Validation of Survival Models

Terry Therneau

Nov 2021

- Joint work with TG6
- David McLernon, Daniele Giardiello, Ben Van Callister, Laure Wynants, Marten van Smeden, Terry Therneau, Ewout Steyerberg

Most important question

- Need a clear view of the *purpose* of the model — how will it be used?
- Risk stratification vs. patient counseling.
- Often leads to a time range
  - Advanced cancer with median = 3 years. One subject dies at year 7 and another at year 8; does it matter whether the model put them in the right order?
  - Which is the greater error: $(t = 6m\ ,\ \hat{t} = 1y)$ $(t = 5y,\ \hat{t} = 7y)$
  - The data itself has an upper limit

Most important question

- Need a clear view of the *purpose* of the model — how will it be used?
- Risk stratification vs. patient counseling.
- Often leads to a time range
  - Advanced cancer with median = 3 years. One subject dies at year 7 and another at year 8; does it matter whether the model put them in the right order?
  - Which is the greater error: $(t = 6m\ ,\ \hat{t} = 1y)$ $(t = 5y,\ \hat{t} = 7y)$
  - The data itself has an upper limit
- "If you don't know where you are going, you might end up someplace else" – Yogi Berra

Most important question

- Need a clear view of the *purpose* of the model — how will it be used?
- Risk stratification vs. patient counseling.
- Often leads to a time range
  - Advanced cancer with median = 3 years. One subject dies at year 7 and another at year 8; does it matter whether the model put them in the right order?
  - Which is the greater error: $(t = 6m\ ,\ \hat{t} = 1y)$ $(t = 5y,\ \hat{t} = 7y)$
  - The data itself has an upper limit
- "If you don't know where you are going, you might end up someplace else" – Yogi Berra
- Altman and Royston, What do we mean by validating a prognostic model?, Stat Med, 2000
- Korn and Simon, Measures of explained variation in survival data, Stat Med, 1990.

Targets

- Hazard ratio
- Time to event
  - Scaling issue
  - No good suggestions
- Observed vs expected number of deaths
  - standarized incidence ratio (SIR) of epidemiology
  - extends to subsets and/or regression
- P(alive at τ)
  - Natural for users
  - Parallel to binomial methods
  - Have to pick a time τ
- Measures of association
  - Concordance $P(t_i > t_j | \hat{t}_i > \hat{t}_j)$
  - R^2
  - Royston and Sauerbrei D
  - ...

The skeleton in the closet

- Interested in P(alive at 4 years)
- Validation subject censored at 2: we have $\hat{p}_i(4)$ but no $y_i(4)$
  - Redistribute-to-the-right (IPCW)
    - sensitivity, specificity, AUROC, ...
    - IPCW + R^2 = Brier score
    - assumes independent censoring
    - loss of covariates
  - $\hat{p}$ vs $\hat{p}$
    - $\hat{p}_1$ = predictions from reference model
    - $\hat{p}_2$ = Cox model + validation data, η as only predictor
    - assumes independent censoring given η
  - Ignore the censored observations
    - Biased (Berkson 1952, Kaplan and Meier 1958)
    - Prolific nonetheless
    - A Kamarudin, T Cox, D Kolamunnage-Dona, BMC 2017, Time-dependent ROC curve analysis in medical research: current methods and applications.

Pessimissm

- Take a favorite binomial/logisitic measure and fix it up a bit
  - "If it works for logistic it will work for censored data"
- This more or less holds when P(event) $< .25$
  but not in general
- "ROC" now raises warning flags, "Time dependent ROC"
  frank suspicion
- About 1/2 the papers that I read, carefully, are methods that I can not recommend in good conscience. (Good journals)

Frustration

- Much of the above is not in the draft paper: there is not space for it
  - My enthusiasm for taking good work and then fitting it for a straightjacket (a journal) has become low.
  - My enthusiasm for burying that work behind a paywall has gone to zero.
  - If someone does want to listen to us, we've made doing so a lot of work.
  - We are using an ineffective and obsolete model (IMHO).

Frustration

- Much of the above is not in the draft paper: there is not space for it
  - My enthusiasm for taking good work and then fitting it for a straightjacket (a journal) has become low.
  - My enthusiasm for burying that work behind a paywall has gone to zero.
  - If someone does want to listen to us, we've made doing so a lot of work.
  - We are using an ineffective and obsolete model (IMHO).
- I became involved as an adjuct to a TG6 effort
  - They had looked at and summarized/selected the literature
  - Many papers for me to read and absorb
  - Different perspective to absorb (of course)
  - Exciting to interact with good minds
  - (Guilt for significantly slowing their process)