Study design and analysis for self-reported time-to-event outcomes with the `icensmis` R package

Kate Hoff Shutta

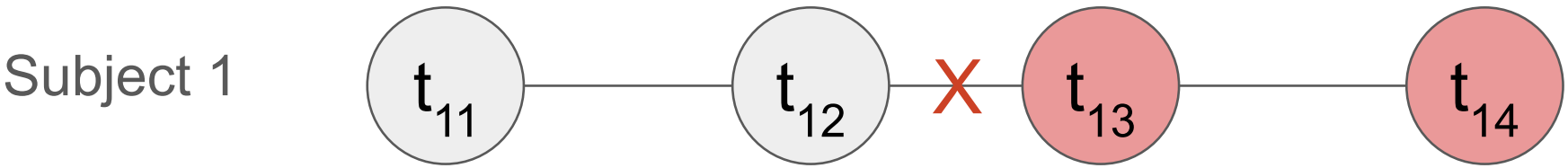
Ryan Sheehan, Yibai Zhao, Yukun Li, Minsu Kim,
Xiangdong Gu, Raji Balasubramanian

ASA Conference on Statistical Practice
New Orleans, LA
February 29, 2024

Outline

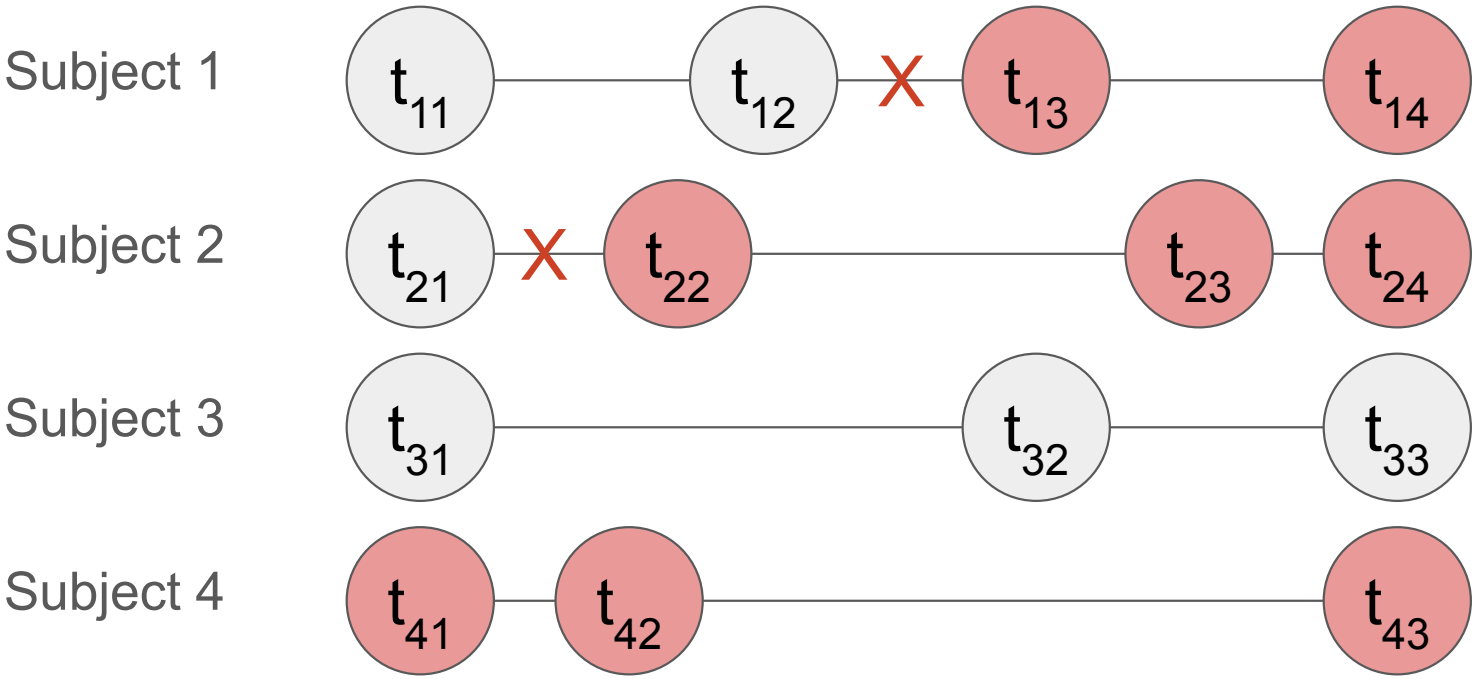
Table of Contents

Interval-censored time-to-event data



- A subject participates in a longitudinal study with multiple follow-up visits at times $t_{11}, \ldots, t_{14}$
- The event of interest is experienced between the 2nd and 3rd visit
- The exact time-to-event is unknown, but we know the interval in which it happened
- Once the event has occurred, the person is considered "positive" for the remainder of the study

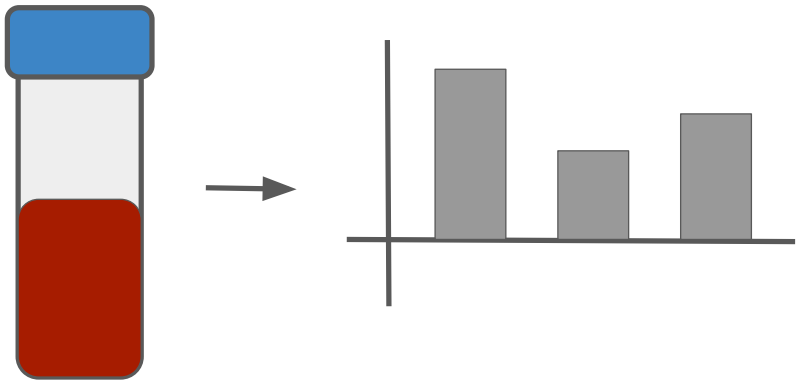
Data may include left- and right-censored cases



- Subject 3 is **right-censored**; they never experience the event
- Subject 4 is **left-censored**; they experienced the event before entering the study
- By setting $t_{i0} = 0$ and $t_{i5} = \infty$ for all subjects, these can be treated similarly to interval-censored data

Using self-reports to study time-to-event data

- In large cohort studies, gold-standard diagnostic testing can be expensive and infeasible
- Alternative: self-reported measurement of an event
- Example: diabetes in the NHANES Epidemiological Follow-up Study (NHEFS)
  - Gold-standard tests are based on blood draws
  - A questionnaire-based measure serves as a proxy[a]



B38 SUGRDIAB 721-722 DID DOCTOR EVER SAY SUBJECT HAD DIABETES?

729	-1	INAPPLICABLE
0	-7	REFUSED
13	-8	DON'T KNOW
4	-9	NOT ASCERTAINED
247	1	YES
8931	2	NO
74	95	BORDERLINE

INAPPLICABLE. (COL 90-91=1) OR (COL 90-91=-1,2 AND COL 547-548=2) OR (COL 90-91=-1,2 AND COL 717-718=2)

Q.B38 WAS ASKED OF THE RESPONDENT IF NO DIABETES WAS REPORTED IN A PREVIOUS INTERVIEW OR THE SUBJECT HAD NOT BEEN PREVIOUSLY INTERVIEWED. IF THE RESPONDENT VOLUNTEERED BORDERLINE DIABETES, COL 721-722 IS CODED 95.

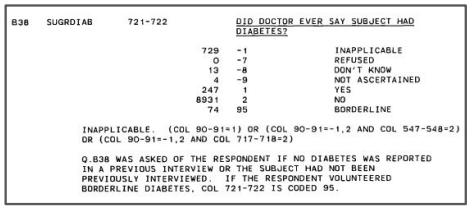
[a] Image: 1997 NHEFS interview, https://wwwn.cdc.gov/nchs/nhanes/nhefs/

Limitations of self reports

- Self reports can have varying accuracy and precision
  - Stigma may prevent a person from reporting an event that happened
  - Misunderstanding of a conversation with a physician could lead to a false positive report
  - A person may not remember a diagnosis they received
- In contrast, gold standard diagnostics are typically highly accurate and precise
  - Well-established technology
  - FDA diagnostic approval process
  - Clinicians are experienced in their use
- Common methods for survival analysis assume error-free reporting; `icensmis` addresses the error directly

Quantifying error in self-reports

Self-Report by Questionnaire



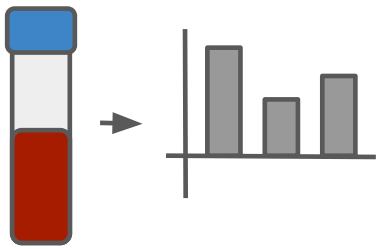
B38 SUGRDIAB 721-722 DID DOCTOR EVER SAY SUBJECT HAD DIABETES?

729	-1	INAPPLICABLE
0	-7	REFUSED
13	-8	DON'T KNOW
4	-9	NOT ASCERTAINED
247	1	YES
8931	2	NO
74	95	BORDERLINE

INAPPLICABLE, (COL 90-91=1) OR (COL 90-91=-1,2 AND COL 547-548=2) OR (COL 90-91=-1,2 AND COL 717-718=2)

Q.B38 WAS ASKED OF THE RESPONDENT IF NO DIABETES WAS REPORTED IN A PREVIOUS INTERVIEW OR THE SUBJECT HAD NOT BEEN PREVIOUSLY INTERVIEWED. IF THE RESPONDENT VOLUNTEERED BORDERLINE DIABETES, COL 721-722 IS CODED 95.

Gold Standard Diagnosis



	Self-report that event has occurred	Self-report that event has not occurred
Event has occurred	True Positive (TP)	False Negative (FN)
Event has not occurred	False Positive (FP)	True Negative (TN)

Sensitivity and specificity are key characteristics of self-reports

- Sensitivity = TP/(TP + FN) = P(Subject self-reports they experienced the event | the event has occurred)
- Specificity = TN/(TN + FP) = P(Subject does not self-report the event | the event has not occurred)
- These two summary measures can be used to incorporate error into time-to-event models

	Self-report that event has occurred	Self-report that event has not occurred
Event has occurred	True Positive (TP)	False Negative (FN)
Event has not occurred	False Positive (FP)	True Negative (TN)

Table of Contents

Model notation

For each of N subjects, indexed by i, we use the following notation:

- n_i: the number of visits (tests) for person i
- $\mathbf{t_i} = (t_{i1}, \ldots, t_{in_i})$: the set of distinct, ordered visit times
- $\mathbf{R_i} = (R_{i1}, \ldots, R_{in_i})$: the self-reported outcome at each visit
  - $R_{ik} = 0$ if the subject self-reports they have not experienced the event at visit t_{ik}
  - $R_{ik} = 1$ if they report that they have
- X_i: the time at which person i experiences the event

Assume there are J total distinct visit times across the N participants. We represent these ordered times as:

- $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_{J+1})$
- To handle left- and right-censored cases, set $\tau_0 = 0$ and $\tau_{j+1} = \infty$
- Note $\boldsymbol{\tau}$ has the following property:

$$0 = \tau_0 < \tau_1 < \cdots < \tau_{J+1} = \infty \tag{1}$$

Self-report values depend on sensitivity and specificity

Suppose subject i experiences the event at time X_i. What will they report at each of their visit times $t_{i1}, \ldots, t_{in_i}$?

- Case 1, perfect self-reports:
  - R_{ik} is deterministic and depends only on whether the event has happened yet or not

$$R_{ik} = \begin{cases} 0 & t_{ik} \leq X_i \\ 1 & t_{ik} > X_i \end{cases} \tag{2}$$

- Case 2, error-prone self-reports:
  - R_{ik} is now a random variable depending on sensitivity φ_1 and specificity φ_0

$$Pr(R_{ik} = 1) = \begin{cases} 1 - \varphi_0 & t_{ik} \leq X_i \\ \varphi_1 & t_{ik} > X_i \end{cases} \tag{3}$$

- Note that if $\varphi_1 = 1$ and $\varphi_0 = 1$ (perfect sensitivity and specificity), Case 2 reduces to Case 1

Incorporating sensitivity and specificity into the likelihood

For each person i, the probability of observing their sequence of n_i self-reports $\mathbf{R}_i$ at times $\mathbf{t}_i$ can be expressed as:

$$g(\mathbf{R}_i, \mathbf{t}_i, n_i) = \sum_{j=1}^{J+1} Pr(\tau_{j-1} < X_i \leq \tau_j) \prod_{k=1}^{n_i} Pr(R_{ik} | \tau_{j-1} < X_i \leq \tau_j, t_{ik}) \quad (4)$$

- θ_j: Probability that the event occurs between times τ_{j-1} and τ_j
- C_{ij}: Conditional probability of the person's data given the event occurs between τ_{j-1} and τ_j [1]
- We are interested in estimating θ_j; C_{ij} is a function of the sensitivity φ_1 and specificity φ_0

[1] Under the assumption that $\mathbf{R}_i$ are independent given X_i

A Cox PH framework enables incorporation of covariates

- Consider a covariate vector $\mathbf{z}_i$ for each individual
- The Cox PH model assumes that the hazard function has the form $\lambda(t|\mathbf{z}_i) = \lambda_0(t)\exp(\mathbf{z}_i\beta)$
- An equivalent formulation in terms of the survival function is $S(t|\mathbf{z}_i) = S_0(t)\exp(\exp(\mathbf{z}_i\beta))$
- Our modified Cox PH likelihood takes the form

$$\ell(\theta,\beta) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{J+1} D_{ij}(1-\sum_{k=1}^{j}\theta_k)\exp\{\exp(\mathbf{z}_i\beta)\}\right). \quad (5)$$

- $\left(1-\sum_{k=1}^{j}\theta_k\right)$ is the baseline survival function at time τ_j
- D_{ij} is a function of C_{ij}, which is a function of test sensitivity φ_1 and test specificity φ_0

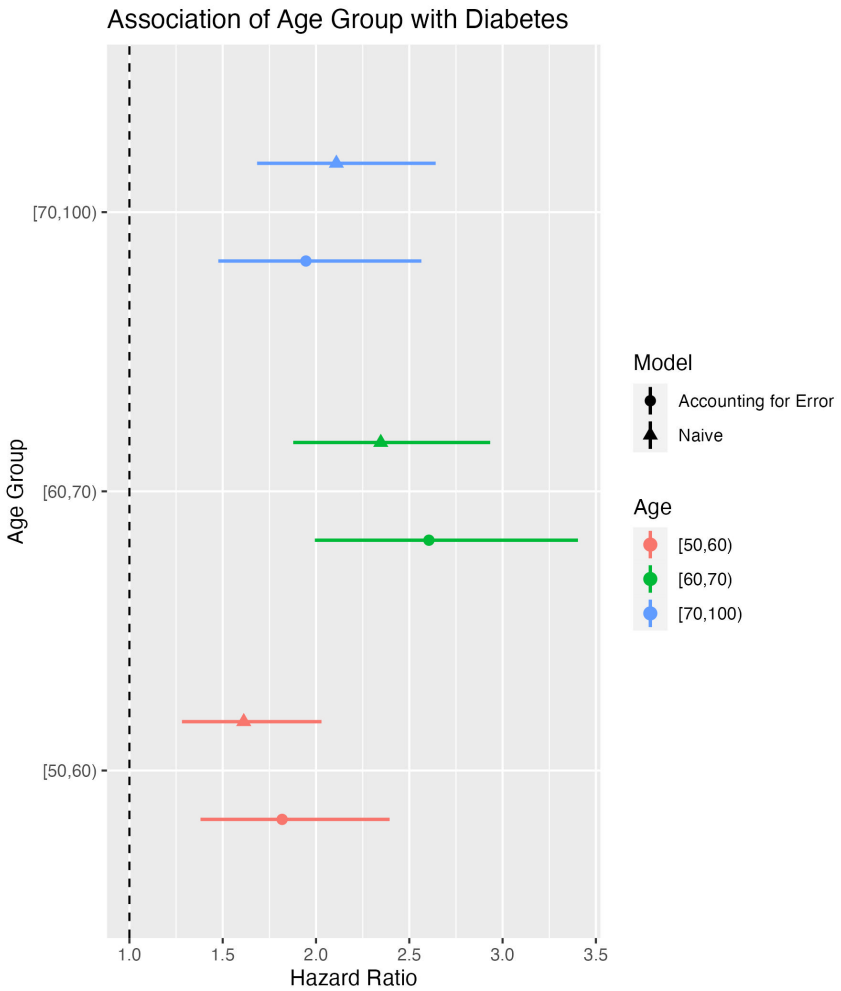
Example: Diabetes self-reports in the NHANES NHEFS

- NHANES: National Health And Nutrition Examination Study, an ongoing study of the US Centers for Disease Control and Prevention
- Conducted in some version since 1959[a]
- NHEFS: NHANES I Epidemiologic Follow-up Study
  - Initial exam in NHANES I (1971-1974)
  - Four follow-up interviews between 1982 - 1992
- At each interview, subjects were asked if a doctor had ever told them they had diabetes

[a] https://wwwn.cdc.gov/nchs/nhanes/

Is there an association between age at study entry and time to diabetes onset?

- $N = 9385$ subjects with diabetes status self-reported at least once in the follow-up interview period
- 698 of these subjects reported onset of diabetes during the study (10-year incidence $\approx 7.4\%$
- Assumed sensitivity $\varphi_1 = 0.55$ and specificity $\varphi_0 = 0.99$ based on previous studies in the Women's Health Initiative [1]



Estimated time to diabetes onset

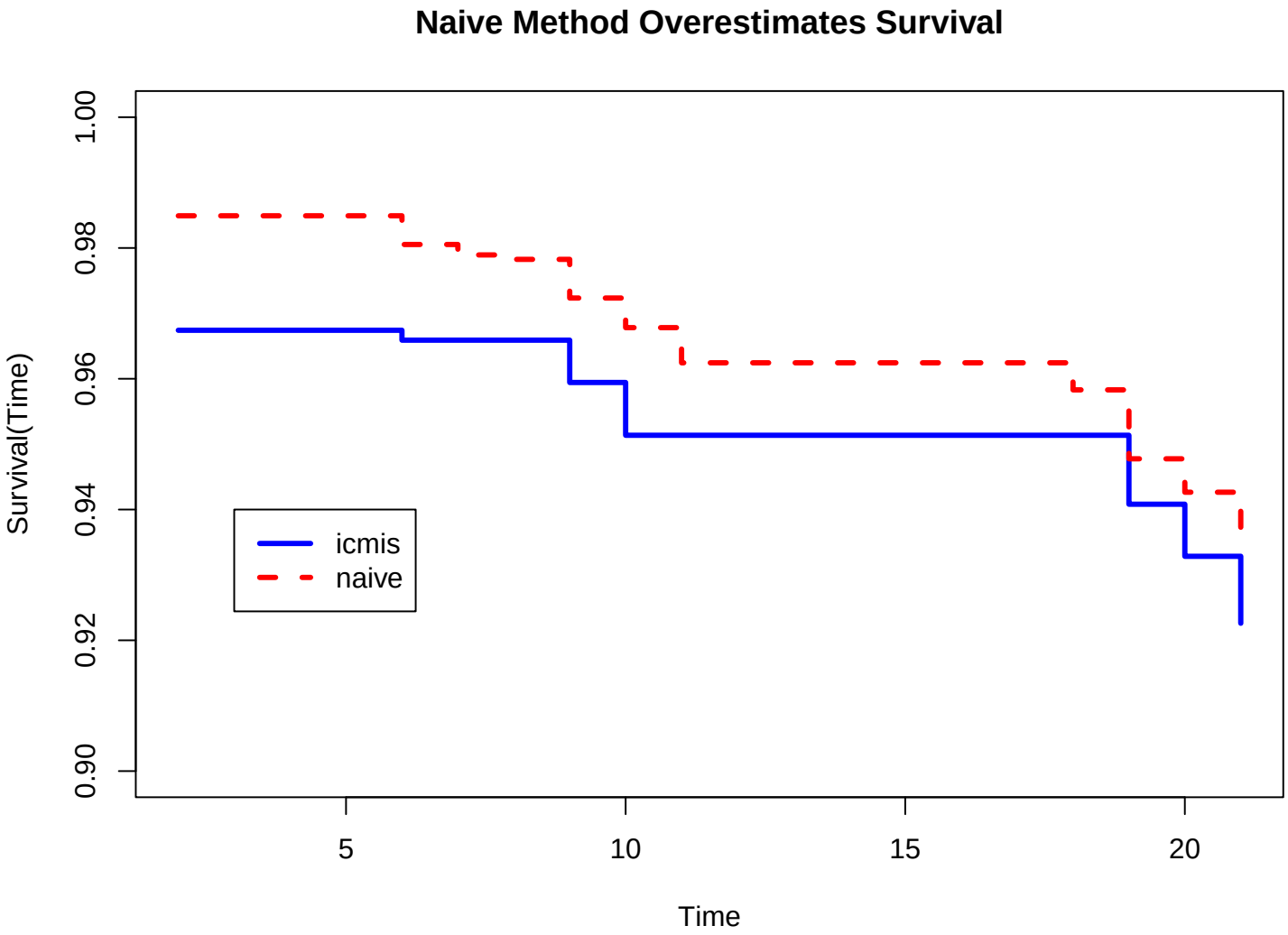
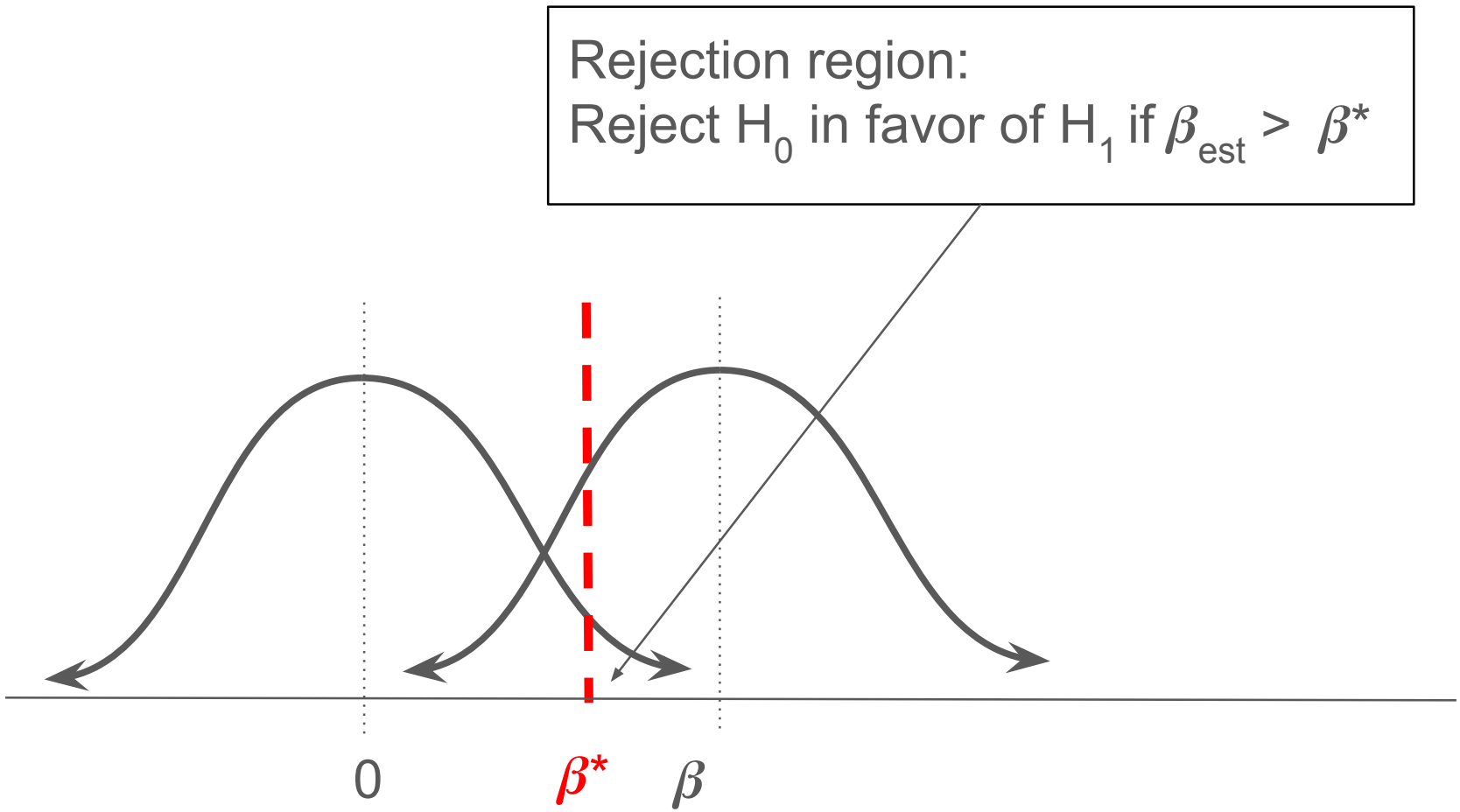


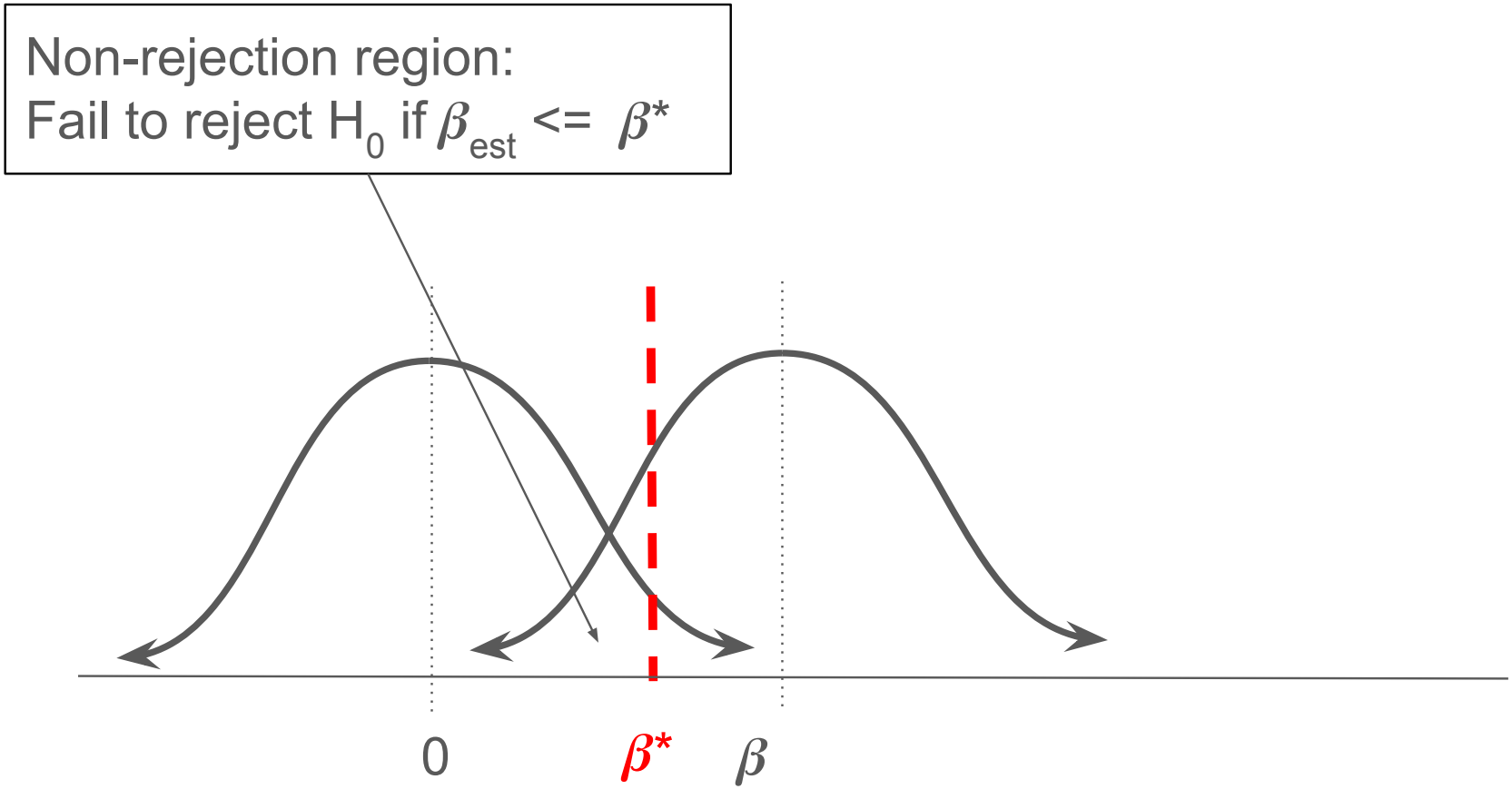
Table of Contents

Review of power



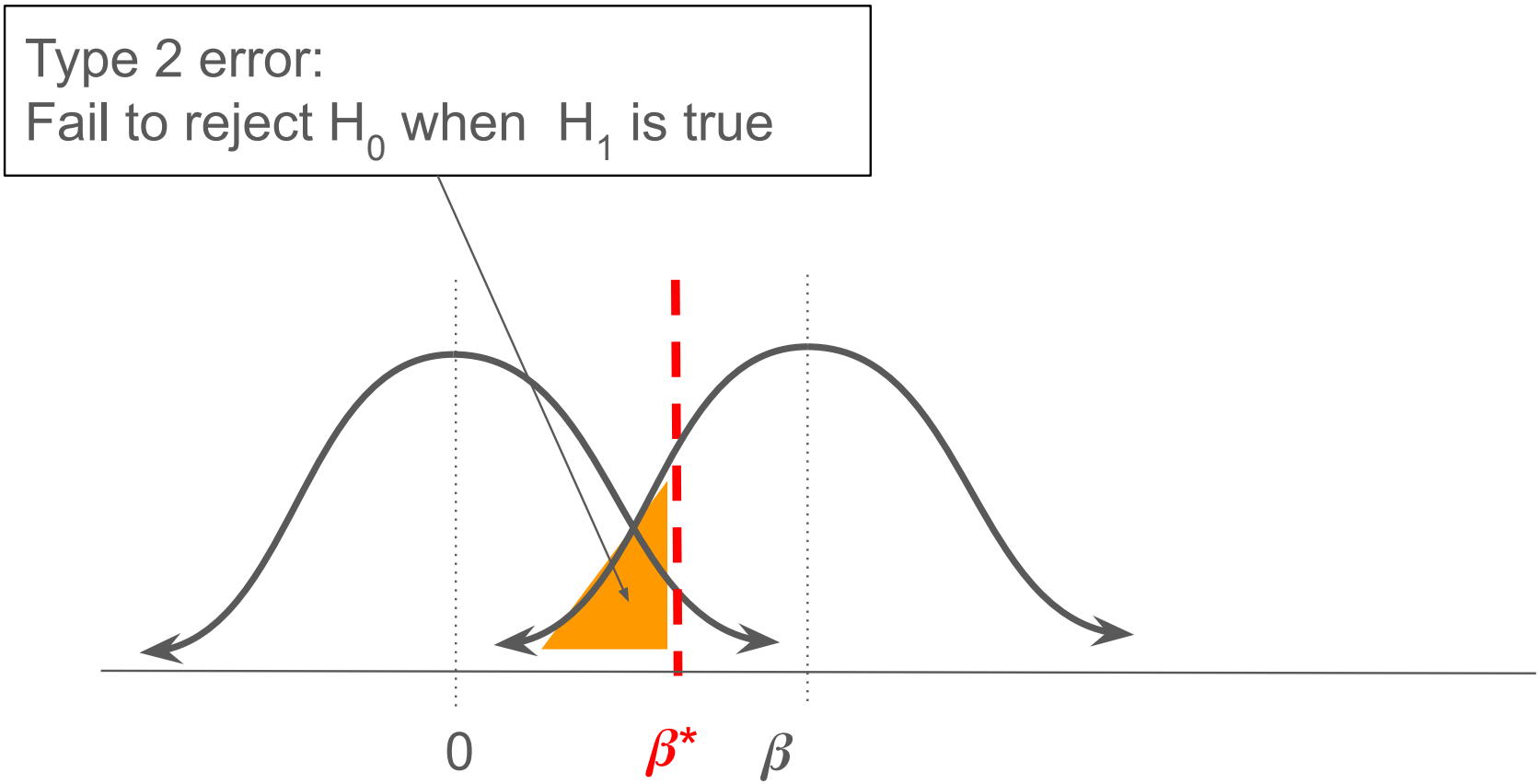
H_0: log hazard ratio = 0 vs. H_1: log hazard ratio = β

Review of power



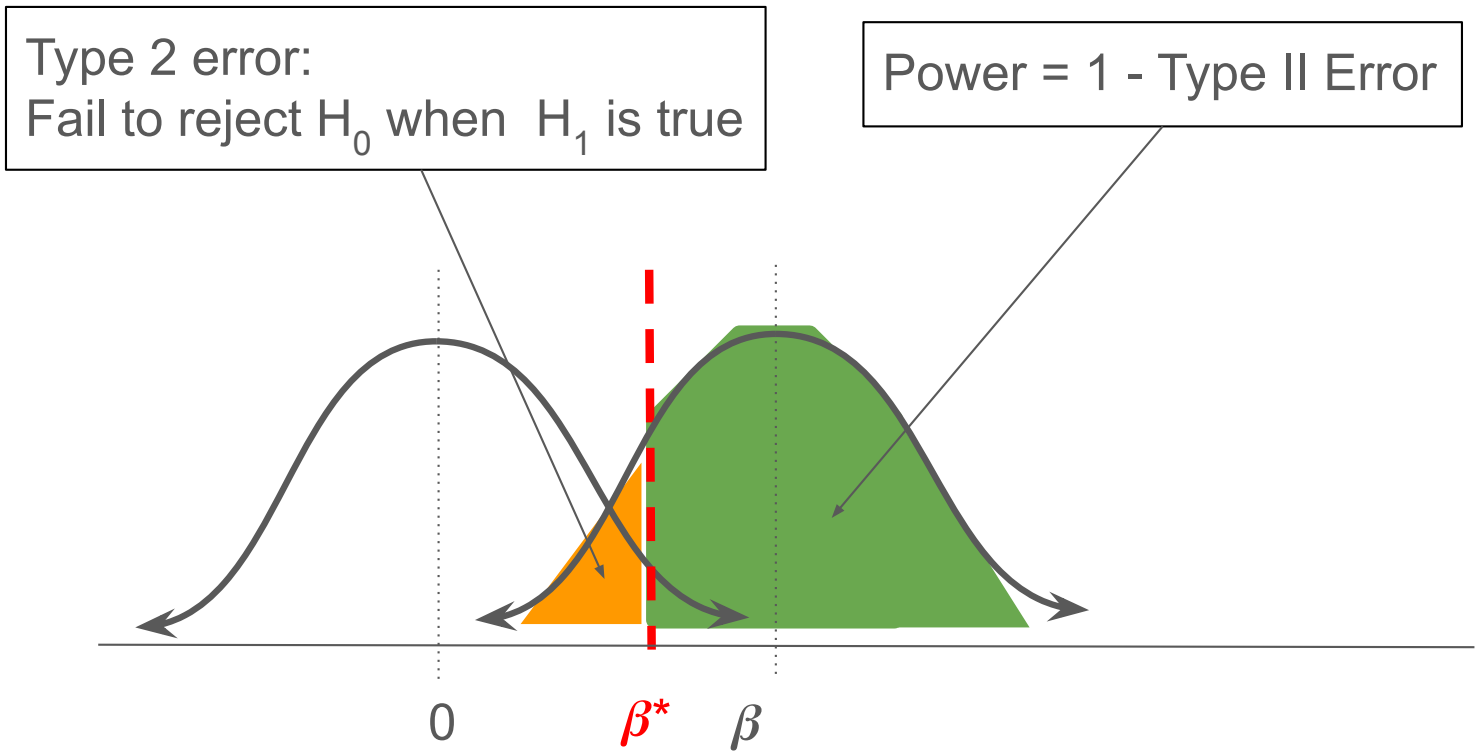
H_0: log hazard ratio = 0 vs. H_1: log hazard ratio = β

Review of power



H_0: log hazard ratio = 0 vs. H_1: log hazard ratio = β

Review of power



H_0: log hazard ratio = 0 vs. H_1: log hazard ratio = β

Power calculations in self-reported time-to-event data

- Estimating the standard error of the null distribution is a critical step
- This depends on the sampling distribution of β
- If self-reports are assumed to be perfect, the null distribution will be incorrectly estimated
- Incorporating sensitivity and specificity of self-reports is essential for correct power calculations

Using the modified Cox PH likelihood to estimate $SE(\beta)$

- Earlier, we showed the likelihood derived by Gu et al. (2015) [2], noted as $\ell(\theta, \beta)$
- To estimate the standard error of the sampling distribution of β under the null hypothesis, we use the Fisher information:

$$\mathcal{I}(\beta) = \mathbb{E}\left[-\frac{\partial^2}{\partial \beta^2}\ell(\theta, \beta)\right] \tag{6}$$

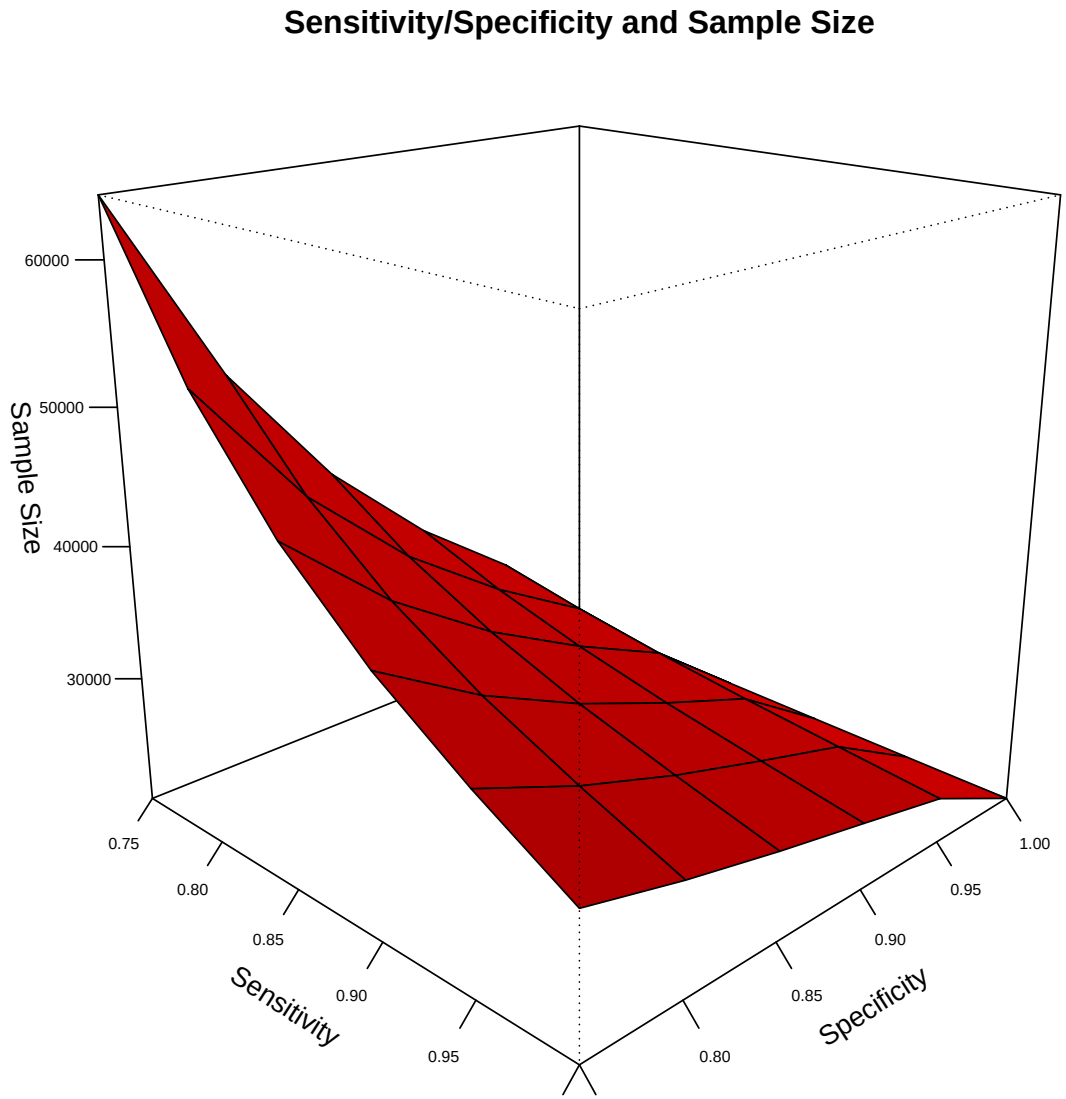
- Specific examples of this matrix are derived by Gu et al. 2016 [3]
- Estimation is implemented in the `icpower` function of the icmis package.

Other study characteristics affecting power

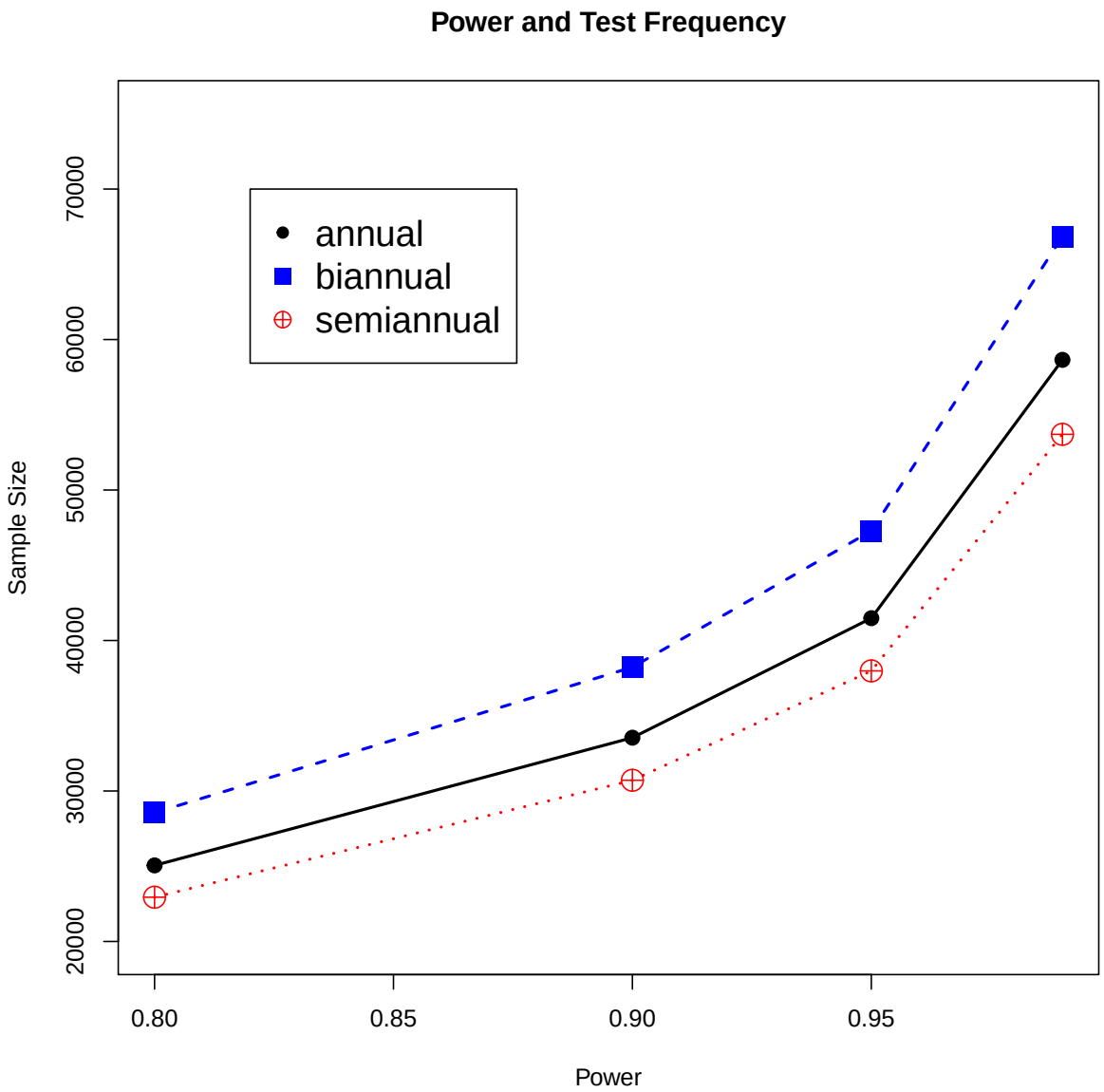
In the setting of error-prone, time-to-event data, there are two key factors affecting power beyond the usual questions of sample size and effect size:

- Testing strategy
  - "No Test after First Positive" (NTFP): Once a subject self-reports a positive result, they have no further tests.
  - "Missing Completely At Random" (MCAR): Testing continues throughout the study regardless of the value of the self reports. Any missing reports are assumed to be completely at random.
- Censoring
  - At any point in the study, individuals may drop out of the study or be lost to follow-up.
  - Censoring probabilities at each test time point can be specified when running power calculations in `icensmis`.
  - Independent censoring is assumed throughout.

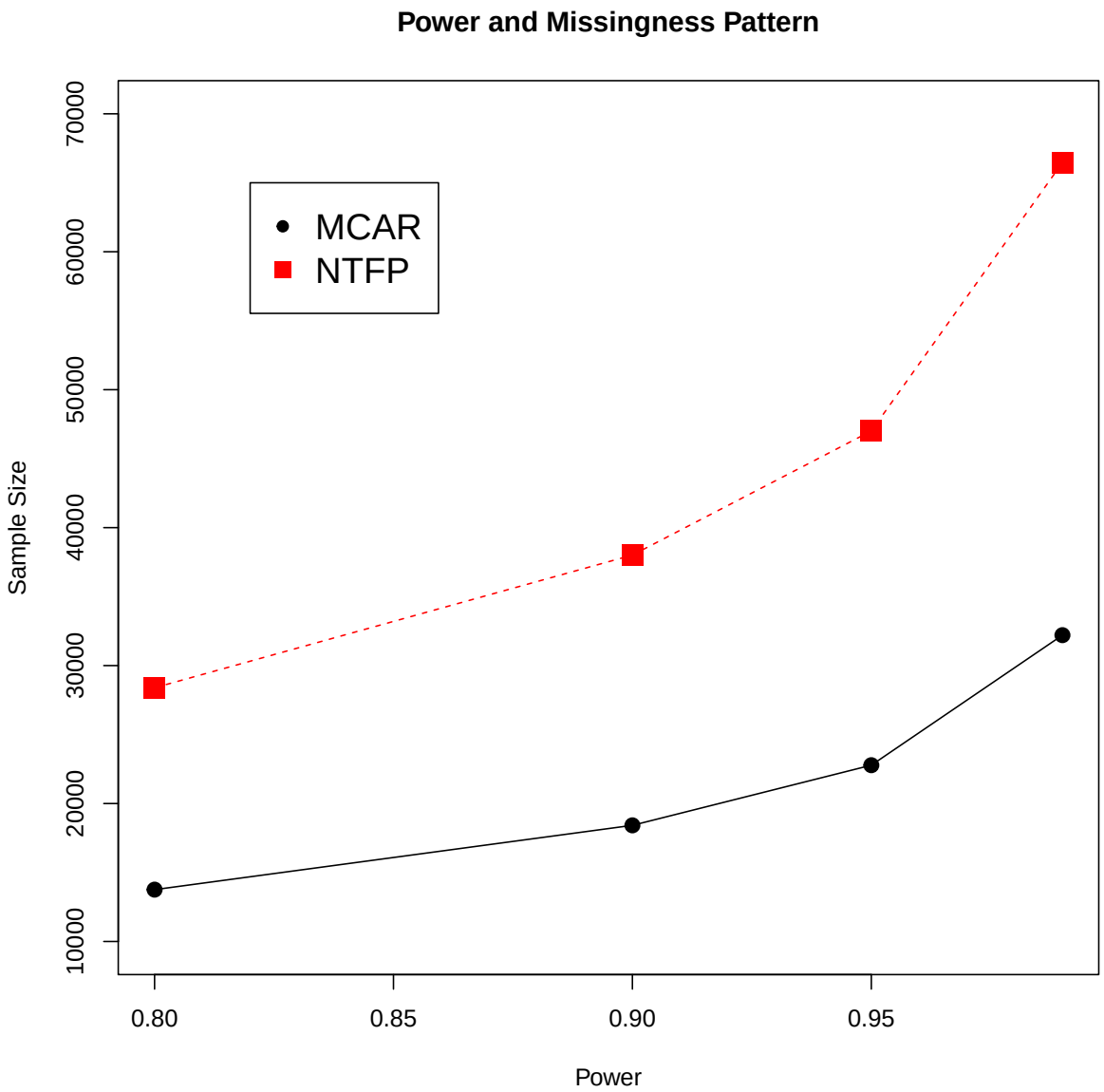
Necessary sample size varies non-linearly with test sensitivity and specificity



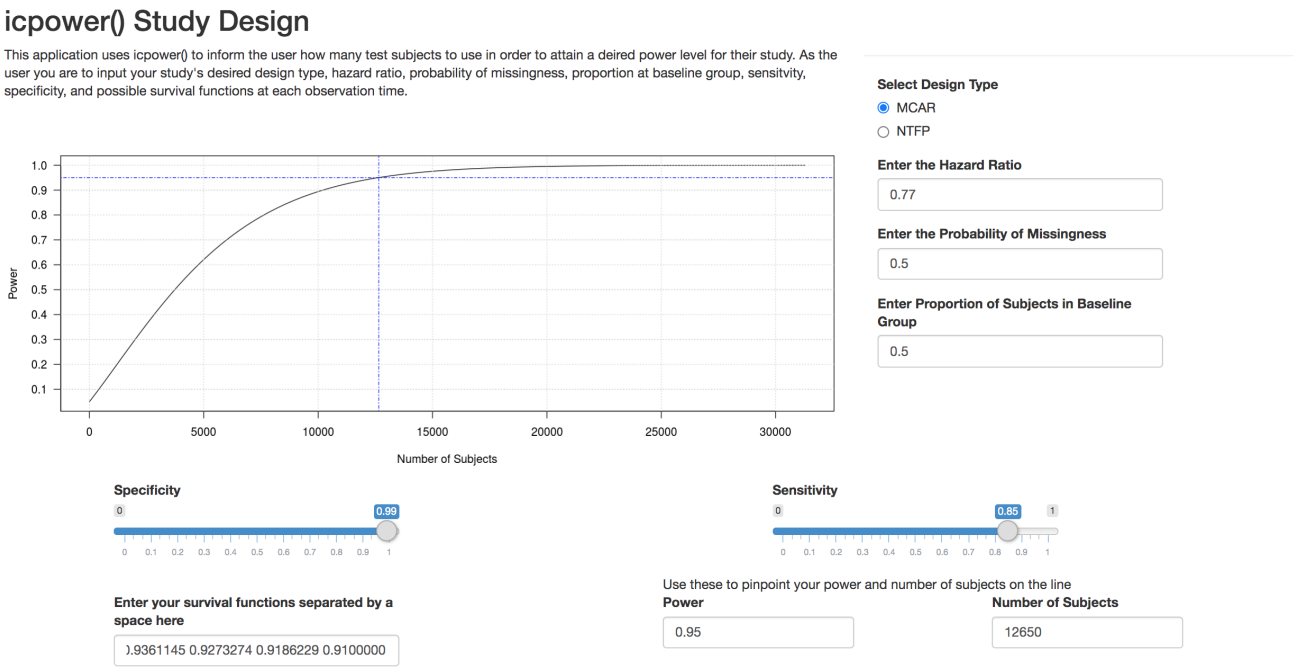
More frequent visits yield increased power



Power is higher for MCAR than NTFP study designs



Using icpower in our Shiny app



Ryan Sheehan

- Shiny app available at `https://raji-balasubramanian.shinyapps.io/icpower_graph/`

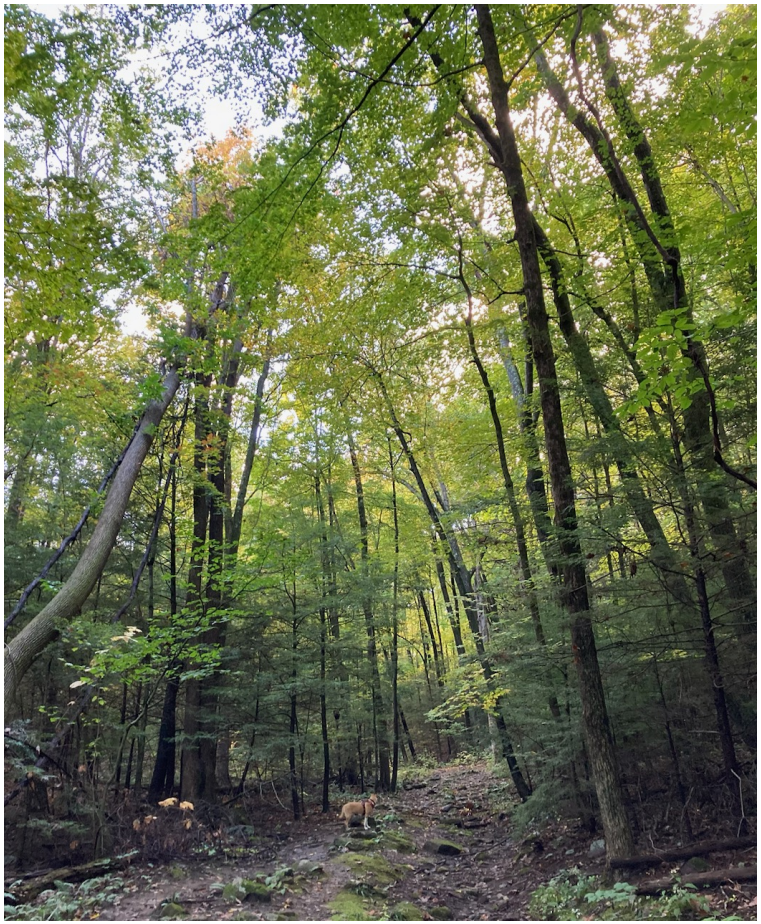
Table of Contents

High-dimensional predictors in self-reported, time-to-event data

- High-dimensional: number of samples n is smaller than the number of predictors p (usually $n << p$)
- Examples:
  - Study of plasma metabolomics profiles consisting of 500 assayed metabolites and cardiovascular disease onset in a cohort of ~ 100 individuals
  - Study of genetic variants and the time to diabetes onset, using a SNP array with hundreds of thousands of SNPs measured on ~ 1000 individuals
- We integrate tools for variable selection in high-dimensional data with the error-prone Cox PH model introduced above
  - `icRSF`: Random survival forests
  - `bayes_fit`: Bayesian variable selection

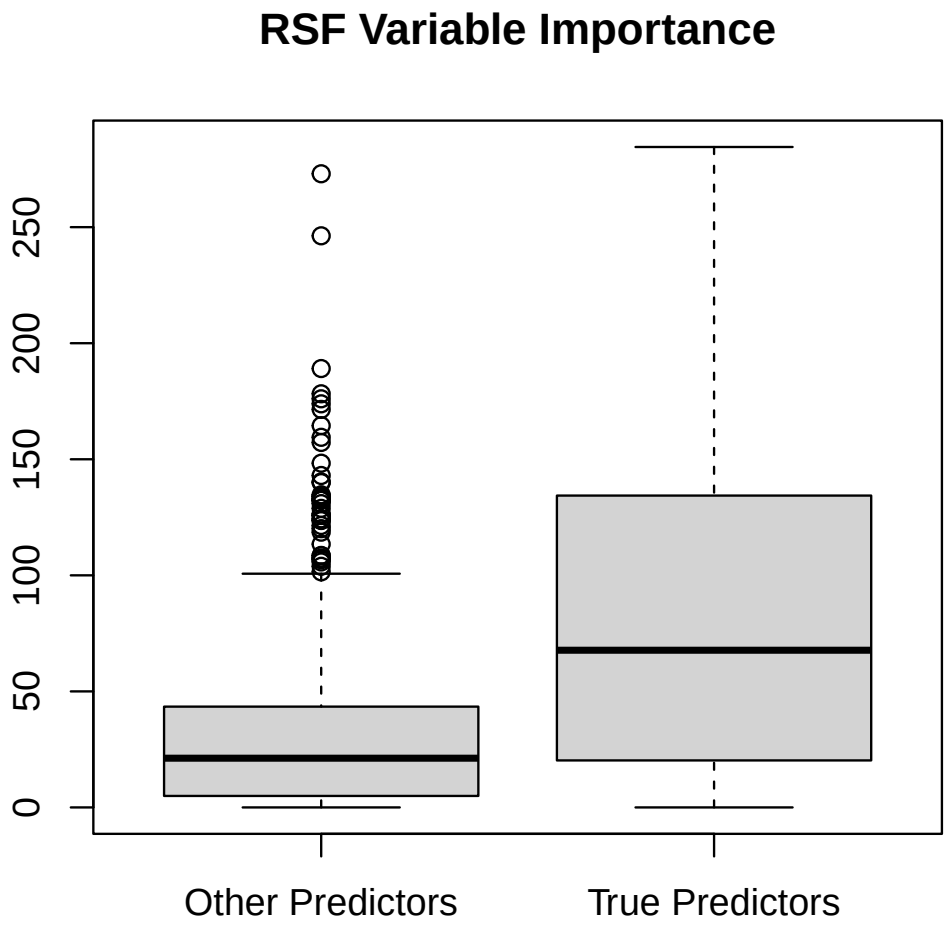
Interval-Censored Random Survival Forests (icRSF)

- A random survival forest is an ensemble of tree models
- In icRSF, trees are grown based on using our modified Cox PH likelihood to define a splitting criterion
- Variable importance is quantified based on these likelihood differences across all the trees
- RSF is a previously established method (Ishwaran 2008) [4]
- The contribution of icRSF is to account for the error-prone nature of the data



RSF on Simulated Data

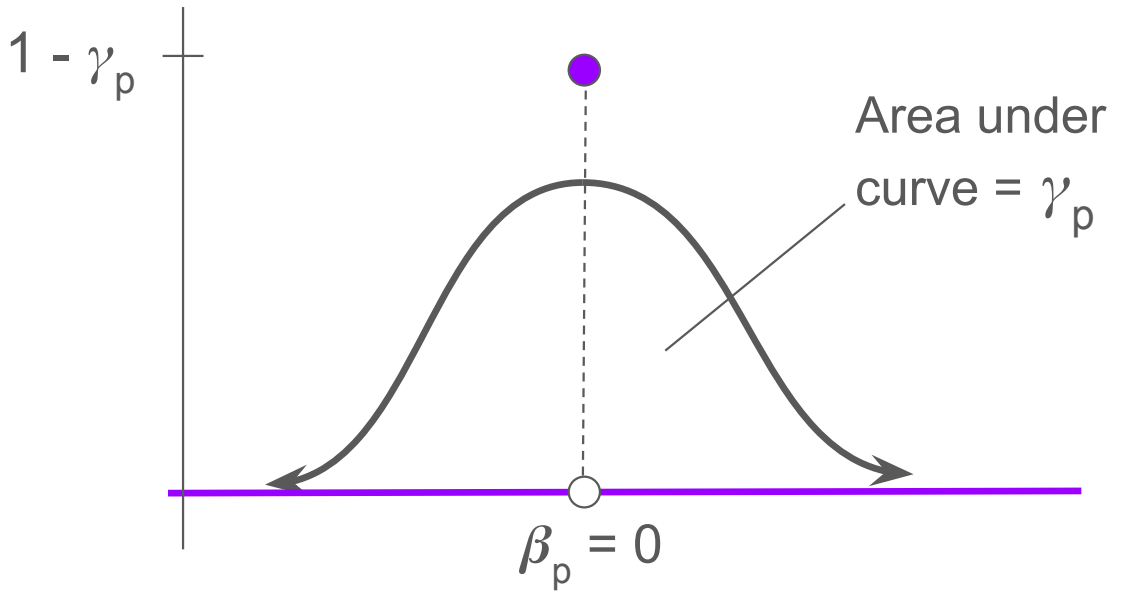
Figure 1: icRSF variable importance in simulated data with $N = 500$ and $P = 1000$, with 50 true predictors



Bayesian Variable Selection (BVS) Framework

Survival probabilities θ are assumed to have a Dirichlet(1) prior, while a spike-and-slab prior is assumed for each predictor's effect β_p:

$$\theta \sim Dirichlet(\mathbf{1})$$
$$\beta_p|\gamma_p \sim \gamma_p N(0, b^2) +$$
$$\gamma_p|\omega \sim Bernoulli(\omega)$$
$$\omega \sim Beta(w_1, w_2)$$



- Note that BVS is previously established work [5, 6]
- The contribution of `icensmis` is to account for the error-prone nature of the data by model fitting with our modified Cox PH likelihood

BVS on Simulated Data

Figure 2: BVS variable importance in simulated data with $N = 500$ and $P = 1000$, with 50 true predictors

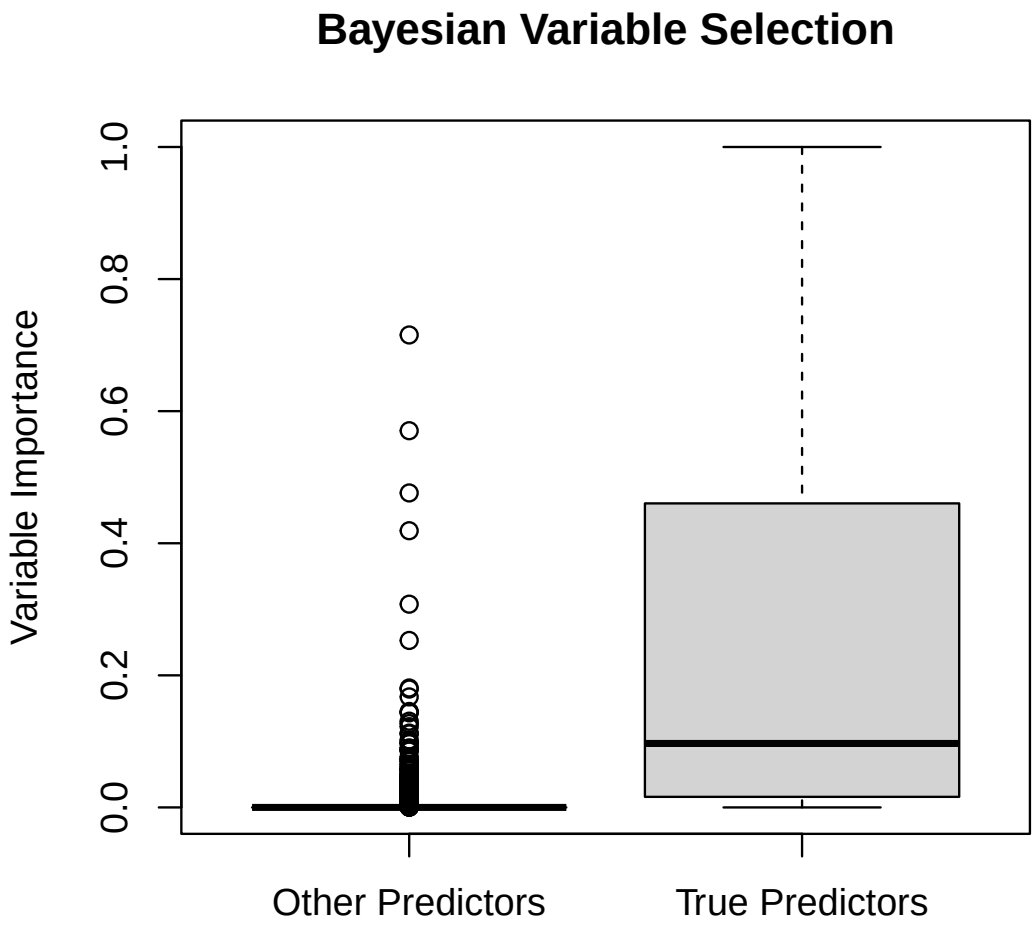


Table of Contents

Demonstration: Using the `icmis` function in R

```
mod_fit <- icmis(subject = ID,
                 testtime = time,
                 result = result,
                 data = icmis_data_with_trt,
                 sensitivity = sens,
                 specificity = spec,
                 formula = ~ factor(trt_groups),
                 param = 3,
                 control = list(maxit = 1000))
```

- `subject`: a unique identifier for each individual
- `testtime`: time of the self-report
- `result`: 0 for no event, 1 for event
- `data`: data frame including covariates
- `sensitivity`: φ_1
- `specificity`: φ_0
- `formula`: an R formula object specifying the covariates for the modified Cox PH model
- `param`:
- `control`: Arguments passed on to `optim` for numerical optimization

Demonstration: Designing a study of sex differences in stroke

Let's design a study that will use self-reports to assess sex differences in stroke incidence.

- Endpoint: incident stroke
- Study duration: 10 years

Sensitivity and specificity for self-reported stroke based on estimates from a cardiovascular health study (Eliassen et al. 2016 [7]):

- Sensitivity $\varphi_1 \approx 0.81$
- Specificity $\varphi_0 \approx 0.995$

Disease-specific characteristics, estimated from a previous study of sex differences in stroke among older individuals (Tsodak et al. 2012[8]):

- Estimated incidence of stroke: 1.6 strokes per 100 person-years, averaged across males and females
- Effect size to detect: Hazard ratio of 1.14 for stroke in women vs. men

Other study characteristics:

- Missingness mechanism: MCAR
- Missingness probabilities: 0.01 at each visit
- Censoring probabilities: 0.05 at each visit

Demonstration: using `icpower` in R

```
power80 = icpower(survivals = mySurv,
                  HR = 1.14,
                  sensitivity = 0.81,
                  specificity = 0.995,
                  power = 0.8,
                  rho = 0.5,
                  alpha = 0.05,
                  pmiss = 0.02,
                  pcensor = 0.01,
                  design = "MCAR" )
```

- `survivals`: baseline survival probabilities, usually estimated from cumulative incidence
- `HR`: Hazard ratio to detect (effect size)
- `sensitivity`: φ_1, estimated sensitivity
- `specificity`: φ_0, estimated specificity
- `power`: Desired power level
- `rho`: Percentage of participants in the treatment group
- `alpha`: Type 1 error rate
- `pmiss`: Probability of a randomly missing self-report at each test time
- `pcensor`: Censoring probability at each test time
- `design`: Missingness design, one of NTFP or MCAR

Demonstration: icRSF

```
rsf_res = icrsf(data=simdata,
                subject=ID,
                testtimes=time,
                result=result,
                sensitivity=0.8,
                specificity=0.995,
                Xmat=Xmat,
                root.size=10,
                ntree=100,
                ns=sqrt(ncol(Xmat)),
                node=8)
```

- `data`: a data frame containing `subject`,`testtimes`, and `result`
- `subject`: column name indexing subject in `data`
- `testtimes`: column name indexing test times in `data`
- `result`: column name indexing whether or not the test has occurred in `data`

- `sensitivity`: φ_1
- `specificity`: φ_0
- `Xmat`: $N \times P$ matrix of samples and covariates
- `root.size`: node size at which to stop growing the tree
- `ntree`: number of bootstrapped trees to grow
- `ns`: number of predictors p^* to search for splitting criteria at each node
- `node`: for parallel computing, number of compute nodes to use

Demonstration: Bayesian Variable Selection

```
bvs_result = bvs_fit(data = simdata,
                     Xmat = Xmat,
                     sensitivity = 0.8,
                     specificity = 0.995,
                     b = 1,
                     om1 = 5,
                     om2 = 100,
                     niter = 50000,
                     psample = 0.3,
                     initsurv = 0.9,
                     nreport = 1000,
                     nburn = 10000)
```

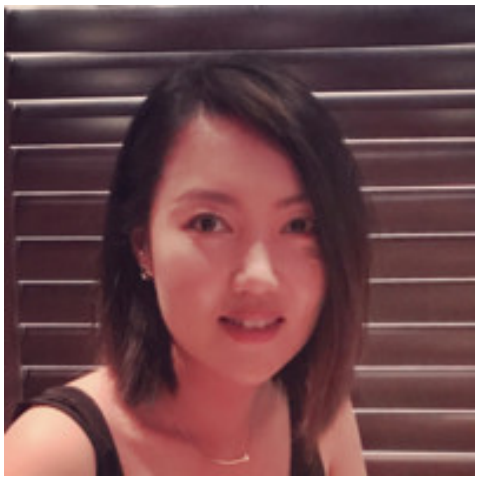
- `data`: a data frame containing `subject`,`testtimes`, and `result`
- `Xmat`: $N \times P$ matrix of samples and covariates
- `sensitivity`: φ_1
- `specificity`: φ_0
- `b`: hyperparameter; standard deviation of the spike portion of the spike-and-slab prior for the coefficients β
- `om1`: hyperparameter; shape parameter of the Beta prior distribution for ω
- `om2`: hyperparameter; second shape parameter of the Beta distribution for ω
- `niter`: number of MC iterations to run
- `psample`: Monte Carlo sampling parameter for updating coefficients
- `initsurv`: proportion of subjects not experiencing the event by end of study
- `nreport`: console log from variable fitting; parameters will be printed every `nreport` iterations
- `nburn`: number of MC iterations to discard as a burn-in period

Table of Contents

Summary of tools for self-reported time-to-event data

Function	Purpose	Reference
`icmis`	Estimate survival functions and hazard ratios	Gu et al. 2015 [2]
`icpower`	Study design and power calculations	Gu et al. 2016 [3]
`icRSF`	Variable selection with random survival forests	Xu et al. 2018 [9]
`bvs_fit`	Variable selection with Bayesian methods	Gu et al. 2020 [10]

Note that all of the tools above are also applicable to the more general case of interval censored data by setting sensitivity and specificity to 1 (i.e., assuming perfect self-reports).

Xiangdong Gu; Hui Xu

Resources

- The R packages `icensmis` and `icRSF` are available via CRAN
- Source code is available on Github at `https://github.com/XiangdongGu/icensmis` and `https://github.com/cran/icRSF`
- Our Shiny app for power calculations is at: `https://raji-balasubramanian.shinyapps.io/icpower_graph/`
- Slides and code from this presentation are at: `https://katehoffshutta.github.io/web/icensmis/20240229_CSP_selfreports.html`

Acknowledgements

The Balasubramanian Lab and Alumnae (UMass Amherst)

- Raji Balasubramanian
- Xiangdong Gu
- Hui Xu
- Ryan Sheehan
- Minsu Kim
- Yukun Li
- Yibai Zhao

This research was funded in part by National Heart, Lung, and Blood Institute of the NIH under awards:

- R01HL122241 (RB)
- T32HL007427 (KHS)

UMassAmherst | School of Public Health & Health Sciences
Biostatistics and Epidemiology

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

References

[1] Karen L Margolis et al. "Validity of diabetes self-reports in the Women's Health Initiative: comparison with medication inventories and fasting glucose measurements". In: *Clinical trials* 5.3 (2008), pp. 240–247.

[2] Xiangdong Gu, Yunsheng Ma, and Raji Balasubramanian. "Semiparametric time to event models in the presence of error-prone, self-reported outcomes—with application to the women's health initiative". In: *The annals of applied statistics* 9.2 (2015), p. 714.

[3] Xiangdong Gu and Raji Balasubramanian. "Study design for non-recurring, time-to-event outcomes in the presence of error-prone diagnostic tests or self-reports". In: *Statistics in medicine* 35.22 (2016), pp. 3961–3975.

[4] Hemant Ishwaran et al. "Random survival forests". In: (2008).

References

[5] Toby J Mitchell and John J Beauchamp. "Bayesian variable selection in linear regression". In: *Journal of the american statistical association* 83.404 (1988), pp. 1023–1032.

[6] Edward I George and Robert E McCulloch. "Variable selection via Gibbs sampling". In: *Journal of the American Statistical Association* 88.423 (1993), pp. 881–889.

[7] Bent-Martin Eliassen et al. "Validity of self-reported myocardial infarction and stroke in regions with Sami and Norwegian populations: the SAMINOR 1 Survey and the CVDNOR project". In: *BMJ open* 6.11 (2016), e012717.

[8] Meytal Avgil Tsadok et al. "Sex differences in stroke risk among older patients with recently diagnosed atrial fibrillation". In: *Jama* 307.18 (2012), pp. 1952–1958.

References

[9] Hui Xu et al. "A modified random survival forests algorithm for high dimensional predictors and self-reported outcomes". In: *Journal of Computational and Graphical Statistics* 27.4 (2018), pp. 763–772.

[10] Xiangdong Gu et al. "Bayesian variable selection for high dimensional predictors and self-reported outcomes". In: *BMC medical informatics and decision making* 20 (2020), pp. 1–11.