Biostat 537: Survival Analaysis

TA Session 7

Ethan Ashby

February 27, 2024

Review from Last Time

1. Residual-based diagnostics can help diagnose covariate functional form misspecification, assess goodness-of-fit, detect influential observations, and check the proportional hazards assumption in a Cox model.
2. Time-dependent covariates can be accommodated in the Cox model by creating “pseudo-observations” using the start-stop format.
3. Accelerated failure time (AFT) models impose a common parametric shape on the distribution of survival times and allow covariates to speed up/slow down the event process.

Presentation Overview

1. Overview of RCTs & Power Analysis
2. Sample Size & Power Calculations w/ Survival Data

RCTs: a gold standard of evidence-based medicine

Randomized, placebo-controlled trials can yield causal conclusions about the effect of treatment on an outcome.

Dangers to the validity of an RCT

1. Poor choice of outcome: outcome doesn't capture how a patient "feels, functions, or survives".
2. Lack of prespecified analysis plan: multiple outcomes, comparisons, "looks" at the data can lead to invalid results.
3. Trial is inadequately designed (too small, too much missing data, etc.) to answer the scientific question of interest.

RCTs: a gold standard of evidence-based medicine

1. Poor choice of outcome $\Longrightarrow$ low clinical relevance.
   1. Solution: consult clinicians, stakeholders, domain experts, and patients re: what constitutes a clinically relevant effect.
2. Lack of prespecified analysis plan $\Longrightarrow$ false positives.
   1. Solution: rigorously develop analysis plan to answer main question(s) before seeing the data. Build in contingencies for anticipated issues.
3. Trial is inadequately designed to answer question of interest $\Longrightarrow$ false negatives.
   1. Solution: conduct power analysis to determine trial size needed to reliably answer scientific question.

Type I Errors

Clinical trials are often designed around rejecting/failing to reject a single null hypothesis of treatment futility.

Statistical tests are designed to control the Type I error rate, or the probability of *incorrectly rejecting* a true null.

For example, consider the null where an experimental drug A has no effect on patient survival. $H_0 : S_{A=1}(t) = S_{A=0}(t)$.

An $\alpha = 0.05$-level logrank test controls the probability of mistakenly claiming an ineffective drug is effective at 5%.

Complications to controlling Type I Error

All reasonable statistical tests are designed to control Type I error rate. Additional thought needs to be given when...

1. There are multiple outcomes of interest.
2. There are multiple treatment groups to be compared.
3. There are multiple times during the study where the hypothesis H_0 will be tested.

Type II Error

Recall statistical tests control Type I error rates, or the probability of rejecting the null when it is true.

We also want to minimize the probability of Type II error, or the probability of *failing to reject* the null when it is false. *Power* is (1 - Type II error rate).

E.g., consider the null where an experimental drug A has no effect on patient survival. $H_0 : S_{A=1}(t) = S_{A=0}(t)$.

A Type II error occurs when a logrank test fails to reject H_0 when the drug does affect survival.

The Importance of Well-Powered Research

	H_0 True	H_0 False
Fail to Reject	✓	T2E
Reject	T1E	✓

Well-powered research increases the probability reaching the correct conclusions ✓.

E.g., consider the case of finding an effective HIV vaccine where H_0 is true for most candidates. Even with valid statistical tests, the *False Discovery Rate* = $\frac{\text{T1E}}{\text{T1E}+\checkmark}$ can be small over many trials.

High power increases the probability that rejecting H_0 is a true ✓ and not a T1E.

Sample Size Calculations: One Arm Study

Suppose survival times $\sim$ Exp(λ). Suppose we wish to test $H_0 : \lambda = \lambda_0$ versus $H_A : \lambda = \lambda_A$.

Goal: determine how many patients (events) we need to detect a certain hazard ratio λ_0/λ_A with a specified power $(1 - \beta)$ at significance level α.

Sample Size Calculations: One Arm Study

Under the exponential model, the log-likelihood is

$$\ell = \sum_{i=1}^{n} \delta_i \log(\lambda) - \lambda \sum_{i=1}^{n} t_i \equiv -d\theta - Ve^{-\theta}$$

The MLE is $\hat{\theta} = \log(V/d)$ with $\text{Var}(\hat{\theta}) \approx 1/d$.
To control the significance level at α under H_0, we find k such that

$$\alpha = P\left(Z > \frac{k - \theta_0}{1/\sqrt{d}}\right) \implies k = \theta_0 + \frac{z_\alpha}{\sqrt{d}}$$

Sample Size Calculations: One Arm Study

To ensure the power of the test is at level $1-\beta$ under H_A

$$1-\beta = P(\hat{\theta} > k|\theta = \theta_A) = P\left(Z > \frac{k-\theta_A}{1/\sqrt{d}}\right)$$

$$\implies z_{1-\beta} = \sqrt{d}(k-\theta_A)$$

$$\implies z_{1-\beta} = \sqrt{d}\left(\theta_0 + \frac{z_\alpha}{\sqrt{d}} - \theta_A\right)$$

Now we can solve for d, the required number of events d required to detect hazard ratio Δ at significance level α with power $1-\beta$.

$$d = \frac{(z_\beta + z_\alpha)^2}{(\theta_A - \theta_0)^2}$$

Sample Size Calculations: One Arm Study

How do we go from number of events to total number of participants?

Need to estimate the proportion of patients who will experience the event.

Suppose patients are recruited Unif$(0, a)$ and followed until time $a + f$. Then the event probability is

$$\pi = 1 - \frac{1}{a}\int_{a}^{a+f} S(u,\lambda)du = 1 - \frac{1}{a\lambda}\{e^{-\lambda f} - e^{-\lambda(a+f)}\}$$

d/π yields the estimated number of participants enrolled.

Sample Size Calculations: Two Arms

In most practical cases, we wish to compare an experimental arm to a control arm.

$H_0 : S_0(t) \geq S_1(t)$ versus $H_A : S_0(t) < S_1(t)$.

Assume T are exponentially distributed. Let $0 < p < 1$ be the proportion randomized to treatment versus control.

Let $\log(\hat{\lambda}_0) - \log(\hat{\lambda}_1)$ be the estimate of the log hazard ratio where $\hat{\lambda}_a = \frac{d_a}{V_a} = \frac{\sum_{i=1}^{n} d_i I(A=a)}{\sum_{i=1}^{n} t_i I(A=a)}$.

Sample Size Calculations: Two Arms

Let $\log(\hat{\lambda}_0) - \log(\hat{\lambda}_1)$ be the estimate of the log hazard ratio.

$$\text{Var}(\log(\hat{\lambda}_0) - \log(\hat{\lambda}_1)) = \frac{1}{n_0\pi_0} + \frac{1}{n_1\pi_1} = \frac{1}{np(1-p)}\left(\frac{p}{\pi_1} + \frac{1-p}{\pi_0}\right)^{-1}$$

To control the test at level α, we must find k such that

$$\alpha = P(\log(\hat{\lambda}_0) - \log(\hat{\lambda}_1) > k|H_0) = P(Z \geq k/\sigma)$$
$$\implies k = z_\alpha \sigma$$

To ensure Power is $1-\beta$, we have

$$1-\beta = P(\log(\hat{\lambda}_0) - \log(\hat{\lambda}_1) > k|H_A) = P(Z \geq k/\sigma)$$
$$\implies z_{1-\beta} = \frac{k-\delta_A}{\sigma} \implies z_\beta = \frac{\delta_A}{\sigma} - z_\alpha$$

Sample Size Calculations: Two Arms

Finally we have

$$
\begin{aligned}
\frac{\delta_A^2}{(z_\beta + z_\alpha)^2} = \sigma^2 &= \frac{1}{np(1-p)} \left(\frac{p}{\pi_1} + \frac{1-p}{\pi_0} \right)^{-1} \\
\implies n &= \frac{(z_\beta + z_\alpha)^2}{\delta_A^2 p(1-p)} \left(\frac{p}{\pi_1} + \frac{1-p}{\pi_0} \right)^{-1} \\
\implies d &= \frac{(z_\beta + z_\alpha)^2}{\delta_A^2 p(1-p)}
\end{aligned}
$$

Sample Size Calculations: Simulation

Analytical sample size calculations are often based on simplistic assumptions which may not hold in practice.

An alternative is to *simulate* plausible datasets based on scientific knowledge, apply your statistical test over repeated simulations, and evaluate the power empirically.

Can be useful to characterize sample size estimates under varying parameters.

Sample Size Calculations: Simulation

Question: how large of a trial would we need to run to have 90% power to detect if a bNAb is effective in preventing mother to child transmission of HIV via breastmilk in South Africa?

- $H_0 : \frac{\lambda_{\text{bNAb}}}{\lambda_{\text{Placebo}}} \geq 1$, $H_A : \frac{\lambda_{\text{bNAb}}}{\lambda_{\text{Placebo}}} = 0.2$.
- $T \sim \text{Exp}\left(\lambda_{\text{Placebo}}(t) * \left(\frac{\lambda_{\text{bNAb}}}{\lambda_{\text{Placebo}}}\right)^A\right)$ where $\lambda_{\text{Placebo}}(t) = 0.015$ from $t \in \{0, 6\}$ months, $\lambda_{\text{Placebo}}(t) = 0.0075$ thereafter.
- $C_{\text{LFU}} \sim \text{Exp}(\lambda = 0.10)$ (10% annual LFU).
- $C_{\text{Weaning}} \sim N(16, 3.5)$ using prior data.
- Analysis method: Cox model score test of H_0 at level $1 - \alpha$.

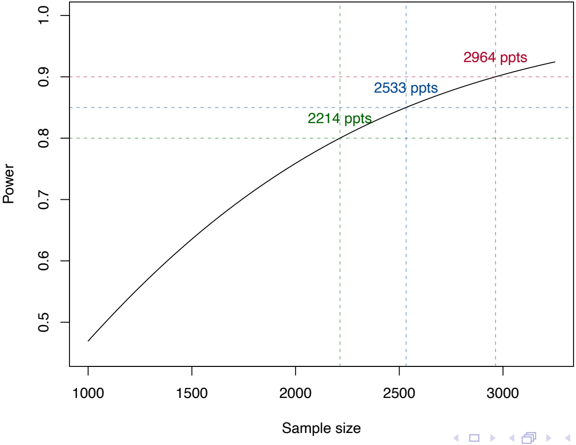
Sample Size Calculations: Simulation

Construct a grid of sample sizes for n. For each value in the grid, repeat the following 1000 times.

1. Generate data $(A, \tilde{T} := \min(T, C_{\text{LFU}}, C_{\text{Weaning}}), \Delta_T)$ under H_A where $\frac{\lambda_{\text{bNAb}}}{\lambda_{\text{Placebo}}} = 0.2$ under sample sample size n.
2. Run test: Cox model score test, of $H_0 : \frac{\lambda_{\text{bNAb}}}{\lambda_{\text{Placebo}}} \geq 1$ at level $\alpha = 0.025$. Record whether H_0 is rejected.

$\widehat{\text{Power}}(n)$ is the proportion of replicates where we rejected H_0 at sample size n. Determine the minimum n in your grid such that $\widehat{\text{Power}}(n) > 1 - \beta$.

Sample Size Calculations: Simulation



Review

1. When designing a study, careful attention should be paid to selecting a meaningful outcome, controlling the probability of Type I error, adequate power to detect plausible effects.
2. There exist analytical approaches to calculate the number of events and number of participants to enroll to ensure adequate power under assumptions.
3. Simulation is a flexible and powerful tool for sample size calculations and power analysis.