \approx \Delta_{\text{SACE}}$, and, therefore, the differential survivor causal effect can be approximately equivalently defined as $\mathcal{D}_i = \mathbb{P}\{\Delta_{\text{CSACE}}(X_i) \leq \Delta_{\text{SACE}}|\text{Obs. Data}, \boldsymbol{\theta}\}$.

The differential survivor causal effect, $\mathcal{D}_i$, is a measure of the evidence that the CSACE, $\Delta_{\text{CSACE}}(X_i)$, is less than or equal to the average of CSACE among the set of likely always-survivors, and thus, we should expect both high and low values of $\mathcal{D}_i$ in settings where nonnegligible heterogeneity of treatment effects exists. The closely-related quantity, $\mathcal{D}_i^*$, approaches 1 as the value of $\mathcal{D}_i$ approaches either 0 or 1, and $\mathcal{D}_i^* = 0$ when $\mathcal{D}_i = 1/2$. For a given participant $i \in \mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$, we, therefore, consider there to be strong evidence of heterogeneity in CSACEs if $\mathcal{D}_i^* > 0.9$ (equivalently, if $\mathcal{D}_i < 0.05$ or $\mathcal{D}_i > 0.95$), moderate evidence of heterogeneity provided that $\mathcal{D}_i^* > 0.8$ (equivalently, if $\mathcal{D}_i < 0.1$ or $\mathcal{D}_i > 0.9$), and mild evidence of heterogeneity if $\mathcal{D}_i^* > 0.7$ (equivalently, if $\mathcal{D}_i < 0.15$ or $\mathcal{D}_i > 0.85$). In the simulation study by Henderson et al. (2020) without truncation by death and for cases with treatment effect homogeneity, they found that the proportion of participants exhibiting high values of the $\mathcal{D}_i^*$ should, ideally, be zero or quite close to zero. For this reason the proportion of participants with $\mathcal{D}_i^* > 0.9$ can potentially be a useful summary measure for detecting heterogeneity in CSACEs. In the ARMA trial, approximately 0.4% of participants had strong evidence of heterogeneity in CSACEs (i.e., $\mathcal{D}_i^* > 0.9$), approximately 1.3% of participants had moderate evidence of heterogeneity (i.e., $\mathcal{D}_i^* > 0.8$), and approximately 6.1% of participants had mild evidence of heterogeneity (i.e., $\mathcal{D}_i^* > 0.7$). Web Figure A4 in Web Appendix A3 presents the histogram and density describing the distribution of $\mathcal{D}_i^*$.

Second, the heterogeneity in CSACEs can also be assessed via the proportion of always-survivors benefiting from the treatment (Henderson et al. (2020)), where we directly infer the number of participants benefiting from the low tidal volume treatment from the set of participants who are likely always-survivors. In specific, the proportion of always-survivors benefiting from the low tidal volume treatment can be defined as

$$\mathcal{Q} = \frac{1}{N_{11}} \sum_{i \in \mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}} \mathbb{I}\{\Delta_{\text{CSACE}}(X_i) < 0\}.$$

The posterior mean of $\mathcal{Q}$ is an average of the posterior probabilities of treatment benefit, $\hat{q}_i = \mathbb{P}\{\Delta_{\text{CSACE}}(X_i) < 0|\text{Obs. Data}, \boldsymbol{\theta}\}$, which summarizes the treatment benefit of a participant from a probabilistic perspective. Trial participants who are more likely to benefit from

TABLE 3
Tabulation of proportions of participants in $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$ benefiting from the low tidal volume treatment to different degrees

Benefiting degree	Proportion (%) among $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$
$\mathbb{P}\{\Delta_{\mathrm{CSACE}}(X_i) < 0 \mid \text{Obs. Data}, \boldsymbol{\theta}\} > 0.99$	19.0
$\mathbb{P}\{\Delta_{\mathrm{CSACE}}(X_i) < 0 \mid \text{Obs. Data}, \boldsymbol{\theta}\} > 0.95$	68.4
$\mathbb{P}\{\Delta_{\mathrm{CSACE}}(X_i) < 0 \mid \text{Obs. Data}, \boldsymbol{\theta}\} > 0.9$	88.9
$\mathbb{P}\{\Delta_{\mathrm{CSACE}}(X_i) < 0 \mid \text{Obs. Data}, \boldsymbol{\theta}\} > 0.8$	98.5

the low tidal volume treatment will have higher chances of a negative CSACE. A tabulation of participants among the likely always-survivors according to their likelihood of benefiting from the low tidal volume treatment is presented in Table 3, where 68.4% of participants in $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$ exhibit a posterior probability of benefiting from the low tidal volume treatment greater than 0.95, and 88.9% exhibit a posterior probability of benefiting from the low tidal volume treatment greater than 0.9. Web Figure A5 in Web Appendix A3 presents the histogram and density describing the distribution of $\hat{q}_i$.

4.4. *Exploring effect moderation.* We adopt the Bayesian "fit-the-fit" strategy (Hahn, Murray and Carvalho (2020)) to explore the relationship between CSACEs and covariates among the likely always-survivors. This approach amounts to first applying our proposed method to estimate the CSACEs for each likely always-survivor (in $\mathbb{S}_1 \cup \mathbb{S}_{(0.8,1)}$) and then, using these estimated CSACEs as a new response variable in an exploratory analysis, to identify important effect moderators and possible subgroups defined by such effect moderators. Specifically, in our exploratory analysis, a classification and regression tree (CART) model was used to regress the posterior means of the CSACE on the covariates. We fit a sequence of CART models, with covariates (standardized to have zero mean and unit variance) sequentially added to the CART model in a stepwise manner to improve the model fit measured by R^2. At each step the variable leading to the largest R^2 improvement was selected into the model, and the procedure was terminated when the percent improvement in R^2 was less than 1%. Results showed that covariates with the five largest estimated standardized coefficients in absolute value were (from high to low): $AaDO_2$, sex, FIO_2, PtoF, and systolic BP. Subgroup treatment effects were estimated by averaging CSACEs among individuals falling into each node of the final CART model, and the branch decision rules suggest final combination rules of covariates. Figure 3 presents the final tree estimates, based on the top two covariates, that are the main drivers of the heterogeneity in CSACE, where the final R^2 between the tree fit and the posterior mean CSACE is 78.9%. The 95% credible interval (CrI) of each subgroup causal effect is obtained by projecting the posterior draws of CSACE onto the predictive space of the final CART fit and hence comes with a valid Bayesian uncertainty interpretation (Woody, Carvalho and Murray (2021)).

In Figure 3 the first splitting variable was sex. Female always-survivors had approximately 28.4 (95% CrI: 18.2–40.2) days shorter in DTRH, on average, under the low tidal volume treatment, whereas male always-survivors had approximately 20.9 (95% CrI: 11.4–26.9) days shorter in DTRH, on average, under the low tidal volume treatment. The second level of variable splitting by the value of $AaDO_2$, the first alveolar-arterial oxygen gradient, provided further resolution on the magnitude of the treatment benefit for participants. The most beneficial subgroup was female always-survivors with $AaDO_2 \geq 258.9$, where the average reduction in DTRH is 32.3 (95% CrI: 20.4–45.7) under the low tidal volume treatment. Among male always-survivors, those with $AaDO_2 < 296.6$ experience treatment benefit from the low tidal volume treatment with an average DTRH of approximately 17.9 (95% CrI: 6.9–27.8) days
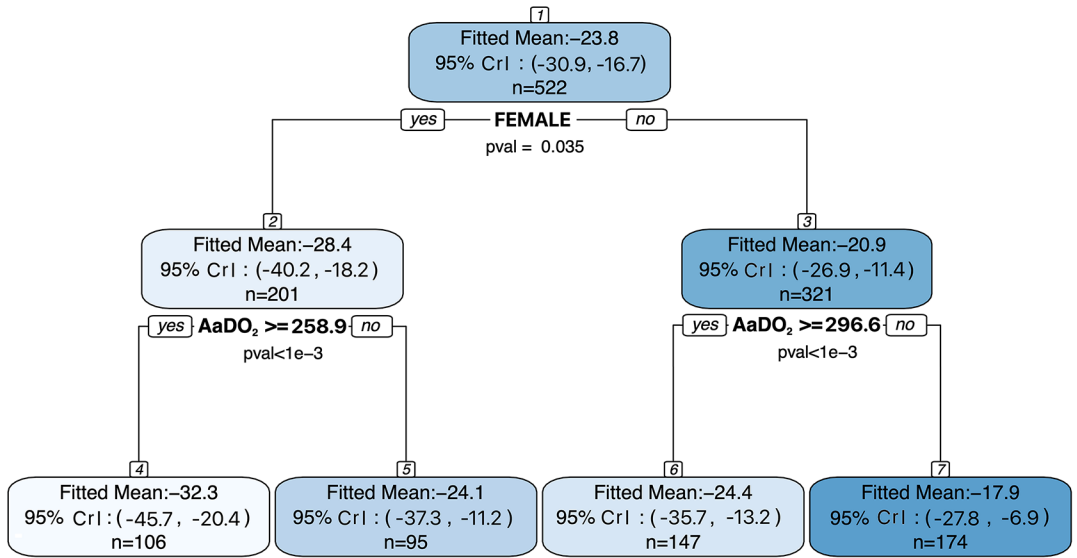
FIG. 3. *Final CART model fit to the posterior mean DTRHs (in days) between the low tidal volume treatment and the traditional tidal volume treatment. Values in each node correspond to the posterior mean and 95% credible intervals for the average CSACE for the subgroup of individuals represented in that node.*

shorter; in comparison, male always-survivors with $AaDO_2 \geq 296.6$ experience even greater treatment benefit from the low tidal volume treatment with an average DTRH of approximately 24.4 (95% CrI: 13.2–35.7). Concordant with Figure 3, we also visualize the posterior distribution of each pairwise difference in the subgroup causal effects in Figure 4. It is apparent that the female always-survivors with $AaDO_2 \geq 258.9$ have the largest benefit from the low tidal volume treatment, as the majority of posterior mass in treatment effect difference is below zero when contrasting this subgroup to the others.

Finally, we also explore the patterns between the estimated CSACE and the posterior probability of being an always-survivor within each covariate-defined subgroup in Figure 5. This exploration reveals a slight tendency that a subgroup with larger treatment benefits may have higher probability of being always-survivors. For example, among the set of likely always-survivors (Section 4.3.1), there are only four participants with estimated posterior probability of being an always-survivor lower than 90%—one female with $AaDO_2 < 258.9$, two males with $AaDO_2 \geq 296.6$, and one male with $AaDO_2 < 296.6$. The most beneficial subgroup correspond to the least uncertainty in being always-survivors; that is, except for one participant, all female with $AaDO_2 \geq 258.9$ have estimated posterior probability of being an always-survivor being at least 99%.

Overall, our exploratory analyses indicate that the reduction in DTRH was greatest among female always-survivors with $AaDO_2 \geq 258.9$ at baseline and is smallest among male always-survivors with $AaDO_2 < 296.6$. The effect among females is consistent with prior findings in existing observational studies. For instance, a study from the Large Observational Study to Understand the Global Impact of Severe Acute Respiratory Failure (LUNG SAFE) (McNicholas et al. (2019)), which is an international, multicenter, prospective cohort study, conducted for four consecutive weeks in the winter of 2014 in a convenience sample of 459 ICUs from 50 countries across six continents, and found that surviving females had a shorter duration of invasive mechanical ventilation and reduced length of stay, compared with males. Second, participants with more severe acute respiratory disease syndrome have lower $PaO_2:FiO_2$ ratios and larger $AaDO_2$ gradients (Helmholz Jr. (1979)). Thus, there is some speculation that individuals with severe acute respiratory distress syndrome may be more
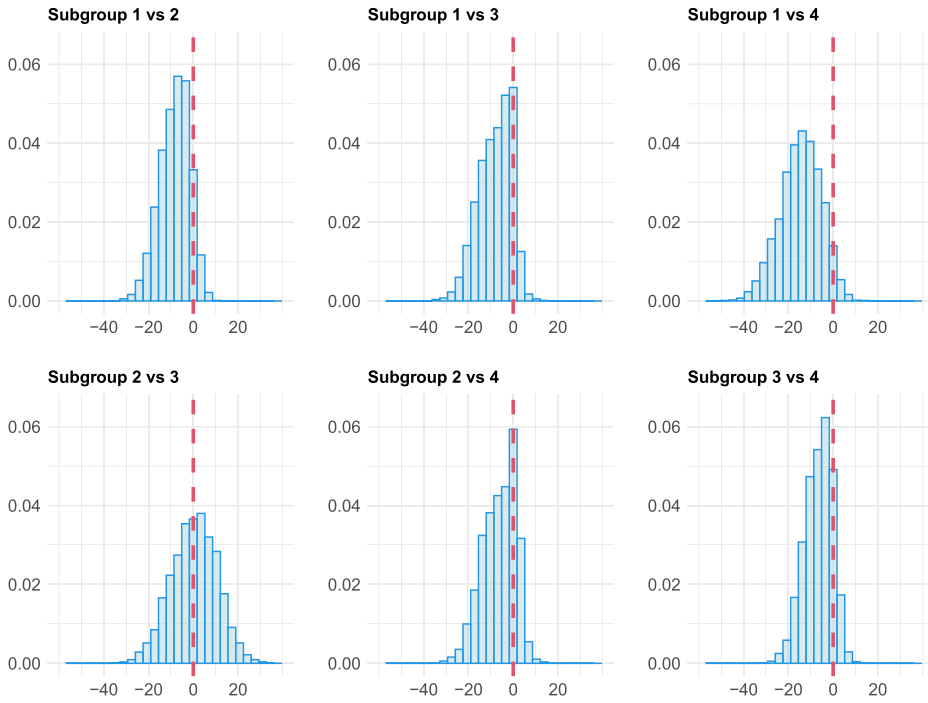
FIG. 4.　*Posterior distributions of the difference in treatment effects between any two always-survivor subgroups—Subgroup 1: Female always-survivors with* $AaDO_2 \geq 258.9$, *Subgroup 2: Female always-survivors with* $AaDO_2 < 258.9$, *Subgroup 3: Male always-survivors with* $AaDO_2 \geq 296.6$, *and Subgroup 4: male always-survivors with* $AaDO_2 < 296.6$.
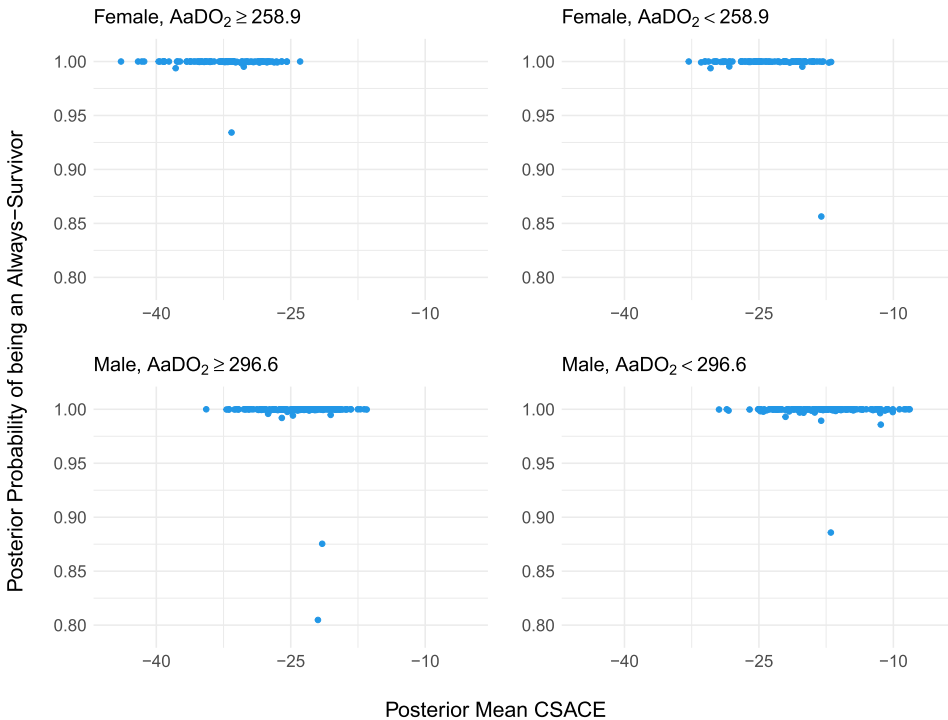

FIG. 5.　*Scatter plots of the posterior probability of being in the always-survivor stratum against the posterior mean CSACE by each subgroup.*

likely to benefit from the intervention, whereas those with smaller gradients would be more strongly associated with poor clinical outcomes, such as death, or in our context, discharge to a long-term acute care hospital, skilled nursing facility, or hospice, thereby delaying time to get home. In other words, the always-survivors with relatively higher $AaDO_2$ had more "opportunity to benefit" (Goligher et al. (2021)). Taken together, female always-survivors appear to benefit more from the low tidal volume treatment than their male counterparts. Thus, while the exact mechanisms may not be clear, our findings do seem plausible and directly engage with current debates in the treatment of acute lung injury and acute respiratory disease syndrome and the associated research literature (Del Sorbo et al. (2017), Fan et al. (2017), Shen et al. (2019)).

5. Discussion. Recent advancements in Bayesian machine learning have provided important tools to flexibly specify the outcome model to reduce the potential estimation bias that occurs when estimating the average treatment effect and has enabled researchers to estimate heterogeneous causal effects among the study population. This article advances the application of BART to quantify the SACE and CSACE within the principal stratification framework when a nonmortality outcome is subject to truncation by death and thus opens the door to a wide range of causal discoveries that could inform individualized care delivery in the motivating critical care use case. We applied our proposed approach to operationalize considerations for *exploratory* heterogeneity of treatment effect analysis among the likely always-survivors in the ARMA trial and identified key effect moderators using a data-driven approach that aligns with several clinical prior findings, as we explicate in Section 4.4.

Beyond effect moderation due to sex and $AaDO_2$, we found in our analysis of the ARMA trial, the Bayesian "fit-the-fit" strategy that we employed also identified pressure of arterial oxygen, the ratio of PaO_2 to FiO_2, and systolic blood pressure as three additional factors that weakly moderate the causal effects among the always-survivors. However, the subgroup structure with more effect moderators necessarily becomes more complex and less interpretable; however, it is worth noting that these findings all still align with the clinical literature. We, therefore, decided to prioritize the top two effect moderators in our final exploratory analysis but fully acknowledge the value of future work for better synthesizing more than two effect moderators to generate interpretable subgroup findings. To the best of our knowledge, this is the first study that employed Bayesian machine learning tools to study effect moderation for mechanical ventilation treatments among the always-survivors population in a critical care intervention study. The investigation of the true causal mechanisms of such effect moderation will be left for future studies and necessitates structured engagement with a wider set of clinical colleagues.

In exploring the variation among the CSACE estimates, we have implemented a decision to first identify the likely always-survivors, which include 260 survivors receiving the high tidal volume treatment and a subset of survivors (262 out of 303) receiving the low tidal volume treatment but having the highest posterior probability of being an always-survivor. In Section 4.3.1 we have compared the baseline characteristics among the likely always-survivors and those among the latent always-survivors (generated through our posterior sampling algorithm) and found no systematic difference. The SACE estimates are also identical between these two sets of participants, suggesting no clear evidence against the adequacy of using the likely always-survivors to approximate the latent always-survivors. Subsequently, pursuing the CSACE analysis with the likely always-survivors is based on two practical considerations. First, having a tangible subset of participants helps us directly study the variation in treatment effect for the nonmortality outcome (response heterogeneity) without the complications, due to variation in the conditional probability, of being an always-survivor (membership heterogeneity). Had we estimated CSACE based on covariates of the entire trial, it

would be necessary to disentangle response heterogeneity from membership heterogeneity, which is challenging. Second, we recognize that an alternative approach is to focus on the 260 survivors receiving the high tidal volume treatment. Under Assumptions 1 and 2, this smaller subset is always a valid approximation to the latent always-survivor subpopulation. However, this alternative approach comes at a cost of substantially reduced sample size for exploring heterogeneity of treatment effect. Since we did not find systematic differences between the likely always-survivors and the latent always-survivors, we considered an analysis with the largest possible sample size. In cases where the stratum membership can not be easily predicted and hence the use of likely always-survivors may be questionable, it would then be preferable to focus the analysis on the smaller set of survivors under the usual care condition.

To estimate the SACE for nonmortality outcomes truncated by death, typically, both structural assumptions and parametric modeling assumptions are invoked. The structural assumptions are necessary to identify the causal parameter with observed data, whereas the parametric assumptions are useful in modeling the observed data and summarizing information from observed data. Under the principal stratification framework, the proposed Bayesian machine learning approach differs from the existing methods in that we considered a finite mixture of BART Y-models (with mixture probability also given by nested Probit BART models) rather than a finite mixture of fully parametric Y-models, thus relaxing some of the parametric modeling assumptions. In simpler settings without any intermediate outcomes, the BART approach has shown to be a flexible and robust tool to estimate the average treatment effect and its conditional counterpart with minimum bias and high precision (Dorie et al. (2019), Hahn, Murray and Carvalho (2020), Hill (2011), Hu, Ji and Li (2021)). Under this perspective our work represents a generalization of the BART approach to additionally account for an intermediate variable through a mixture model framework. While relaxing the parametric assumptions, our approach still maintains standard structural assumptions to estimate the SACE. The SUTVA and randomization assumptions are generally plausible in applications to randomized trials, but the monotonicity assumption may not always be plausible, such as in noninferiority or comparative effectiveness trials where there is an active comparator. In this case one potential solution is to allow for an additional harmed strata (Zhang, Rubin and Mealli (2009)) by extending the nested Probit BART with another layer and including an additional Y-model for the harmed population under the usual care condition. Alternatively, it may be interesting to consider the monotone probit BART for mixture weights similar to Papakostas et al. (2023) to reflect a stochastic monotonicity constraint that one active treatment does not mitigate the risk of mortality compared to the other. While theoretically appealing, such extended approaches may be overparameterized and lead to semiparametric mixture models that are only weakly identified in the sense that the posterior distributions of SACE and CSACE remain flat around the region of highest density. Of note, integrating BART into the mixture model framework for principal stratification analysis is not the only approach to address heterogeneity of treatment effect under truncation by death. In future work it would be interesting to compare the proposed BART approach with alternative Bayesian nonparametric priors, such as the dependent Dirichlet process-Gaussian process prior (Xu et al. (2016, 2022), Roy, Lum and Daniels (2017)), for estimating CSACE. In addition, the extent to which alternative identification strategies (Ding and Lu (2017), Hayden, Pauler and Schoenfeld (2005)) might improve the current mixture model framework to estimate CSACE can also form the scope of further research.

A relevant extension of our approach is to view both DTRH and time-to-death as two time-to-event outcomes under the semicompeting risks framework. Under this framework, a continuous-time principal stratification approach has been developed in Comment et al. (2019), Xu et al. (2022), and Nevo and Gorfine (2022) to define always-survivors up to each follow-up time point, based on which a time-varying version of SACE (TV-SACE) is proposed. Our SACE estimand, defined in Section 2, can thus be viewed as a "snapshot" version

of TV-SACE at $u = 180$ days; and we focus on a snapshot version of principal stratification to facilitate the exploration of CSACE without addressing complications due to temporal heterogeneity of treatment effect. As a supplementary analysis, we implemented the Bayesian nonparametric approach of Xu et al. (2022) using the *BaySemiCompeting* R package to estimate the TV-SACE from the ARMA data; the details are summarized in Web Appendix A5. The results indicate that low tidal volume treatment strategy leads to consistently higher chances of returning home prior to day u among the always-survivors up to day u, for each $0 < u \leq 180$. However, a full development of formal methodology for estimation and interpretation of time-varying conditional survivor average causal effect (TV-CSACE), possibly through BART for time-to-event outcomes (Henderson et al. (2020), Hu, Ji and Li (2021)), can be a fruitful direction for further investigation.

SUPPLEMENTARY MATERIAL

Web appendix of "A Bayesian machine learning approach for estimating heterogeneous survivor causal effects: Applications to a critical care trial" (DOI: 10.1214/23-AOAS1792SUPPA; .pdf). The Web Appendix contains the Gibbs sampler of the parametric model (A1), a Monte Carlo simulation study (A2), related Web Figures (A3), and two aforementioned sensitivity analyses (A4 and A5).

Code (DOI: 10.1214/23-AOAS1792SUPPB; .zip). R Code. (The R code contains scripts for the simulation study in Web Appendix A2, where the comparison methods, i.e., 'BART', 'YBSP', 'YPSB', and 'Parametric', as well as the level of difficulty in identifying always-survivors (corresponding to different data generating processes), i.e., 'easy' and 'difficult', are included in the name of the script.)

REFERENCES

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

AUSTIN, P. C. and STUART, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat. Med.* **34** 3661–3679. MR3422140 https://doi.org/10.1002/sim.6607

BARGAGLI-STOFFI, F. J., DE WITTE, K. and GNECCO, G. (2022). Heterogeneous causal effects with imperfect compliance: A Bayesian machine learning approach. *Ann. Appl. Stat.* **16** 1986–2009. MR4455908 https://doi.org/10.1214/21-aoas1579

BIA, M., MATTEI, A. and MERCATANTI, A. (2022). Assessing causal effects in a longitudinal observational study with "truncated" outcomes due to unemployment and nonignorable missing data. *J. Bus. Econom. Statist.* **40** 718–729. MR4410893 https://doi.org/10.1080/07350015.2020.1862672

BROWER, R. G., MATTHAY, M. A., MORRIS, A., SCHOENFELD, D., THOMPSON, B. T., WHEELER, A. et al. (2000). Acute respiratory distress syndrome network. Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. *N. Engl. J. Med.* **342** 1301–1308.

CHEN, X., HARHAY, M. O., TONG, G. and LI, F. (2024). Supplement to "A Bayesian machine learning approach for estimating heterogeneous survivor causal effects: Applications to a critical care trial." https://doi.org/10.1214/23-AOAS1792SUPPA, https://doi.org/10.1214/23-AOAS1792SUPPB

CHIBA, Y. and VANDERWEELE, T. J. (2011). A simple method for principal strata effects when the outcome has been truncated due to death. *Amer. J. Epidemiol.* **173** 745–751. https://doi.org/10.1093/aje/kwq418

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

COMMENT, L., MEALLI, F., HANEUSE, S. and ZIGLER, C. (2019). Survivor average causal effects for continuous time: A principal stratification approach to causal inference with semicompeting risks. arXiv preprint arXiv:1902.09304, 1–28.

DEL SORBO, L., GOLIGHER, E. C., MCAULEY, D. F., RUBENFELD, G. D., BROCHARD, L. J., GATTINONI, L., SLUTSKY, A. S. and FAN, E. (2017). Mechanical ventilation in adults with acute respiratory distress syndrome. Summary of the experimental evidence for the clinical practice guideline. *Ann. Amer. Thorac. Soc.* **14** S261–S270. PMID: 28985479. https://doi.org/10.1513/AnnalsATS.201704-345OT

DENG, Y., GUO, Y., CHANG, Y. and ZHOU, X.-H. (2021). Identification and estimation of the heterogeneous survivor average causal effect in observational studies. arXiv preprint arXiv:2109.13623, 1–23.

DING, P., GENG, Z., YAN, W. and ZHOU, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *J. Amer. Statist. Assoc.* **106** 1578–1591. MR2896858 https://doi.org/10.1198/jasa.2011.tm10265

DING, P. and LI, F. (2018). Causal inference: A missing data perspective. *Statist. Sci.* **33** 214–237. MR3797711 https://doi.org/10.1214/18-STS645

DING, P. and LU, J. (2017). Principal stratification analysis using principal scores. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 757–777. MR3641406 https://doi.org/10.1111/rssb.12191

DORIE, V., HILL, J., SHALIT, U., SCOTT, M. and CERVONE, D. (2019). Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statist. Sci.* **34** 43–68. MR3938963 https://doi.org/10.1214/18-STS667

EGLESTON, B. L., SCHARFSTEIN, D. O., FREEMAN, E. E. and WEST, S. K. (2006). Causal inference for non-mortality outcomes in the presence of death. *Biostatistics* **8** 526–545. https://doi.org/10.1093/biostatistics/kxl027

FAN, E., DEL SORBO, L., GOLIGHER, E. C., HODGSON, C. L., MUNSHI, L., WALKEY, A. J., ADHIKARI, N. K. J., AMATO, M. B. P., BRANSON, R. et al. (2017). An official American thoracic society/European society of intensive care medicine/society of critical care medicine clinical practice guideline: Mechanical ventilation in adult patients with acute respiratory distress syndrome. *Am. J. Respir. Crit. Care Med.* **195** 1253–1263. PMID: 28459336. https://doi.org/10.1164/rccm.201703-0548ST

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039 https://doi.org/10.1111/j.0006-341X.2002.00021.x

FRUMENTO, P., MEALLI, F., PACINI, B. and RUBIN, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *J. Amer. Statist. Assoc.* **107** 450–466. MR2980057 https://doi.org/10.1080/01621459.2011.643719

GOLIGHER, E. C., COSTA, E. L. V., YARNELL, C. J., BROCHARD, L. J., STEWART, T. E., TOMLINSON, G., BROWER, R. G., SLUTSKY, A. S. and AMATO, M. P. B. (2021). Effect of lowering Vt on mortality in acute respiratory distress syndrome varies with respiratory system elastance. *Am. J. Respir. Crit. Care Med.* **203** 1378–1385. https://doi.org/10.1164/rccm.202009-3536OC

HAHN, P. R., MURRAY, J. S. and CARVALHO, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Anal.* **15** 965–1056. Includes comments and discussions by 25 discussants and a rejoinder by the authors. MR4154846 https://doi.org/10.1214/19-BA1195

HAHN, P. R., MURRAY, J. S. and MANOLOPOULOU, I. (2016). A Bayesian partial identification approach to inferring the prevalence of accounting misconduct. *J. Amer. Statist. Assoc.* **111** 14–26. MR3494635 https://doi.org/10.1080/01621459.2015.1084307

HARHAY, M. O., RATCLIFFE, S. J., SMALL, D. S., SUTTNER, L. H., CROWTHER, M. J. and HALPERN, S. D. (2019). Measuring and analyzing length of stay in critical care trials. *Med. Care* **57** e53–e59. PMID: 30664613. https://doi.org/10.1097/MLR.0000000000001059

HARHAY, M. O., WAGNER, J., RATCLIFFE, S. J., BRONHEIM, R. S., GOPAL, A., GREEN, S., COONEY, E., MIKKELSEN, M. E., KERLIN, M. P. et al. (2014). Outcomes and statistical power in adult critical care randomized trials. *Am. J. Respir. Crit. Care Med.* **189** 1469–1478.

HAYDEN, D., PAULER, D. K. and SCHOENFELD, D. (2005). An estimator for treatment comparisons among survivors in randomized trials. *Biometrics* **61** 305–310. MR2135873 https://doi.org/10.1111/j.0006-341X.2005.030227.x

HELMHOLZ JR., H. F. (1979). The abbreviated alveolar air equation. *Chest* **75** 748. https://doi.org/10.1378/chest.75.6.748

HENDERSON, N. C., LOUIS, T. A., ROSNER, G. L. and VARADHAN, R. (2020). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *Biostatistics* **21** 50–68. MR4043845 https://doi.org/10.1093/biostatistics/kxy028

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

HIRANO, K., IMBENS, G. W., RUBIN, D. B. and ZHOU, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics* **1** 69–88.

HU, L., JI, J. and LI, F. (2021). Estimating heterogeneous survival treatment effect in observational data using machine learning. *Stat. Med.* **40** 4691–4713. MR4315446 https://doi.org/10.1002/sim.9090

IMAI, K. (2008). Sharp bounds on the causal effects in randomized experiments with "truncation-by-death". *Statist. Probab. Lett.* **78** 144–149. MR2382067 https://doi.org/10.1016/j.spl.2007.05.015

KADANE, J. B. (1975). The role of identification in Bayesian theory. In *Studies in Bayesian Econometrics and Statistics* (*in Honor of Leonard J. Savage*). *Contrib. Econom. Anal.* **86** 175–191. North-Holland, Amsterdam–Oxford. MR0483124

KIM, C., DANIELS, M. J., HOGAN, J. W., CHOIRAT, C. and ZIGLER, C. M. (2019). Bayesian methods for multiple mediators: Relating principal stratification and causal mediation in the analysis of power plant emission controls. *Ann. Appl. Stat.* **13** 1927–1956. MR4019162 https://doi.org/10.1214/19-AOAS1260

KIM, C., DANIELS, M. J., MARCUS, B. H. and ROY, J. A. (2017). A framework for Bayesian nonparametric inference for causal effects of mediation. *Biometrics* **73** 401–409. MR3665957 https://doi.org/10.1111/biom.12575

LI, F. and LI, F. (2019). Propensity score weighting for causal inference with multiple treatments. *Ann. Appl. Stat.* **13** 2389–2415. MR4037435 https://doi.org/10.1214/19-aoas1282

LONG, D. M. and HUDGENS, M. G. (2013). Sharpening bounds on principal effects with covariates. *Biometrics* **69** 812–819. MR3146777 https://doi.org/10.1111/biom.12103

MATTEI, A., LI, F. and MEALLI, F. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *Ann. Appl. Stat.* **7** 2336–2360. MR3161725 https://doi.org/10.1214/13-AOAS674

MATTEI, A. and MEALLI, F. (2007). Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics* **63** 437–446. MR2370802 https://doi.org/10.1111/j.1541-0420.2006.00684.x

MATTHAY, M. A., MCAULEY, D. F. and WARE, L. B. (2017). Clinical trials in acute respiratory distress syndrome: Challenges and opportunities. *Lancet Respir. Med.* **5** 524–534. https://doi.org/10.1016/S2213-2600(17)30188-1

MCCAFFREY, D. F., GRIFFIN, B. A., ALMIRALL, D., SLAUGHTER, M. E., RAMCHAND, R. and BURGETTE, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat. Med.* **32** 3388–3414. MR3074364 https://doi.org/10.1002/sim.5753

MCNICHOLAS, B. A., MADOTTO, F., PHAM, T., REZOAGLI, E., MASTERSON, C. H., HORIE, S., BELLANI, G., BROCHARD, L. and LAFFEY, J. G. (2019). Demographics, management and outcome of women and men with acute respiratory distress syndrome in the LUNG SAFE prospective cohort study. *Eur. Respir. J.* **54**. PMID: 31346004. https://doi.org/10.1183/13993003.00609-2019

NEVO, D. and GORFINE, M. (2022). Causal inference for semi-competing risks data. *Biostatistics* **23** 1115–1132. MR4496371 https://doi.org/10.1093/biostatistics/kxab049

PAPAKOSTAS, D., HAHN, P. R., MURRAY, J., ZHOU, F. and GERAKOS, J. (2023). Do forecasts of bankruptcy cause bankruptcy? A machine learning sensitivity analysis. *Ann. Appl. Stat.* **17** 711–739. MR4539050 https://doi.org/10.1214/22-aoas1648

ROY, J., LUM, K. J. and DANIELS, M. J. (2017). A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics* **18** 32–47. MR3612272 https://doi.org/10.1093/biostatistics/kxw029

SHEN, Y., CAI, G., GONG, S., DONG, L., YAN, J. and CAI, W. (2019). Interaction between low tidal volume ventilation strategy and severity of acute respiratory distress syndrome: A retrospective cohort study. *Crit. Care* **23** 254. https://doi.org/10.1186/s13054-019-2530-6

TAN, Y. V. and ROY, J. (2019). Bayesian additive regression trees and the general BART model. *Stat. Med.* **38** 5048–5069. MR4022845 https://doi.org/10.1002/sim.8347

TONELLI, A. R., ZEIN, J., ADAMS, J. and IOANNIDIS, J. P. A. (2014). Effects of interventions on survival in acute respiratory distress syndrome: An umbrella review of 159 published randomized trials and 29 meta-analyses. *Intens. Care Med.* **40** 769–787. https://doi.org/10.1007/s00134-014-3272-1

WANG, L., ZHOU, X.-H. and RICHARDSON, T. S. (2017). Identification and estimation of causal effects with outcomes truncated by death. *Biometrika* **104** 597–612. MR3694585 https://doi.org/10.1093/biomet/asx034

WENDLING, T., JUNG, K., CALLAHAN, A., SCHULER, A., SHAH, N. H. and GALLEGO, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Stat. Med.* **37** 3309–3324. MR3856345 https://doi.org/10.1002/sim.7820

WOODY, S., CARVALHO, C. M. and MURRAY, J. S. (2021). Model interpretation through lower-dimensional posterior summarization. *J. Comput. Graph. Statist.* **30** 144–161. MR4235972 https://doi.org/10.1080/10618600.2020.1796684

XU, Y., MÜLLER, P., WAHED, A. S. and THALL, P. F. (2016). Bayesian nonparametric estimation for dynamic treatment regimes with sequential transition times. *J. Amer. Statist. Assoc.* **111** 921–950. MR3561917 https://doi.org/10.1080/01621459.2015.1086353

XU, Y., SCHARFSTEIN, D., MÜLLER, P. and DANIELS, M. (2022). A Bayesian nonparametric approach for evaluating the causal effect of treatment in randomized trials with semi-competing risks. *Biostatistics* **23** 34–49. MR4366034 https://doi.org/10.1093/biostatistics/kxaa008

YANG, F. and SMALL, D. S. (2016). Using post-outcome measurement information in censoring-by-death problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78** 299–318. MR3453657 https://doi.org/10.1111/rssb.12113

ZHANG, J. L. and RUBIN, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *J. Educ. Behav. Stat.* **28** 353–368. https://doi.org/10.3102/10769986028004353

ZHANG, J. L., RUBIN, D. B. and MEALLI, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *J. Amer. Statist. Assoc.* **104** 166–176. MR2663040 https://doi.org/10.1198/jasa.2009.0012