Does Bayesian approach perform better than Frequentist approach for modeling the survival of rare disease?

Anahita Saeedi, Jinghan Cui

Rare disease and orphan drugs

Rare Disease is a condition in which the prevalence is not more than 50 per 100,000 people.

Orphan drug is a drug used to treat, prevent, or diagnose an orphan disease/rare disease.

Collectively, RDs affect as many as 2,500,000 Americans, according to NIH.

The nature of RDs raises challenges because of

- Small participant numbers and small number of endpoint events.
- Limited interpretation of the treatment effect due to the lack of precision
- Traditional statistical methods yield overly conservative results

Orphanet (www.orpha.net) is a 37-country network, cofunded by the European Commission that aims to increase knowledge on RDs so as to improve the diagnosis, care, and treatment of people with RDs

Frequentist approach: survival analysis

- One sample: Kaplan Meier Estimator, log-rank test (nonparametric)
- Two or more groups: Proportional hazard model, Accelerated failure time. Depending on which parametric family best describe the survival time, we can choose from:
    - M-spline model
    - Exponential model
    - Weibull model
    - Gompertz model
    - Log-normal
    - Log-Logistic
- For our study, we will start with exponential model as it is the most commonly used one

Notation

t_i denotes the observed or censoring time.

$z_1, ..., z_p$ denotes the p covariates

β denotes the regression coefficients

$R(t) = \{i : T_i \geq t\}$: the set of individuals who are "at risk" at time t, **risk set**

Survival models

Proportional hazards model: $\lambda(t \mid \boldsymbol{z}) = \lambda_0(t) e^{\boldsymbol{\beta}^\top \boldsymbol{z}} = \lambda_0(t) \exp(\beta_1 z_1 + \cdots + \beta_p z_p)$

Partial likelihood: $L_j(\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{z}_j)}{\sum_{l \in R(t_j)} \exp(\boldsymbol{\beta}^\top \boldsymbol{z}_l)}$

Accelerated Failure time: $S(t \mid \boldsymbol{z}) = S_0(t \exp(\boldsymbol{\beta}^\top \boldsymbol{z}))$

Which is equivalent to: $\lambda(t \mid \boldsymbol{z}) = \exp(\boldsymbol{\beta}^\top \boldsymbol{z}) \lambda_0(t \exp(\boldsymbol{\beta}^\top \boldsymbol{z}))$

Bayesian approach

- Informative priors attempt to incorporate knowledge from other sources such as past studies in order to realistically capture one's state of knowledge about β
- Choices of prior distribution for β_0 include the normal, t, or Cauchy distributions.
- Choices of prior distribution for the regression coefficients include normal, t, gamma and Cauchy distributions as well as several shrinkage prior distributions.

Bayesian survival estimation

S_d and h_d denote the survival and hazard function of a parametric family d

M_d denotes the survival model

$\theta_d = \{\alpha_d, \beta_d\}$

Likelihood: $p\left(\text{data}|\theta_d, \mathcal{M}_d\right) = \prod h_d\left(t_i|x_i, \theta_d\right)^{I(c_i=1)} \times S_d\left(t_i|x_i, \theta_d\right)$.

According to Bayes theorem, we can obtain the posterior parameter distributions:

$$p\left(\theta_d|\text{data}, \mathcal{M}_d\right) = \frac{p\left(\text{data}|\theta_d, \mathcal{M}_d\right) \times p\left(\theta_d|\mathcal{M}_d\right)}{p\left(\text{data}|\mathcal{M}_d\right)}$$

Bayesian hypothesis testing

$$p\left(\mathcal{M}_{0,d}|\text{data}, d\right) = \frac{p\left(\text{data}|\mathcal{M}_{0,d}, d\right) \times p\left(\mathcal{M}_{0,d}|d\right)}{p(\text{data}|d)},$$

$$p\left(\mathcal{M}_{1,d}|\text{data}, d\right) = \frac{p\left(\text{data}|\mathcal{M}_{1,d}, d\right) \times p\left(\mathcal{M}_{1,d}|d\right)}{p(\text{data}|d)},$$

Bayesian factor: $$\text{BF}_{10} = \frac{p\left(\text{data}|\mathcal{M}_{1,d}, d\right)}{p\left(\text{data}|\mathcal{M}_{0,d}, d\right)},$$

Simulation settings

The true model is given by $Y = \alpha_0 + \alpha_1 X + \boldsymbol{\alpha}_2^T \mathbf{Z} + \epsilon$, with a_0=0.5, a_1=0.5, a_2=-0.5.

Y denotes the true survival time, X denotes the treatment/control, and Z denotes the covariates.

We will consider different sample sizes from 50 to 500 and generate Y ~ Exp(3), X~Bernoulli(0.5), Z ~ Unif(1,6), e ~ Normal(0,0.75^2). The censoring time C will be generated from Exp(lambda), where we will test lambda from 0 to 5 to calculate the corresponding censoring rates.

Empirical bias, standard deviation and standard errors will be calculated for each different simulation.

Next step

1. Applying bayesian and frequentist method to simulated datasets
    a. Rare events and small sample size usually happen at the same time, test the performance of both methods in scenarios when both issues appear or when either of the issues appear
    b. Try survival time from parametric distributions other than exponential
2. Applying the bayesian vs the frequentist method on the rare disease dataset and compare the results as well as identifying the advantages and disadvantages in a real world setting
3. Are there other approaches to improve the estimation of survival for rare events other than the two we explored?

Reference

- https://www.sciencedirect.com/science/article/pii/S0957417409002917
- https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01676-9
- https://arxiv.org/pdf/2002.09633.pdf
- https://ojrd.biomedcentral.com/articles/10.1186/s13023-022-02342-5#Sec2
- https://zenodo.org/record/3866164