PLOS ONE

RESEARCH ARTICLE

Bayesian credible subgroup identification for treatment effectiveness in time-to-event data

Duy Ngo[1,3], Richard Baumgartner[1]*, Shahrul Mt-Isa[2,4], Dai Feng[1], Jie Chen[1], Patrick Schnell[5]

1 Merck & Co., Inc., Kenilworth, NJ, United States of America, **2** MSD Research Laboratories, MSD, London, United Kingdom, **3** Department of Statistics, Western Michigan University, Kalamazoo, Michigan, United States of America, **4** School of Public Health, Imperial College London, London, United Kingdom, **5** The Ohio State University College of Public Health, Columbus, Ohio, United States of America

* richard_baumgartner@merck.com

Check for updates

OPEN ACCESS

Citation: Ngo D, Baumgartner R, Mt-Isa S, Feng D, Chen J, Schnell P (2020) Bayesian credible subgroup identification for treatment effectiveness in time-to-event data. PLoS ONE 15(2): e0229336. https://doi.org/10.1371/journal.pone.0229336

Editor: Alan D Hutson, Roswell Park Cancer Institute, UNITED STATES

Received: July 12, 2019

Accepted: February 4, 2020

Published: February 26, 2020

Copyright: © 2020 Ngo et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data used in this manuscript can be found at https://www.imbi.uni-freiburg.de/Royston-Sauerbrei-book.

Funding: This study was funded from a commercial source: Merck & Company. The funder provided support in the form of salaries for authors Duy Ngo, Richard Baumgartner, Shahrul Mt-Isa, Jie Chen and Dai Feng. The specific roles of these authors are articulated in the 'author contributions' section. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Due to differential treatment responses of patients to pharmacotherapy, drug development and practice in medicine are concerned with personalized medicine, which includes identifying subgroups of population that exhibit differential treatment effect. For time–to–event data, available methods only focus on detecting and testing treatment–by–covariate interactions and may not consider multiplicity. In this work, we introduce the Bayesian credible subgroups approach for time–to–event endpoints. It provides two bounding subgroups for the true benefiting subgroup: one which is likely to be contained by the benefiting subgroup and one which is likely to contain the benefiting subgroup. A personalized treatment effect is estimated by two common measures of survival time: the hazard ratio and restricted mean survival time. We apply the method to identify benefiting subgroups in a case study of prostate carcinoma patients and a simulated large clinical dataset.

1 Introduction

A goal of clinical trials is to evaluate primary endpoints that describe comprehensive characteristics of the disease under study and allow for comparisons of treatments in an entire population. However, trial populations are often heterogeneous due to different demographics, medical history or genetic makeup among patients. In some cases, the efficacy of marketed treatments could not be replicated in follow–up clinical trials [1]. The inability to replicate study results in follow-up trials may be caused by different proportions of benefiting and non–benefiting subgroups of patients from experimental treatment compared to control. Recently, regulators and health technology assessment agencies worldwide have had a growing interest in identifying subgroups of patients who benefit from a treatment. Several methods to find such subgroups in clinical trials have been proposed in the literature [2–4].

Our study is motivated by a practical need for identifying subgroups of patients with improved time-to-event or survival outcomes. Many tree–based and model–based methods have been developed for time–to–event subgroup analysis [5–8]. Ballarini et al. [9] recently introduced a multiple regression model with a Lasso–type penalty to estimate benefiting

Competing interests: We have read the journal's policy and the authors of this manuscript have the following competing interests: D. Ngo, R. Baumgartner, S. Mt-Isa, D. Feng, and J. Chen are currently (or were at the time the study was conducted) employees of Merck & Co., Inc., Kenilworth, NJ, who may own stock in Merck. Patrick Schnell is an employee of The Ohio State University College of Public Health and received consultancy fees from Merck related to the study in general, but not specific to the publication/ authoring of the manuscript. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

subgroups based on estimates of the personalized treatment effect (PTE) and its post–selection confidence intervals. Traditionally, log–rank tests and Cox proportional hazard models have been used to compare treatment effects on an entire population. For example, researchers can identify subgroups with an overall *positive* treatment effect such as hazard ratio (HR) <1. However, this approach does not identify a benefiting subgroup in which *all* members defined by a set of observed baseline characteristics have a positive treatment effect. Likewise, the average treatment effect (ATE) is the average over the entire population of individual treatment effects, and it does not accurately represent each patient's treatment effect.

Recently, the personalized treatment effects (PTEs) have been considered as a suitable alternative to the ATE for determining subpopulations of interest that benefit from a given treatment. Researchers have been focusing on estimating PTE at each predictive covariate point, that is, a set of baseline characteristics that predicts the patient's response to a particular treatment. In a regression model, predictive covariates are incorporated in treatment–covariate interaction terms, and a hypothesis test of a null PTE is considered for each predictive covariate point. Two main issues with this approach are high multiplicity and low power to detect a treatment–covariate interaction [10–13]. In addition to these issues, Pocock et al. [3] points out that biological plausibility should be assessed along with consideration of the strength of evidence for heterogeneity in the treatment effect.

In this paper, we develop a Bayesian approach for subgroup analysis with time–to–event data based on recent advances in subgroup identification methodology proposed by Schnell et al. [14–16]. In a Bayesian framework, Schnell et al. [14] provide a two-step procedure to estimate a benefiting subgroup: (1) fit a regression model, and (2) construct bounding subgroups based on the posterior distribution of PTEs. Compared to previous methods, Schnell et al.'s method has several advantages, such as controlling for multiplicity and easily making statistical inferences from the full posterior distribution of the PTEs. This construction furnishes a pair of credible subgroups: one that is likely to be contained by the benefiting subgroup and one that is likely to contain the benefiting subgroup. The corresponding inferential statement is that every type of patient in one bounding subgroup benefits, and no type of patient outside the other subgroup benefits. These inferences are simultaneous [14] in contrast to non-simultaneous inferences available from tree-based methods. Here the simultaneous inferences mean that all covariate points corresponding to a specific subpopulation simultaneously have a treatment effect exceeding a specified threshold.

Inspired by the two–step procedure, our approach to identify benefiting subgroups for time–to–event endpoints is to define a Cox proportional hazard model and make statistical inferences from the full posterior distribution of the HR between two arms. In a randomized clinical trial with a time-to-event endpoint, HR is the common efficacy measure which represents the relative difference between two survival curves based on the proportional hazard (PH) assumption. When the PH assumption is violated, the HR does not appropriately represent the PTEs. An alternative approach is the restricted mean survival time (RMST) which is the area under the survival curve up to a particular time point. As shown in previous studies [17, 18], RMST is a robust and clinically interpretable measure of the survival time distribution without PH assumption. Moreover, Uno et al. [19, 20] advocate for using RMST to estimate treatment effects as an alternative to the HR. In our approach, we consider the difference in RMST (RMSTd) between two randomized arms at a certain follow–up time point as the PTEs.

Our methods are also tested on two time–to–event datasets: One from a trial in patients with prostate cancer, and another from a simulated Merck Sharp & Dohme, London, UK (MSD) clinical trial in patients with myocardial infarctions. The first dataset is publicly available [21] and has been analyzed in Ballarini et al.'s study [9]. Our findings are similar to those found by Ballarini et al. [9], but, in addition, our methods also identify the non–benefiting

subgroups to the treatment. The second dataset is a simulated data based on a randomized and placebo-controlled study on the effect of vorapaxar in addition with aspirin for secondary prevention of thrombotic events. Scirica et al. [22] applied a Cox proportional hazards model for testing heterogeneous HRs across prespecified subgroups of interest. Our primary interest is searching for benefiting subgroups from vorapaxar treatment, without prespecification of subgroups for testing but only based on covariates which may have predictive value.

Our proposed method is an extension of Schnell et al. [14] to time–to–event endpoints. It is important for practical applications and widens significantly application area of this work. The most important property of our method is handling multiplicity, because it enables control the familywise Type I error rate and thus explicitly controls the probability of making any spurious claims of subgroup benefits. Our method is amenable to the confirmatory setting, rather than the methods that are focused on potential benefiting group discovery. For the discovery, the multiplicity issues are not critical as it would be addressed in subsequent confirmatory analyses. Therefore, these two approaches are complementary to each other.

The organization of the rest of the paper is as follows: In Section 2 we present the Bayesian credible subgroup method for time–to–event endpoint with Section 2.1 introducing the notation and PTEs, defining the log HR and the difference in RMST as PTEs in Section 2.2 and 2.3 respectively. A simulation study is provided in Section 3 which implements our Bayesian credible subgroup with these two difference PTEs. Finally, a detailed analysis of two clinical datasets are presented in Section 4 and 5, while our conclusions are discussed in Section 6. This paper also has accompanying supplementary material containing details simulation study on the performance of HR and difference in RMST.

2 Methods

For the purpose of identifying a benefiting subgroup, traditional approaches begin with a test for treatment–covariate interactions [3, 23]. These approaches have well–known limitations including low power due to the smaller sample size within subgroups and multiplicity adjustment for a larger number of subgroups under investigation. More importantly, results of interaction tests do not directly answer the question of which types of patients (covariates profiles) benefit from treatment. Rejecting the no interaction hypothesis detects treatment heterogeneity, but it does not provide you with the information of which covariate points correspond to positive conditional average treatment effect. To overcome these difficulties, in a Bayesian framework, we introduce the method of credible subgroups to simultaneously identify which types of patients benefit from treatment.

2.1 Notation and personalized treatment effect (PTE)

For an event time T, let x be a $p \times 1$ vector of prognostic covariates, and z be a $q \times 1$ vector of predictive covariates. Some covariates may appear in both x and z, and intercept terms may be included. Suppose that T is only partially observed during an experiment due to censoring, such as right censoring denoted by a random variable C. We assume that T is independent of C given x. Let $Y = \min(T, C)$ and κ be a failure indicator, i.e. $\kappa = 1$ for $T \leq C$ and 0 otherwise (See S1 File Moreover, we consider the time–to–event data to consist of n subjects who were randomly assigned to one of two treatments, i.e. $\theta = \{0, 1\}$, and only predictive covariates z interact with treatment indicator θ in our model fit. In this scenario, the observed data consist of n independent realizations of $\{(Y_i, x_i, z_i, \kappa_i, \theta_i)\}$ for $i = 1, \ldots, n$.

For a two–arm study with censoring, a common non–parametric approach to compare the survival distributions between two treatment groups is the log rank test. This test is based on a series of 2×2 contingency tables constructed at each observed failure time. A semi–parametric

approach using the Cox regression model is commonly applied to investigate the effect of covariates on the HR. These two approaches measure the average treatment effect on the entire study population, which cannot be used to identify which patient benefits from a treatment. Personalized treatment effects (PTEs) are becoming widely used to determine subgroups of patients who most benefit from a treatment. We denote Δ as a PTE for a subject with covariate vectors x_i and z_i, and precisely define Δ for the log HR and RMST differences as measures for PTE in the following sections.

2.2 The log HR as a PTE

The Cox model, which is commonly used for the analysis of time–to–event data, has the following form:

$$\lambda(t|x_i, z_i, \theta_i) = \lambda_0(t) \exp(x_i'\beta + \theta_i z_i'\gamma), \tag{1}$$

where $\lambda(t|x_i, z_i, \theta_i)$ is the hazard function at time t for the ith subject with covariates x_i and z_i, $\lambda_0(t)$ is the unspecified baseline hazard function, and β and γ are $p \times 1$ and $q \times 1$ vectors of regression coefficients, respectively. Here x' denotes the transpose of vector x. We include the interaction terms between z and θ in the model, and the PTE for a patient with covariate z is

$$\Delta_H(z_i) = \frac{\lambda(t|x_i, z_i, \theta_i = 1)}{\lambda(t|x_i, z_i, \theta_i = 0)} = \exp(z_i'\gamma), \tag{2}$$

which is a ratio between the hazards of a patient with treatment $\theta = 1$ and $\theta = 0$.

If a given element of γ is positive, then higher values of the corresponding element of z would indicate that the subject has a higher hazard for treatment $\theta_i = 1$ or shorter survival than the subject with treatment $\theta_i = 0$. Therefore, to determine the characteristics of subjects who benefit from treatment $\theta_i = 1$, we set a predetermined threshold of clinical significance $0 < \delta_H \leq 1$, and search for the points z_i such that

$$\Delta_H(z_i) = \exp(z_i'\gamma) < \delta_H. \tag{3}$$

Alternatively,

$$\log \Delta_H(z_i) = z_i'\gamma < \log \delta_H, \tag{4}$$

where $\log\Delta_H(z_i)$ is the log HR evaluated at points z_i. In subgroup analysis for time-to-event data, we want to find a subgroup in which *every* covariate point for a subject has a conditional log HR less than log δ_H. This approach is distinguished from finding a subgroup whose overall log HR is less than log δ_H in the sense that such subgroup can contain members with higher log HR than log δ_H.

From Eq (2), the HR from the Cox regression model between two subjects is constant over time. This assumption often fails in time-to-event data. For example, non-proportional hazard is present in immuno-oncology trials due to delayed treatment effect and/or functional cure. When the proportional hazard assumption does not hold, the $\Delta_H(z_i)$ may no longer provide suitable PTE for a subject. An alternative approach is to use the RMST which is a robust measure of the survival time distribution without relying on the proportional hazard (PH) assumption. We present the RMST in the next section and describe how it may be used to define the PTE for a subject.

2.3 Difference in restricted mean survival times (RMSTd) as a measure of PTE

The RMST ψ of a random variable failure time T is the mean of the time–to–event $\zeta = \min(T, v)$ limited to some cutoff time point $v > 0$. In other words, the RMST is the area under the survival curve $S(t)$ between $t = 0$ to $t = v$ and can be expressed as

$$\psi = E(\zeta) = \int_0^v S(t)\, dt, \tag{5}$$

In a randomized two–arm clinical trial, let $S(t|\theta = 1)$ and $S(t|\theta = 0)$ be the survival functions for the treatment $\theta = 1$ and $\theta = 0$ respectively. The RMSTd between two arms from $t = 0$ to $t = v$ is defined as

$$\Delta_{Rd} = \psi_{\theta=1} - \psi_{\theta=0} = \int_0^v [S(t|\theta = 1) - S(t|\theta = 0)]\, dt, \tag{6}$$

which is the difference in area between the two survival curves. Alternatively, one can also define the ratio of RMST between two arms such as $\Delta_{Rr} = \frac{\int_0^v S(t|\theta=1)\ dt}{\int_0^v S(t|\theta=0)\ dt}$.

In this paper, we use the Δ_{Rd} as the PTE for a subject, and estimate the two survival functions $S(t|\theta = 0)$ and $S(t|\theta = 1)$ in Eq 6 on the grid of subgroup–defining covariates in order to compute the Δ_{Rd}. A common approach is to obtain a Kaplan–Meier estimator, but that would not take the covariates into account. Thus we employ the conventional Cox proportional hazard model to estimate these two survival functions. Note that we could still employ fixes to PH violations (e.g., time-dependent covariates or effects) without having to worry about reporting time-dependent hazard ratios.

Moreover, if T is years to death and $S(t|\theta = 1) > S(t|\theta = 0)$ for $t \in [0, v]$, the interpretation of Δ_{Rd} is that a subject has the v-year life expectancy higher in treatment $\theta = 1$ than $\theta = 0$, so this subject could be benefiting from treatment $\theta = 1$. Similar to the HR, to identify benefiting subjects from treatment $\theta = 1$, we set Δ_{Rd} to be greater than some predetermined threshold of clinical significance $\delta_R > 0$. Compared to Δ_H, the advantage of Δ_{Rd} would not rely on the PH assumption. However, if it were to hold, possible questions of interest would be "is there relationship between the two PTE measurements?" and "if the investigators knows δ_H, what is the corresponding δ_R (or vice versa)?" To address these questions, we show in S1 File that in parametric settings and when $\delta_H = 1$ and $\delta_R = 0$, Δ_H and Δ_{Rd} are the same for determining whether a patient benefits from the treatment. Thus we use these predetermined significance values in our simulation study (Section 3). In the following section, we show how these PTE measurements are related to subgroup identification problem in the Bayesian framework.

2.4 Bayesian credible subgroups for time-to-event data

There are numerous non-parametric and parametric methods for PTEs to identify who benefits from a treatment given their baseline characteristics. Among them, Berger et al. [23] proposed a Bayesian model selection using tree–based priors that provide the posterior distribution for use in statistical inference. Recently, Ballarini et al. [9] developed the predicted individual treatment effect (PITE) using a multiple regression framework with a Lasso–type penalty for model selection, and provided confidence intervals for the PTEs. Subgroup identification will be evaluated based on these confidence intervals. While the inferences from tree–based methods are not simultaneous, the PITE methods does not address the multiplicity issue. To overcome these limitations, Schnell et al. [14] proposed a Bayesian credible

subgroups methods for continuous endpoints which addressed both simultaneous inference and multiplicity. In this paper, we extend their approach to survival endpoints by using two summaries commonly used in clinical trials: the log HR and RMSTd. In the following sections, we first present the Bayesian credible subgroups methods, and introduce a Bayesian approach to obtain the inference for the time–to–event PTEs. Then we construct simultaneous credible bands from those inferences.

2.4.1 Bayesian credible subgroups. A goal of the Bayesian credible subgroups method [14] is to estimate a set of subject baseline covariate points for which a subject would benefit from the treatment according to a PTE. More precisely, let $\mathcal{Z}$ be a covariate space, this approach searches for the set of covariate points $B_H = \{z \in \mathcal{Z} : \Delta_H(z) < \log(\delta_H)\}$ when the PTE is measured as the log HR, or $B_{Rd} = \{z \in \mathcal{Z} : \Delta_{Rd}(z) > \delta_R\}$ for the RMSTd. In a Bayesian framework, a common estimator for B includes the points $z \in \mathcal{Z}$ whose posterior probability of having the log HR less than $\log(\delta_H)$ (or greater than δ_R for Δ_{Rd}) given observed data is greater than $(1 - \alpha)$, where $1 - \alpha$ is a credible level, and can be expressed as

$$\hat{B}_{H,\alpha} = \{z \in \mathcal{Z} : P(\Delta_H(z) < \delta_H \mid \text{Data}) > 1 - \alpha\} \tag{7}$$

for the PTE of log HR or

$$\hat{B}_{Rd,\alpha} = \{z \in \mathcal{Z} : P(\Delta_{Rd}(z) > \delta_R \mid \text{Data}) > 1 - \alpha\} \tag{8}$$

for the PTE of RMSTd. To control for multiplicity, the credible subgroup pair (D, S) consists of an exclusive credible subgroup D and inclusive credible subgroup S such that $P(D \subseteq B \subseteq S|\text{Data}) > 1 - \alpha$. This means that with posterior probability $(1 - \alpha)$, D contains only covariate points z for which the types of subjects benefit from the treatment, while S includes all types of subjects who benefit.

Fig 1 illustrates the covariate space $\mathcal{Z}$ divided into three regions. The region B enclosed by dashed circle represents the true benefiting subgroup which we wish to estimate by a pair (D, S). Here the green region D includes the types of patients who have evidence of benefit whereas patients in the red region S^C have no benefit. Finally, the blue region is a region of uncertainty that needs more information.

The Bayesian credible subgroups method described above is a two-step procedure: (1) define a model, fit a regression and obtain the marginal posterior of the PTE onto the given covariates; and (2) compute the bounds and obtain a pair (D, S). In the first step, we fit a Cox regression model in a Bayesian framework to get the marginal posterior of coefficients of predictive covariates that interact with treatment choice (in Section 2.4.2). In the second step, we describe the method to compute the credible subgroups (in Section 2.4.3).

2.4.2 Bayesian estimation in time–to–event analysis. Several methods have been proposed for Bayesian analysis of the Cox proportional hazards model with right censored data [24–26]. In this section, we review the commonly used gamma process prior for the Cox regression model. In Eq 1, we need to specify priors on $\boldsymbol{\beta} = (\beta', \gamma')'$ and the baseline hazard function $\lambda_0(t)$. A typical prior for β is a $N_{p+q}(\mu_0, \Sigma_0)$ which is a $p + q$-variate normal distribution with mean vector μ_0 and covariate Σ_0. We choose a nonparametric gamma process prior on the cumulative baseline hazard. This prior partitions the observed survival time into intervals, such as $0 < s_1 < s_2 < \ldots < s_J$ where $s_J > y_i$ for $i = 1, \ldots, n$. Thus, we have J disjoint intervals. The observed data $\mathcal{D} = \{x, z, R_j, M_j \text{ for } j = 1, \ldots, J\}$ will be grouped within these intervals where R_j and M_j are the risk set and failure set of the jth interval $(s_{j-1}, s_j]$, respectively. Let $h_j = H_0(s_j) - H_0(s_{j-1})$ be the increment in the cumulative baseline hazard

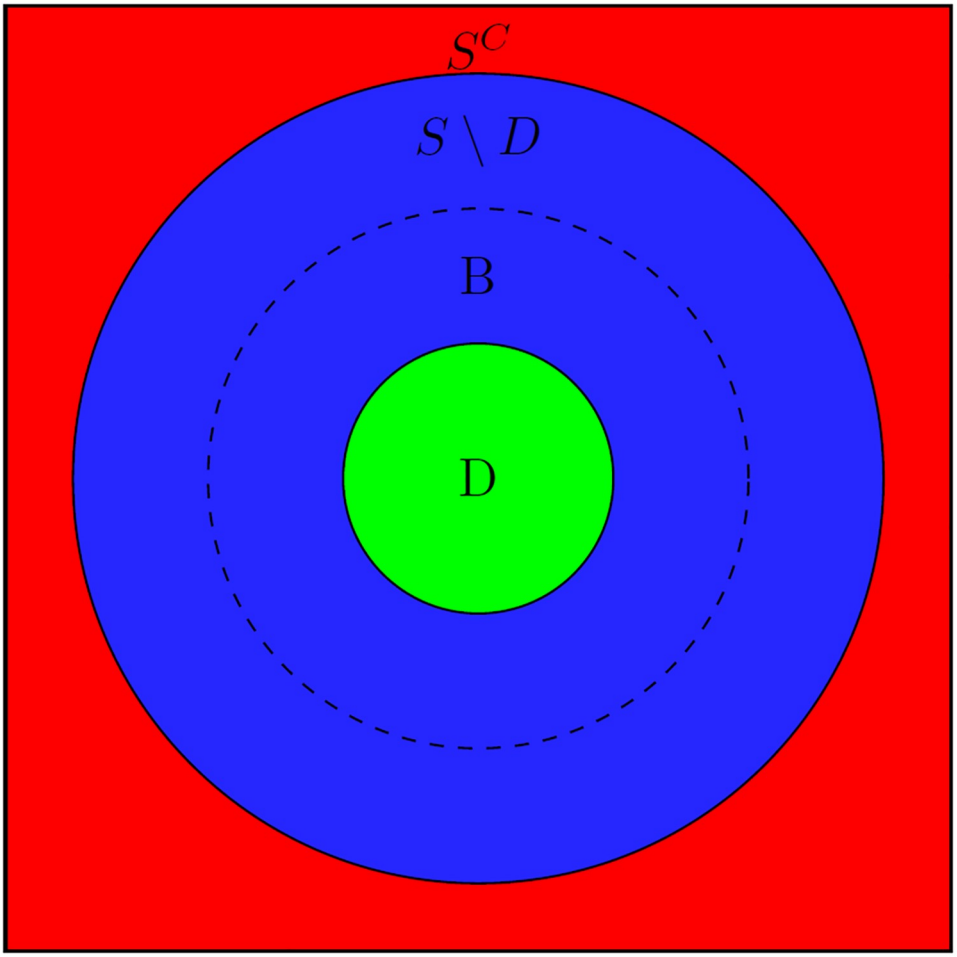

Fig 1. Illustration of credible subgroups. B contains true type of patients who benefit (enclosed by dashed line) while D includes only type of patients who benefit (green). Moreover, type of patients in $S\backslash D$ require more information (blue), and those in S^C have no benefit (red).

https://doi.org/10.1371/journal.pone.0229336.g001

$H_0(t) = \int_0^t \lambda_0(k)\, dk$ in the jth interval. Thus, the h_j's are independent increments in disjoint intervals and

$$h_j \sim \mathcal{G}(\alpha_j - \alpha_{j-1}, b), \tag{9}$$

where $\mathcal{G}(\alpha, b)$ denote the gamma distribution with shape parameter $\alpha > 0$ and scale parameter $b > 0$, such that $\alpha_j = bH(s_j)$ with an increasing function $H(s_j)$. H and b are hyperparameter for h_j, and H_b is a specified parametric cumulative hazard function evaluable at the endpoints of the time intervals, and the scalar b is a weight about the mean. Therefore, the observed likelihood function is

$$L(\boldsymbol{\beta}, \mathbf{h} \mid \mathcal{D}) \propto \prod_{j=1}^{J} l_j, \tag{10}$$

where $l_j = \exp\left[-h_j \sum_{k \in R_j M_j} \exp\left(\mathbf{x}_k' \boldsymbol{\beta}\right)\right] \prod_{m \in M_j} [1 - \exp\left(-h_j \exp\left(\mathbf{x}_m' \boldsymbol{\beta}\right)\right)]$, and $\mathbf{x}$ is a combined vector of $(x_i', \theta z_i')'$.

From the prior distribution for $\boldsymbol{\beta}$ and $\mathbf{h} = (h_1, \ldots, h_J)'$ described above, the joint posterior of $\boldsymbol{\beta}$ and $\mathbf{h}$ is

$$P(\boldsymbol{\beta}, \mathbf{h} \mid \mathcal{D}) \propto \prod_{j=1}^{J} \left[l_j h_j^{\alpha_j - \alpha_{j-1} - 1} \exp(-b h_j) \right] \exp\left(\frac{-1}{2} (\boldsymbol{\beta} - \mu_0) \Sigma_0^{-1} (\boldsymbol{\beta} - \mu_0) \right). \tag{11}$$

From Eq 11, the conditional distribution of β_i given $\mathbf{h}$, and $\boldsymbol{\beta}^{(-i)}$ denoted the $\boldsymbol{\beta}$ vector without the ith component is

$$P\left(\beta_i \mid \boldsymbol{\beta}^{(-i)}, \mathbf{h}, \mathcal{D}\right) \propto \prod_{j=1}^{J} l_j \exp\left(\frac{-1}{2} (\boldsymbol{\beta} - \mu_0) \Sigma_0^{-1} (\boldsymbol{\beta} - \mu_0) \right). \tag{12}$$

where $i = 1, \ldots, p + q$. Similarly, the conditional distribution of h_j is

$$P(h_j \mid \mathbf{h}^{(-j)}, \boldsymbol{\beta}, \mathcal{D}) \propto h_j^{\alpha_j - \alpha_{j-1} - 1} \exp\left[-h_j \left(\sum_{k \in R_j} M_j \exp(\mathbf{x}_k' \boldsymbol{\beta}) + b \right) \right]. \tag{13}$$

The posterior distribution of the parameter of interest can be obtained from these full conditional distributions in Eqs 12 and 13 by Gibbs sampling in a straightforward way. Based on the posterior of coefficients of predictive covariates, we can obtain the posterior of $\Delta_H(z)$. For $\Delta_{Rd}(x, z)$, we first compute the posterior distribution for survival function $S(t) = \exp(-H_0(t) \exp(\mathbf{x}' \beta))$ for each treatment and then obtain the difference between two treatments (See S1 File for more details of constructing the posterior of $\Delta_H(Z)$ and $\Delta_{Rd}(x, Z)$). We implement our proposed method from an R package **spBayesSurv** provided by Zhou et.al. [27]. For ease of notation, we denote $\Delta(z)$ for a general PTE and proceed to construct credible subgroups in the next section. **Remark**: We presented Bayesian estimation of the Cox regression model by using the gamma process prior. However, there are several advanced Markov chain Monte Carlo sampling techniques that were proposed that could be used in this context. They include slice sampling [28] and Hamiltonian Monte Carlo [29]. There are also associated software packages available, e.g. Stan [30], or the R packages such as MfUSampler [31], and sns [32] which provide a wider range of model specifications and which can be used for Bayesian survival analysis. A notable example represents the R package BSGW [33].

2.4.3 Credible subgroups estimation. A goal of constructing credible subgroups based on the posterior of $\Delta(z)$ is to control the multiplicity in testing $\Delta(z)$ at every covariate point and to provide two credible subgroups (D, S) which bound the benefiting subgroup B. These simultaneous credible bands over the covariate space $\mathcal{Z}$ can be constructed as

$$\Delta(z) \in \hat{\Delta}(z) \pm \sqrt{W_\alpha \mathrm{Var}(\Delta(z))}, \tag{14}$$

where W_α is the $1 - \alpha$ quantile of the distribution of $W = \sup_{z \in \mathcal{Z}} \frac{(\Delta(z) - \hat{\Delta}(z))^2}{\mathrm{Var}(\Delta(z))}$ and $\hat{\Delta}(z)$ is the posterior mean of $\Delta(z)$. Therefore, in a case of $\Delta(z) \equiv \Delta_H(z)$, the exclusive credible subgroup D is given by

$$D = \{ z \in \mathcal{Z} : \hat{\Delta}_H(z) + \sqrt{W_\alpha \mathrm{Var}(\Delta(z))} < \delta_H \} \tag{15}$$

and inclusive credible subgroup S is

$$S = \{ z \in \mathcal{Z} : \hat{\Delta}_H(z) - \sqrt{W_\alpha \mathrm{Var}(\Delta(z))} \leq \delta_H \}. \tag{16}$$

Note that $\Delta(z)$ was sampled from the exact posterior as described in previous section, and then we used Gaussian approximation to obtain the simultaneous credible bands in

Eqs 15 and 16. In this paper, we construct the asymptotic credible bands since the posterior distributions for representative covariate points in our applications were approximately Gaussian (See S1 File). Moreover, Schnell et al. [15] proposed a quantile–based simultaneous credible band when $\Delta(z)$ is non–Gaussian.

3 Simulation study

In this section, we conduct extensive simulation studies to evaluate the performance of Bayesian credible subgroups in simulated time–to–event data under the PH and non PH assumptions.

3.1 Simulation study under proportional hazard assumption

For the Cox proportional hazard model, we assume that the hazard function for the i^{th} individual ($i = 1, \ldots, n$) is

$$\lambda_i(t) = \lambda_0(t) \exp(x_i'\beta + \theta_i z_i'\gamma), \tag{17}$$

where $x_i = (x_{i1}, x_{i2})'$ and $z_i = (1, z_{i1}, z_{i2})'$ are the vectors of prognostic and predictive covariates respectively. Then $\beta = (\beta_1, \beta_2)'$ and $\gamma = (\gamma_1, \gamma_2, \gamma_3)'$ are vector of coefficients of x_i and z_i, respectively. Moreover, we assume a Weibull baseline hazard, i.e. $\lambda_0(t) = \lambda\nu(\lambda t)^{\nu-1}$ where λ and ν are the scale and shape parameters respectively. Then the inverse of the cumulative hazard function is $H_0^{-1}(t) = (\lambda^{-1}t)^{1/\nu}$. If U is uniformly distributed on [0, 1], the survival time T_i can be generated as

$$T_i = H_0^{-1}\left[-\frac{\log(U)}{\exp(x_i'\beta + \theta_i z_i'\gamma)}\right] = -\left[\frac{\log(U)}{\lambda \exp(x_i'\beta + \theta_i z_i'\gamma)}\right]^{1/\nu}. \tag{18}$$

Suppose that C_i are the censoring times, drawn from an exponential distribution Exp(a). Due to censoring, we observe $Y_i = \min(T_i, C_i)$ and censoring indicators κ_i. For parameters of the simulation time–to–event data, we set $\lambda = 0.05$ and $\nu = 1.1$ for a Weibull baseline hazard and a rate $a = 0.02$ for the censoring time. Furthermore, we let $x_{i1} = z_{i1} = \{0, 1\}$ with equal probability, and $x_{i2} = z_{i2}$ be uniformly distributed on the interval (−3, 3), and we only consider two arms, i.e. $\theta_i = \{0, 1\}$.

Then we perform diagnostic test for credible subgroups with different sample sizes ($n = 50$, 100, 500, 1000) and at different credible levels (0.4, 0.6, 0.8, 0.95). Finally, we consider three cases of β with different values of γ:

1. The prognostic features have no effect $\beta = (0, 0)$, and we set $\gamma = (0, 0, 0)$, (0.1, 0.1, 0.1), (1, 1, 1), (1, −1, 3),
2. The prognostic features have moderate effect $\beta = (0.2, 0.2)$, and we set $\gamma = (1, 1, 1)$,
3. The prognostic features have higher effect $\beta = (1, -2)$, and we set $\gamma = (1, 0.1, 1)$.

Following the same criteria in Schnell et al.'s simulation study [14], we report five metrics: (1) total coverage measures the frequency with which $D \subseteq B \subseteq S$ for some fixed value z; (2) the pair size measures the proportion of covariate points in the uncertainty region $S \backslash D$; (3) specificity and sensitivity of D measures how well the credible subgroup D aligns with benefiting subgroup B; and (4) mean squared error (MSE) of the treatment effects compares the estimated treatment effects to the true values.

As shown in S1 File, the PTEs measured by the log HR and RMSTd yield the same result when $\delta_H = 1$ and $\delta_R = 0$. We provide three simulation studies as follows:

- Simulation 1: run using the same simulation time–to–event data for log HR and RMSTd when $\delta_H = 1$ and $\delta_R = 0$.
- Simulation 2: run only for log HR at different thresholds $\delta_H = \{0.2, 0.5, 1, 2\}$.
- Simulation 3: only for the RMSTd at $\delta_R = \{-1, 0, 1\}$.

For each simulation study, we simulate 1000 data sets, and for each data set, we use 1000 posterior draws kept after 500 burn-in iterations. Table 1 reports the average summary statistics for Simulation 1 at an 80% credible level. When the effect sizes are relatively small, the benefiting subgroup is empty, and sensitivity of *D*, which is the proportion of the benefiting subgroup *B* included in the exclusive credible subgroup *D*, is not calculable. This is represented with 'NaN' in the table. Overall, we find that the total coverage is always greater than 80% for both PTE approaches. When the sample size and/or effective size are increasing, the credible size is decreasing except when β's are zero. The RMSTd approach has larger credible size than log HR for $n = 50$ and 100 but smaller or similar credible size for larger n. Moreover, both PTE approaches have similar specificity of *D*, which is the proportion of the non–benefiting subgroup *B* not included in the exclusive credible subgroup *D*. Compared to the RMSTd approach, the log HR tends to have higher sensitivity of *D* for a small sample size ($n = 50$) but low sensitivity for a large sample size ($n = 1000$). Finally, the RMSTd approach has small MSE in most scenarios. In general, both approaches show similar trends. Simulation results for other simulation settings, and comparison between the proposed Bayesian credible subgroup

Table 1. Simulation 1 results: Average summary statistics for 80% credible level.

Sample size	Truth	Total Coverage		Credible Pair Size		Sensitivity of D		Specificity of D		MSE	
		log HR	RMSTd	log HR	RMSTd	log HR	RMSTd	log HR	RMSTd	log HR	RMSTd
50	(0,0,0,0,0)	0.88	0.88	0.94	0.93	NaN	NaN	0.97	0.97	0.44	0.48
	(0,0,0.1,0.1,0.1)	0.9	0.86	0.94	0.93	0.04	0.04	0.99	0.98	0.46	0.51
	(0,0,1,1,1)	0.92	0.94	0.32	0.38	0.45	0.47	0.99	0.99	0.46	0.38
	(0,0,1,-1,3)	0.92	0.94	0.12	0.19	0.97	0.97	0.99	0.99	0.96	0.28
	(0.2,0.2,1,1,1)	0.9	0.92	0.31	0.45	0.45	0.44	0.99	0.99	0.43	0.36
	(1,-2,1,0.1,1)	0.96	0.97	0.35	0.55	0.52	0.12	1	1	0.59	0.26
100	(0,0,0,0,0)	0.9	0.89	0.95	0.95	NaN	NaN	0.98	0.98	0.18	0.25
	(0,0,0.1,0.1,0.1)	0.89	0.91	0.9	0.91	0.08	0.08	0.99	0.99	0.2	0.26
	(0,0,1,1,1)	0.92	0.92	0.21	0.21	0.72	0.73	0.99	0.99	0.22	0.22
	(0,0,1,-1,3)	0.91	0.92	0.08	0.08	1	1	0.99	0.99	0.49	0.21
	(0.2,0.2,1,1,1)	0.89	0.9	0.21	0.22	0.72	0.71	0.99	0.99	0.23	0.21
	(1,-2,1,0.1,1)	0.96	0.96	0.21	0.3	0.79	0.57	1	1	0.28	0.2
500	(0,0,0,0,0)	0.88	0.9	0.95	0.95	NaN	NaN	0.97	0.98	0.03	0.05
	(0,0,0.1,0.1,0.1)	0.94	0.92	0.77	0.77	0.08	0.08	1	1	0.03	0.05
	(0,0,1,1,1)	0.9	0.92	0.13	0.13	1	1	0.99	0.99	0.08	0.06
	(0,0,1,-1,3)	0.88	0.86	0.06	0.05	1	1	0.98	0.98	0.31	0.05
	(0.2,0.2,1,1,1)	0.94	0.92	0.13	0.13	1	1	0.99	0.99	0.08	0.06
	(1,-2,1,0.1,1)	0.92	0.92	0.12	0.12	1	1	0.99	0.99	0.07	0.07
1000	(0,0,0,0,0)	0.9	0.89	0.97	0.96	NaN	NaN	0.98	0.98	0.01	0.03
	(0,0,0.1,0.1,0.1)	0.94	0.94	0.59	0.58	0.15	0.18	1	0.99	0.02	0.03
	(0,0,1,1,1)	0.88	0.88	0.13	0.13	1	1	0.99	0.99	0.06	0.03
	(0,0,1,-1,3)	0.87	0.88	0.06	0.06	1	1	0.98	0.98	0.28	0.03
	(0.2,0.2,1,1,1)	0.86	0.89	0.12	0.13	1	1	0.98	0.99	0.07	0.04
	(1,-2,1,0.1,1)	0.93	0.93	0.11	0.11	1	1	0.99	0.99	0.05	0.04

https://doi.org/10.1371/journal.pone.0229336.t001

with the pointwise method, i.e. without multiplicity correction [14], are also reported in S1 File. We found that moving from pointwise method to our proposed methods, there is increasing in credible pair size, specificity of D, but smaller sensitivity of D.

Multiplicity is incorporated into the simulation framework as the "total coverage" metric. Total coverage is the rate at which the true benefiting subgroup is both contained in the inclusive credible subgroup and contains the exclusive credible subgroup. A coverage failure corresponds to a family–wise error. The credible subgroup method should have a total coverage rate equal to the credible level, whereas a method not accounting for multiplicity would have lower total coverage due to that multiplicity.

3.2 Simulation study under nonproportional hazard assumption

When the PH assumption is violated, the HR may not accurately represent PTEs, so RMST summaries are used to estimate PTEs as an alternative approach to the HR. The simulation study in this section aims to investigate the performance of RMSTd for subgroup identification in a case of the nonproportional hazard.

From Eq 17, we simulated two groups with different hazard rates. The treatment group, i.e. $\theta_i = 1$, had a constant exponential hazard with rate $\lambda_0(t) = 0.01$. The control group, i.e. $\theta_i = 0$, had a piecewise exponential hazard with rate

$$\lambda_0(t) = \begin{cases} 0.01 & 0 \leq t < t_c \\ 0.1 & t_c \leq t. \end{cases} \tag{19}$$

Under this nonproportional hazard model, the hazard ratio between two treatments for subject i is $\exp(z_i'\gamma)$ until time t_c and then there is an abrupt change to a rate of $\exp(z_i'\gamma)/10$.

The first step of determining the two bounded subgroup pairs is to obtain the joint posterior sample of the PTEs at each covariate points. For the RMSTd, we employ a fully nonparametric Bayesian accelerated failure time (AFT) model proposed by Henderson et. al. [34]. It directly models the log-failure time as a sum of a regression function of covariates and residual. The conditional mean function is modeled using Bayesian additive regression trees (BART). The residual is modeled using a location-mixture of Gaussian distributions with a centered Dirichlet process as prior. We compute the survival functions of the non–parametric AFT model for each treatment, then take the difference between these survival functions to obtain the RMSTd at each covariate point.

The settings for the prognostic covariates x, the predictive covariates z, treatment indicator θ and censoring rate are similar to settings in simulation study under PH assumption. We considered the true values of coefficients $\beta = (0.7, 0.7)$ and $\gamma = (0.5, -0.5, -0.5)$. Then we simulated 1000 data sets, and for each data set, we used 1000 posterior draws kept after 500 burn-in iteration. Finally, we chose $\delta_{Rd} = 0$ and $t_c = 30$. The RMSTd were computed at the change point t_c up to $t_c + 50$ when the treatment effect shows up during this time interval. The results for 80% credible subgroup pairs are presented in Table 2, and S1 File provides results for other credible levels.

Table 2. Average summary statistics for 80% credible subgroup pairs under nonproportional hazard assumption.

Sample Size	Total Coverage RMSTd	Credible Pair Size RMSTd	Sensitivity of D RMSTd	Specificity of D RMSTd	MSE RMSTd
50	0.76	0.5	0.52	0.76	0.68
100	0.79	0.36	0.68	0.79	0.58
500	0.86	0.17	0.88	0.86	0.43
1, 000	0.92	0.15	0.91	0.92	0.38

https://doi.org/10.1371/journal.pone.0229336.t002

When the sample size increases, the credible pair size is smaller, and there is greater total coverage and improved sensitivity and specificity of D. Moreover, total coverage is above 80% for large sample size (n = 500, 1000) and close to 80% for small and moderate sample size (n = 50, 100). Table 2 also shows that a RMSTd approach tends to well estimate PTEs with respect to effect MSE. The results of this simulation suggest the RMSTd approach is appropriate to identify benefiting subgroups in a case of nonproportional hazards.

4 Analysis of the prostate cancer dataset

A prostate cancer dataset is publicly available [21] and has been analyzed in Rosenkranz et al. [35] for exploratory subgroup analysis by model selection. Ballarini et al. [9] have proposed the predicted individual treatment effect (PITE) to identify subgroups of patients who benefit from treatment. In this section, we illustrate the Bayesian credible subgroups method using the prostate cancer dataset, and compare our results with published results using PITE [9].

We include the 475 patients with complete data in our analysis as in the previously published studies. Each subject was randomly assigned either to a combination of placebo and the lowest dose level of diethyl stilbestrol (control group) or higher doses (treatment group). Moreover, we included the same covariates and interaction terms as used in Ballarini et al. [9]: existence of bone metastasis (bm), disease stage either 3 or 4 (stage), performance (pf), history of cardiovascular events (hx), age and weight (wt). We denote rx as treatment indicator, and include the two important interactions in the model, i.e. bm:rx and age:rx as in [9]. Table 3 provides the posterior mean and posterior standard deviation of the coefficients. We found that stage and age are not significant at a nominal 95% credible level, but we have a strong interaction with treatment of bone metastasis, and age.

The left panel in Fig 2 shows credible subgroups, for prostate cancer patients, using a log HR and a credible level of 95%. We used the same value $\delta_H = 1$ to define subgroups ([9]). Each bar in each panel represents a particular type of patient with their age and existence of bone metastasis. Members in the green region (D) are benefiting from the treatment. The blue uncertainty region ($S\backslash D$) contains characteristics of patients who are or are not be benefiting from the treatment. The red region (S^C) indicates that these types of patients may not be benefiting from the treatment.

Similarly, the right panel in Fig 2 shows credible subgroups using the RMSTd with a credible level of 95% and $\delta_R = 0$. Notice that the uncertainty region of Δ_{Rd} is smaller than the uncertainty region of Δ_H in patients with existence of bone metastasis. For both summaries, we found that when patients were younger than 67 years old and did not have any bone metastases, are benefiting to the treatment. Compared with the analysis reported by Ballarini et al. [9],

Table 3. Posterior summaries of coefficients of covariates in prostate cancer dataset. The * indicates that the estimates are greater than 1.96 standard errors from 0. This is equivalent to 95% level.

Effect	Posterior Mean	Posterior SD	Significance
bm	0.575	1.170	*
stage	-0.012	0.470	
pf	0.178	0.773	*
hx	0.191	0.608	*
age	-0.021	0.018	
weight	-0.017	-0.001	*
rx	-6.085	-1.838	*
bm:rx	-1.125	-0.291	*
age:rx	0.026	0.082	*

https://doi.org/10.1371/journal.pone.0229336.t003

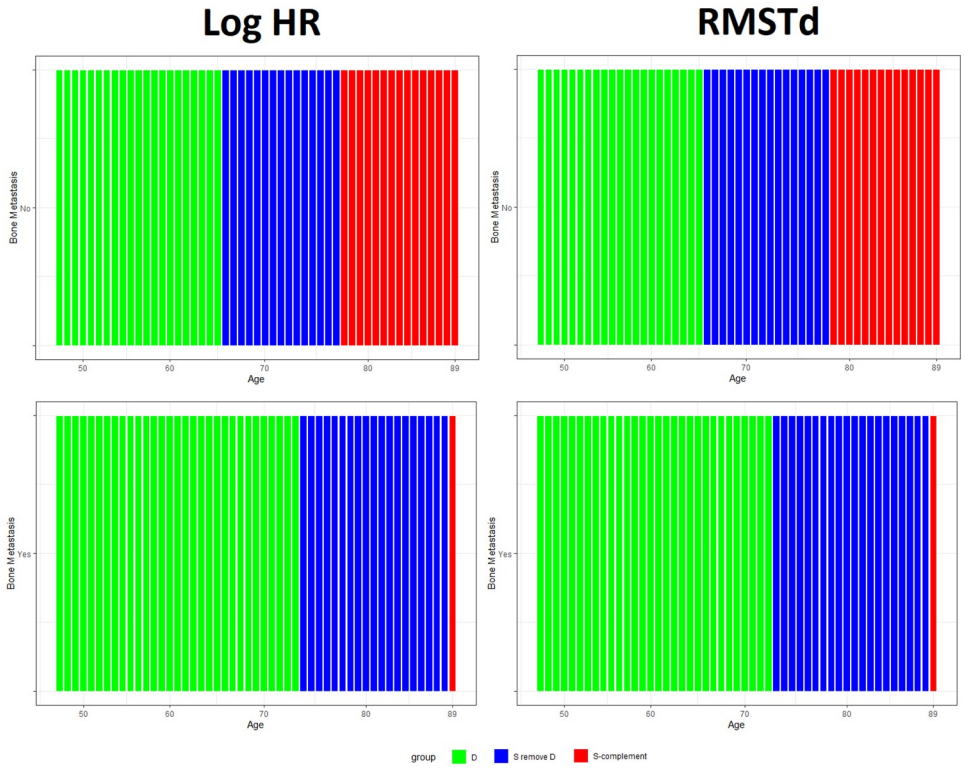

Fig 2. The Bayesian credible subgroups for prostate cancer by using Δ_H (left panel) and Δ_{Rd} (right panel) with credible level 95%.

https://doi.org/10.1371/journal.pone.0229336.g002

the estimated benefiting and uncertainty regions are similar to our regions. However, our method, for both PTE approaches (Δ_H and Δ_{Rd}), identified that patients who were older than 89 years and did not have existence of any bone metastases, may not be benefiting from the treatment. The PITE method did not identify these non–benefiting patients.

5 Analysis of a large simulated clinical trial dataset

We now illustrate our proposed methods on a simulated dateset based on a study published by Scirica et al. [22]. It is the Thrombin Receptor Antagonist in Secondary Prevention of Atherothrombotic Ischemic Events-Thrombolysis in Myocardial Infarction 50 trial. The primary efficacy endpoint is the time of first myocardial infarction, stroke or cardiovascular death. Even though patients with a history of myocardial infarction are treated for secondary prevention, they are still at risk of a recurrent thrombotic events. To reduce recurrent thrombotic events, patients are often treated with platelet inhibitors in addition to aspirin for up to a year, but this treatment also increases bleeding.

Scirica et al. [22] analyzed the vorapaxar dataset using a HR from a Cox proportional hazards model for testing heterogeneous effect across the prespecified subgroups of interest. In contrast to their approach, our aim is to search for benefiting subgroups without prespecifying subgroups of interest. To illustrate our proposed approach, we derived a simulated dataset from the proprietary dataset used by Scirica et al [22] in the following section.

5.1 Simulated dataset based on a large clinical trial dataset

The simulated data mimics the dataset presented at Scirica et al. [22] that pertains to 17,779 patients of whom 8898 were assigned to treatment and 8881 were assigned to placebo. We

Table 4. Summary of baseline characteristics for simulated clinical trial dataset. We report the median with the first and third quartiles for continuous variables, and total count with its percent of the total trial population for categorical variables.

	Treatment (n = 8898)	Placebo (n = 8881)
Age (in years)	59 (52-66)	59 (52-66)
Weight (in kg)	85 (73.5-96)	85 (73-95.5)
Hyperlipidaemia	7568 (85%)	7545 (85%)
Smoking	1729 (19.4%)	1755 (%19.8)
Previous coronary revascularisation	7629 (85.7%)	7645 (86%)

https://doi.org/10.1371/journal.pone.0229336.t004

considered a 5-dimensional prognostic covariate vector which represented patients' characteristics at baseline: age at entry (years), baseline weight (kilograms), history of hyperlipidemia, smoking status and prior coronary revascularization. For each treatment group, we randomly selected 20% of subjects and added a Gaussian noise with zero mean and standard deviation of 1 and 5 for continuous covariates age and baseline weight, respectively. Previous studies [36, 37] found that patients who are younger than 75 years old, with no history of stroke and bodyweight at least 60 kg are likely benefiting from the antiplatelet therapy. Hence we include age, baseline weight and history of prior coronary revascularization as predictive covariates.

The baseline characteristics of 17,779 subjects are summarized in Table 4. The median follow–up was 2.5 years (IQR 2–2.9 years), and the Kaplan–Meier curve [38] of estimated occurrence of the cardiovascular death, myocardial infarction or stroke is showed in Fig 3. The chance of cardiovascular death, myocardial infarction, or stroke was lower in patients in treatment group than those in placebo group over the follow–up time. Moreover, the global test of proportional hazards [39] fails to reject the assumption of proportional hazards (p–value = 0.51).

5.2 Results

We applied Bayesian credible subgroup analysis to the simulated dataset using log HR and RMSTd as described in Section 2. Moreover, we applied cubic B-splines with three degrees of freedom due to their numerical stability [40] for continuous covariates: age at entry and baseline weight, and we chose one knot at medians of these covariates. Fig 4 presents the credible subgroups using Δ_H with credible level at 95% and $\delta_H = 1$, and Fig 5 illustrates the credible subgroups using Δ_{Rd} with the same credible level at 95% and $\delta_R = 0$. For subjects without history of prior coronary revascularization, the results of Δ_H and Δ_{Rd} are similar. Both approaches determine that types of patients younger than 82 years old and with bodyweight at least 80 kg benefit from the treatment versus the control. Moreover, we also have enough evidence to identify non–benefiting subgroup including types of patients who are older than 90 years and have bodyweight less than 78 kg.

For subjects with history of prior coronary revascularization, we found that the two approaches Δ_H and Δ_{Rd} yield similar credible subgroups except that the uncertainty region is slightly larger when using RMSTd. Both approaches find that types of patients aged older than 70 years, with history of prior coronary revascularization, and bodyweight at most 150 kg are not benefiting from a treatment. Note that there is relatively small uncertainty region around age of 22 for log HR approach and not for RMSTd approach. For both dataset, a comparison between the proposed Bayesian credible subgroup with the pointwise method are also reported in S1 File.

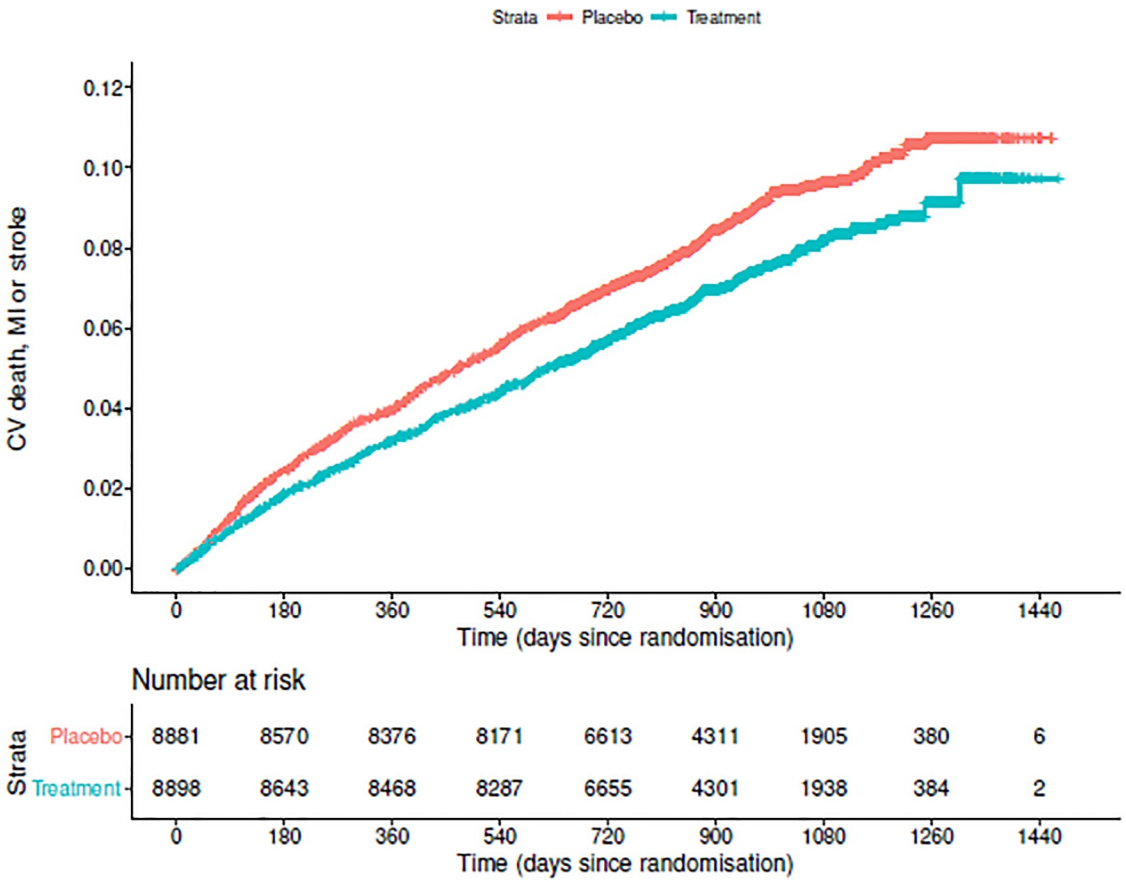

Fig 3. The Kaplan–Meier curve for the time of first myocardial infarction, stroke, or cardiovascular death for the simulated dataset.

https://doi.org/10.1371/journal.pone.0229336.g003

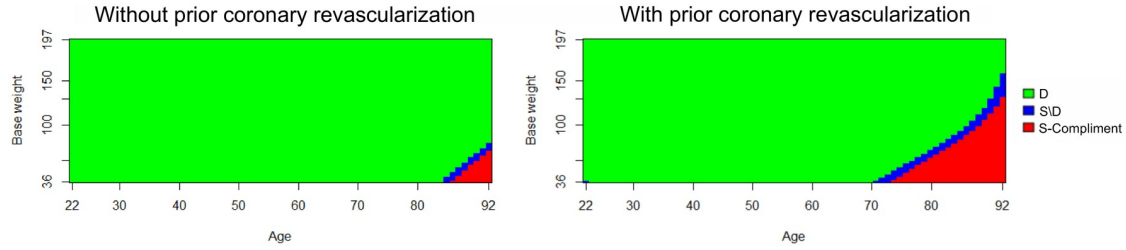

Fig 4. The Bayesian credible subgroups for the simulated dataset by using Δ_H with credible level 95%.

https://doi.org/10.1371/journal.pone.0229336.g004

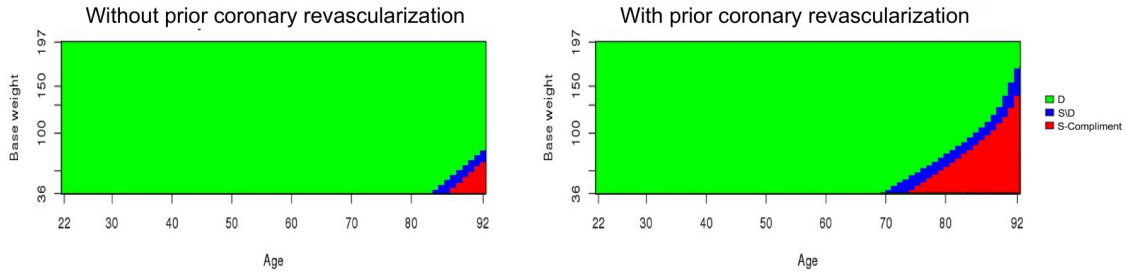

Fig 5. The Bayesian credible subgroups for the simulated dataset by using Δ_{Rd} with credible level 95%.

https://doi.org/10.1371/journal.pone.0229336.g005

6 Discussion

We have presented a Bayesian credible subgroup method for survival endpoints by using two common summaries: log HR and RMSTd. Our proposed methods perform well in simulation studies with respect to frequentist properties for finding credible subgroup pairs *D* and *S*, such as a total coverage, and sensitivity and specificity of *D*. As shown in previous studies [17–20], compared to HR, RMST is a robust and clinically interpretable measure of the survival time distribution without PH assumption. We also demonstrated that an RMSTd approach is appropriate to identify benefiting subgroups for studies with nonproportional hazards. From our applications in the prostate cancer dataset and the simulated large clinical trial data, a Bayesian credible subgroup method, using two common summaries, can identify all member of exclusive subgroup *D* who benefit from a treatment, and non–benefiting subjects who are not members of inclusive subgroup *S*. Moreover, our proposed methods control multiplicity issues in contrast to previous studies.

The major advantage of the Bayesian framework is that it allows us to compute the joint posterior sample of the PTEs and assess the treatment effect across the covariate space. Due to the two–stage approach, our semi–parametric model (Cox proportional hazard regression) in the regression stage can be extended to parametric (accelerated failure time AFT regression) or non–parametric (Dirichlet process priors [41] or Bayesian additive regression trees (BART) [34]). The model choice depends on the flexibility and applicability necessary for the problem as long as the joint posterior sample of the PTEs can be obtained from the model. When using the parametric and semi-parametric model, it would be worth to further investigate the performance of credible subgroups in a case of model misspecification where neither AFT or PH assumption holds.

The first stage of our procedure requires only the list of possible covariates. However, a large number of predictive, especially continuous, covariates makes interpretation of the shape of credible subgroups difficult and reduces power. A possible approach and a topic for future research is to employ variable selection methods in the regression stage such as Bayesian lasso. Another approach is to compute the maximum credible level at which the test of no effect for a given subject's covariate profile is rejected [15].

The methods we proposed in this article focus on a single efficacy endpoint in a clinical trial, but it can be generalized to include more than one endpoint [16]. Multiple efficacy and safety endpoints may be considered simultaneously to establish the actual estimated benefit-risk balance patients may experience depending on their individual characteristics. However, there may be some complexity around the choice of benefit-risk metrics used in combination with the credible subgroup method due to their level of discriminatory abilities [42–44]. Additionally, the endpoints that matter in a decision may also be in different units of measurement, and although the methods of utilities have been proposed as a potential solution, there may be other methodological issues. For example, the uncertainties relating to utilities are more complex to derive. Utilities are also a very specific concept that is context–specific, and may not be intuitive to the general public and decision–makers. Furthermore, there has been some shift in the pharmaceutical industry and regulatory focus on better use of patient preferences data in benefit–risk assessment, as evident by various global initiatives (IMI-PREFER [45, 46], PDUFA VI [47], FDA MDIC [48], PFDD [49]). Patient preferences and perspectives on certain outcomes or treatment options add a unique complexity to the problem because of the heterogeneous nature of patients. It is possible that patients with different characteristics, not only may respond differently to treatments but, may also have different preference values. It is not entirely clear at this time how preferences should be taken into account in relation to patient subgroups. Nevertheless, decisions about a benefit-risk balance of a treatment option

must be made in the context of its benefits and risks, as well as patient preferences; and this presents a wealth of research opportunity to improve decision-making in healthcare.

Finally, our proposed methods in this article only address the scenario when there is no missing value in patients' covariate. Although such case has not been investigated here, regression tree, e.g. BART, can be a potential approach for handling missing data without selection of imputation method [50]. As a closing remark, our Bayesian credible subgroup method for survival endpoint has a broad application in clinical trials as we demonstrated the method in two time–to–event dataset where identifying benefiting subgroups are important in discovering personalized treatment.

Supporting information

S1 File. Supplementary material for Bayesian credible subgroup identication for treatment effectiveness in time–to–event data.
(PDF)

Acknowledgments

The authors would like to acknowledge the contribution of late Joseph Heyse of Merck & Co., Inc., who provided leadership and scientific input to this work. We dedicate this manuscript to his memory. The authors would like to thank also Adam Polis of Merck & Co., Inc. and Jay Horrow of Merck & Co., Inc. for their thoughtful feedback on the manuscript.

Author Contributions

Conceptualization: Richard Baumgartner, Shahrul Mt-Isa, Patrick Schnell.

Data curation: Shahrul Mt-Isa.

Formal analysis: Duy Ngo, Patrick Schnell.

Funding acquisition: Richard Baumgartner, Shahrul Mt-Isa.

Investigation: Richard Baumgartner, Shahrul Mt-Isa.

Methodology: Duy Ngo, Richard Baumgartner, Shahrul Mt-Isa, Dai Feng, Patrick Schnell.

Project administration: Richard Baumgartner, Shahrul Mt-Isa.

Software: Duy Ngo, Dai Feng.

Supervision: Richard Baumgartner, Shahrul Mt-Isa, Patrick Schnell.

Visualization: Duy Ngo.

Writing – original draft: Duy Ngo.

Writing – review & editing: Richard Baumgartner, Shahrul Mt-Isa, Dai Feng, Jie Chen, Patrick Schnell.

References

1. Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. Annals of Internal Medicine. 2000; 133(6):455–463. https://doi.org/10.7326/0003-4819-133-6-200009190-00014 PMID: 10975964
2. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis) uses of baseline data in clinical trials. The Lancet. 2000; 355(9209):1064–1069. https://doi.org/10.1016/S0140-6736(00)02039-0

3. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in Medicine. 2002; 21(19):2917–2930. https://doi.org/10.1002/sim.1296 PMID: 12325108
4. Ruberg SJ, Chen L, Wang Y. The mean does not mean as much anymore: finding sub-groups for tailored therapeutics. Clinical Trials. 2010; 7(5):574–583. https://doi.org/10.1177/1740774510369350 PMID: 20667935
5. Negassa A, Ciampi A, Abrahamowicz M, Shapiro S, Boivin JF. Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. Statistics and Computing. 2005; 15(3):231–239. https://doi.org/10.1007/s11222-005-1311-z
6. Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. The International Journal of Biostatistics. 2008; 4(1). https://doi.org/10.2202/1557-4679.1071 PMID: 20231911
7. Su X, Tsai CL, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. Journal of Machine Learning Research. 2009; 10(Feb):141–158.
8. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Statistics in Medicine. 2011; 30(24):2867–2880. https://doi.org/10.1002/sim.4322 PMID: 21815180
9. Ballarini NM, Rosenkranz GK, Jaki T, König F, Posch M. Subgroup identification in clinical trials via the predicted individual treatment effect. PloS One. 2018; 13(10):e0205971. https://doi.org/10.1371/journal.pone.0205971 PMID: 30335831
10. Berry DA. Subgroup analyses; 1990.
11. Cui L, James Hung H, Wang SJ, Tsong Y. Issues related to subgroup analysis in clinical trials. Journal of Biopharmaceutical Statistics. 2002; 12(3):347–358. https://doi.org/10.1081/bip-120014565 PMID: 12448576
12. Lagakos SW, et al. The challenge of subgroup analyses-reporting without distorting. New England Journal of Medicine. 2006; 354(16):1667. https://doi.org/10.1056/NEJMp068070 PMID: 16625007
13. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine|reporting of subgroup analyses in clinical trials. New England Journal of Medicine. 2007; 357(21):2189–2194. https://doi.org/10.1056/NEJMsr077003 PMID: 18032770
14. Schnell PM, Tang Q, Offen WW, Carlin BP. A Bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. Biometrics. 2016; 72(4):1026–1036. https://doi.org/10.1111/biom.12522 PMID: 27159131
15. Schnell PM, Müller P, Tang Q, Carlin BP. Multiplicity-adjusted semiparametric benefiting subgroup identification in clinical trials. Clinical Trials. 2018; 15(1):75–86. https://doi.org/10.1177/1740774517729167 PMID: 29035083
16. Schnell P, Tang Q, Müller P, Carlin BP, et al. Subgroup inference for multiple treatments and multiple endpoints in an Alzheimer's disease treatment trial. The Annals of Applied Statistics. 2017; 11(2):949–966. https://doi.org/10.1214/17-AOAS1024
17. Zhao L, Claggett B, Tian L, Uno H, Pfeffer MA, Solomon SD, et al. On the restricted mean survival time curve in survival analysis. Biometrics. 2016; 72(1):215–221. https://doi.org/10.1111/biom.12384 PMID: 26302239
18. Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Medical Research Methodology. 2013; 13(1):152. https://doi.org/10.1186/1471-2288-13-152 PMID: 24314264
19. Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. Journal of Clinical Oncology. 2014; 32(22):2380. https://doi.org/10.1200/JCO.2014.55.2208 PMID: 24982461
20. Uno H, Wittes J, Fu H, Solomon SD, Claggett B, Tian L, et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. Annals of Internal Medicine. 2015; 163(2):127–134. https://doi.org/10.7326/M14-1741 PMID: 26054047
21. Royston P, Sauerbrei W. Multivariable Model-building: Advanced prostate cancer dataset; 2008. https://www.imbi.uni-freiburg.de/Royston-Sauerbrei-book.
22. Scirica BM, Bonaca MP, Braunwald E, De Ferrari GM, Isaza D, Lewis BS, et al. Vorapaxar for secondary prevention of thrombotic events for patients with previous myocardial infarction: a prespecified subgroup analysis of the TRA 2 P-TIMI 50 trial. The Lancet. 2012; 380(9850):1317–1324. https://doi.org/10.1016/S0140-6736(12)61269-0
23. Berger JO, Wang X, Shen L. A Bayesian approach to subgroup identification. Journal of Biopharmaceutical Statistics. 2014; 24(1):110–129. https://doi.org/10.1080/10543406.2013.856026 PMID: 24392981
24. Kalbfleisch JD. Non-parametric Bayesian analysis of survival time data. Journal of the Royal Statistical Society: Series B (Methodological). 1978; 40(2):214–221.

25. Hjort NL, et al. Nonparametric Bayes estimators based on beta processes in models for life history data. The Annals of Statistics. 1990; 18(3):1259–1294. https://doi.org/10.1214/aos/1176347749

26. Laud PW, Damien P, Smith AF. Bayesian nonparametric and covariate analysis of failure time data. In: Practical nonparametric and semiparametric Bayesian statistics. Springer; 1998. p. 213–225.

27. Zhou H, Hanson T, Zhang J. spBayesSurv: fitting Bayesian spatial survival models using R. arXiv preprint arXiv:170504584. 2017;.

28. Neal RM, et al. Slice sampling. The annals of statistics. 2003; 31(3):705–767. https://doi.org/10.1214/aos/1056562461

29. Neal RM, et al. MCMC using Hamiltonian dynamics. Handbook of markov chain monte carlo. 2011; 2(11):2.

30. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of statistical software. 2017; 76(1). https://doi.org/10.18637/jss.v076.i01

31. Mahani AS, Sharabiani MT. Multivariate-From-Univariate MCMC Sampler: The R Package MfUSampler. Journal of Statistical Software, Code Snippets. 2017; 78(1), 1–22.

32. Mahani A, Hasan A, Jiang M, Sharabiani M. Stochastic Newton Sampler: The R Package sns. Journal of Statistical Software, Code Snippets. 2016; 74(2):1–33.

33. Mahani A, Sharabiani M. BSGW: Bayesian survival model with lasso shrinkage using generalized weibull regression. R package version 09. 2015; 1.

34. Henderson NC, Louis TA, Rosner GL, Varadhan R. Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. arXiv preprint arXiv:170606611. 2017;.

35. Rosenkranz GK. Exploratory subgroup analysis in clinical trials by model selection. Biometrical Journal. 2016; 58(5):1217–1228. https://doi.org/10.1002/bimj.201500147 PMID: 27230820

36. Wiviott SD, Braunwald E, McCabe CH, Montalescot G, Ruzyllo W, Gottlieb S, et al. Prasugrel versus clopidogrel in patients with acute coronary syndromes. New England Journal of Medicine. 2007; 357 (20):2001–2015. https://doi.org/10.1056/NEJMoa0706482 PMID: 17982182

37. Wiviott SD, Desai N, Murphy SA, Musumeci G, Ragosta M, Antman EM, et al. Efficacy and safety of intensive antiplatelet therapy with prasugrel from TRITON-TIMI 38 in a core clinical cohort defined by worldwide regulatory agencies. The American Journal of Cardiology. 2011; 108(7):905–911. https://doi.org/10.1016/j.amjcard.2011.05.020 PMID: 21816379

38. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American statistical association. 1958; 53(282):457–481. https://doi.org/10.1080/01621459.1958.10501452

39. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994; 81(3):515–526. https://doi.org/10.1093/biomet/81.3.515

40. De Boor C. A practical guide to splines. vol. 27. New York: Springer-verlag; 1978.

41. Müller P, Mitra R. Bayesian nonparametric inference–why and how. Bayesian Analysis. 2013; 8(2).

42. Mt-Isa S, Hallgreen CE, Wang N, Callréus T, Genov G, Hirsch I, et al. Balancing benefit and risk of medicines: a systematic review and classification of available methodologies. Pharmacoepidemiology and Drug Safety. 2014; 23(7):667–678. https://doi.org/10.1002/pds.3636 PMID: 24821575

43. Waddingham E, Mt-Isa S, Nixon R, Ashby D. A Bayesian approach to probabilistic sensitivity analysis in structured benefit-risk assessment. Biometrical Journal. 2016; 58(1):28–42. https://doi.org/10.1002/bimj.201300254 PMID: 25631038

44. Hughes D, Waddingham E, Mt-Isa S, Goginsky A, Chan E, Downey GF, et al. Recommendations for benefit–risk assessment methodologies and visual representations. Pharmacoepidemiology and Drug Safety. 2016; 25(3):251–262. https://doi.org/10.1002/pds.3958 PMID: 26800458

45. de Bekker-Grob EW, Juhaeri J, Kihlbom U, Levitan B. Giving patients' preferences a voice in the medical product lifecycle: why, when and how?: The public-private PREFER project: Work package 2. ISPOR Value & Outcomes Spotlight. 2018;.

46. Soekhai V, Whichello C, Levitan B, Veldwijk J, Pinto CA, Donkers B, et al. Methods for exploring and eliciting patient preferences in the medical product lifecycle: a literature review. Drug Discovery Today. 2019;.

47. Dabrowska A, Thaul S. Prescription Drug User Fee Act (PDUFA): 2012 Reauthorization as PDUFA V; 2018. Available from: https://fas.org/sgp/crs/misc/R44864.pdf.

48. MDIC MDA. Patient Centered Benefit-Risk; 2015. Available from: https://mdic.org/project/patient-centered-benefit-risk-pcbr/.

49. FDA. Plan for issuance of patient focused drug development guidance; 2017. Available from: https://www.fda.gov/media/105979/download.

50. Kapelner A, Bleich J. Prediction with missing data via Bayesian additive regression trees. Canadian Journal of Statistics. 2015; 43(2):224–239. https://doi.org/10.1002/cjs.11248