1.2 Biostatistics meets ML 1

Angela Wood

Outline:

- Equip with key biostatistical tools and demonstrate their use in healthcare research
- Highlight common misconceptions about limitations of biostatistical methods
- Describe common benchmarks for AI and ML methods, esp for;
    - Modeling survival/ time-to-event data
    - Handling missing data
    - Risk predictions
- Objectives: Understanding:
    - Principles of Cox regression
    - Handling missing data using multiple imputation by chained equations (MICE)
    - Principles of developing and evaluating risk prediction tools

Part 1 Survival analysis

- Censored: if they have not experienced the event at the point they stop
    - Event has not occurred when study ends or at cut-off date
    - Lost to follow-up
    - Each individual contributes to time-to-event or time-to-censoring to analysis
        - Standard survival analysis assume non-informative censoring
            - An individual is censored, subsequent risk of even of interest is not affected
- Aims of analysis
    - Combination of continuous variable (time) and binary variable (event/ no event)
    - To estimate the survivor function based on survival data from a sample of individual
    - To compare overall survival experience between different groups of individuals
        - Between randomized group in CT
        - Between individuals with different baseline exposure in a cohort
- **Kaplan Meier** - probability of survival / survivor function S(t) = Pr(T>t)
    - 1. Raw data (survival times, noted for censored observation)
    - 2. Life table calculation: below features over time *t*
        - N alive ("at risk") before time *t*
        - N deaths at time *t*
        - P(death at time *t*)
        - P(survival until time *t*)
    - 3. Kaplan Meier curve: x-axis: time (*t*) vs y-axis: P(survival up to time *t*)
- **Hazard Function** h(t) : rate of event in next small time interval after time t

$$h(t) = \frac{Pr(t \leq T < t + \delta t \mid T \geq t)}{\delta t} \; as \; \delta t -> 0$$

    - Instantaneous rate of an event at time *t*
    - Hazard Ratio = negative log of survivor function differentiated over time
- **Hazard ratio and proportional hazards model**

- Hazard Ratio comparing groups 1 vs group 0:

  $$\psi = \frac{h_1(t)}{h_0(t)}$$

- Proportional hazards (PH) assumption:
  - **ψ** does not depend on *t*
  - Hazard ratio is constant at all time

- Proportional hazards model(PH model): $h_i(t) = h_0(t)\, e^{(\beta x_i)}$
  - $h_i(t)$: hazard function for individual *i* = 1, 2, …, n
  - x_i: binary covariate (group 0 or 1)
    - $h_i(t)$ = $h_0(t)$ if individual *i* is in group 0
    - $h_i(t) = h_0(t)\, e^{(\beta)}$ if individual *i* is in group 1
  - $e^{\beta} =$ Hazard ratio comparing group 1 vs group 0
    - β = log hazard ratio comparing group 1 vs group 0
    - $h_0(t)$ = baseline hazard function
- **Cox model**: a semi-parametric PH model

  $$h_i(t) = h_0(t)\, e^{(\beta x_i)}$$

  - βx_i = as many variables of different type interactions with other variables
    - linear/ non-linear functions (e.g. splines, polynomials, fractional polynomials)
  - β coefficients and standard errors estimated from partial log likelihood function
  - Semi-parametric: no assumption made about shape of $h_0(t)$ or T distribution
  - Could formulate model making assumptions about shape of $h_0(t)$ and distribution of T:
    - E.g. exponential, Weibull, Gompertz, log-normal, log-logistic, Gamma
    - If distributional assumption is valid, inferences more precise
    - Enables extrapolation of survivor functions beyond end of follow-up period
- **Standard extensions**
  - Stratified Cox model
    - Allow different baseline hazard function for different levels of covariate which does not satisfy the PH assumption
      - E.g. hazard function for individual *i* in stratum *s*:
      - $h_{iS}(t) = h_{0S}(t)\, e^{(\beta x_i)}$
  - Interactions with time
    - If PH assumption not met for particular covariates, could fit a model which allows association to differ in different periods of follow-up
    - E.g. hazard ratio differing for years 0-2 and year 2
  - Competing risks
    - Cause-specific hazard models:

        - Used when individuals experience alternative events which later P(main event)
- **Example: quantifying association between COVID-19 and CVD**
    - https://www.hdruk.ac.uk/projects/cvd-covid-uk-project/
    - Key question:
        - What are the long term risks of major arterial and venous diseases after COVID-19 infection?
    - Big data analytics:
        - 48 Mill adults: 1.4 M with COVID
        - Diagnosis between 1/1/2020 - 7/12/2020
        - Cox model
            - calendar -time scale
            - Time-dependent exposure = “weeks since COVID-19 diagnosis/ vaccination”
            - Adjustment for confounders
    - Higher risk of major vascular disease following COVID-19 infection
        - Knight et al., Circulation 2022: https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.120.049252
        - 30 times higher risk of arterial events
            - After around 12 weeks, remain elevated after 1 year
        - Venous events:
            - Remain elevated after 1 year
        - Hazard ratio expected to decrease over time
- **ML for survival data**
    - Approaches:
        - Random survival forest
        - Support vector regression
        - Gradient boosting
        - Multi-task learning
    - “No single model is best across all datasets, and frequently no single model is best across all time horizons within a single dataset”
    - https://www.vanderschaar-lab.com/survival-analysis-competing-risks-and-comorbidities/
        - Automated Machine Learning (AutoML)
- **Summary**
    - features: combines time to event variable and censoring indicator
    - aims: to estimate the probability of surviving beyond a time, and/or compare survival experiences between different groups
    - commonly modelled using cox regression - easily interpretable
    - Final remarks:
        - Often Cox regression is used a comparator for many ML models
        - Rarely the models are optimized

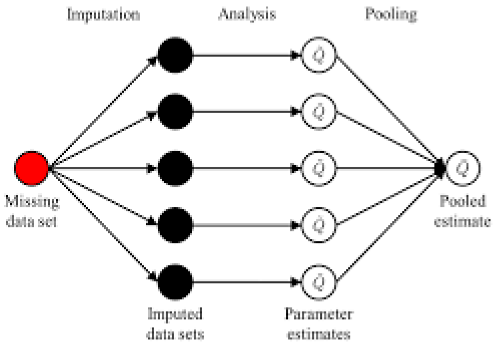
            - Easily allow for non-linear relationships of covariates/features, time interactions or allowance of non-proportional hazards, competing risk etc.

Part 2. How to handle missing data (Multiple imputation by chained equations)

- Recommended readings:
    - Sterne et al., 2009
        - Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls
        - https://www.bmj.com/content/338/bmj.b2393#:~:text=Multiple%20imputation%20is%20a%20general,obtained%20from%20each%20of%20them.
    - White et al., 2011
        - Multiple imputation using chained equations: Issues and guidance for practice
        - https://onlinelibrary.wiley.com/doi/10.1002/sim.4067
- **Effect of missing data**
    - Missing values irreversibly lose power
    - Analyses need untestable assumption
    - Wrong analyses can lead to:
        - Biased estimate, biased standard errors, inefficiency (loss of information)
    - Suitable method of analysis:
        - Give unbiased estimate, SE, and efficient
- **Missing data mechanism**
    - Missing completely at Random
        - P(M) not depend on values of observed or missing data
        - E.g. accidentally dropped sample
    - Missing at Random
        - P(M) depend on values of observed data but not the values of missing data
        - E.g. blood pressure data in younger population
    - Missing not at Random
        - P(M) depend on values of missing data
        - E.g. blood pressure data of individuals with normal bp
- **Missing data assumption**
    - MCAR: can be told at hand
        - Made more likely by collecting explanatory variables, explaining missingness
    - MAR/ MNAR: cannot be told from data
        - MNAR setting is rare to known appropriate model
- **Describing missing data**
    1. Look at % missing for each variable
    2. Look at missing data patterns
    3. Compare % of missing data by variables of interest(in trials - by randomised group, in epi- predictors of missingness)

    4. Compare reasons for missing data(by variable of interest)
    5. In trials, look at baseline imbalance among those with observed outcome

- **Analysis in the presence of missingness**
    - Taking incomplete dataset into a completed rectangular dataset
        - Complete case
        - Missing data indicator
        - Multiple imputation by chained equations
        - Random-Forest imputation
    - Complete-case analysis
        - Inefficient
        - Could be biased under many missing data mechanisms
        - Default
        - Complete case regression analysis:
            - Partially missing single outcome with fully observed covariates:
                - Reasonable:individuals with missing outcome dont carry much information
                - Fine under MCAR
                - Fine under MAR if adjusted for covariates which predic missingness
                - Biased under MNAR
            - Fully observed single outcome with partially missing covariates:
                - Unbiased if P(complete-case) independent of the outcome given the observed covariates
                    - Loss of power
                - Similar (not exactly the same) assumption as MAR
                - Unbiased with MCAR + Loss of power
                - Biased under MNAR + Loss o power
- **Regression imputation**
    - Preserve relationships between variables but exaggerates correlations
    - One stochastic imputation
        - Over-precise as treating imputed value as correct
    - Multiple imputation



        - Combining estimates from multiple imputation analysis
            - Overall estimate: $Q * = \frac{1}{m}\sum_{J-1}^{m} \quad Q_j$

- Let $Q_1, \ldots, Q_m$ be *m* estimates of interest
- $W_1, \ldots, W_m$ be corresponding variance estimates
- Two sources of uncertainty;
  - Within-imputation variance component:
    - $W = (W_1 + \ldots + W_m)/m$
  - Between-imputation variance component:
    - $B = (Q_1 - Q^*)^2]/(m-1)$
  - Total variance of $Q*$
    - $T = var(Q^*) = W + (1+1/m)B$

- **Predictive mean matching (PMM)**
  - Regression approach takes 4 steps in imputing – “draw” method
  - PMM imputes using the observed value of a suitable donor
  - Advantages:
    - Always impute observable values
    - More robust to model mis-specification
  - Morris et al., 2014:
    - Tuning multiple imputation by predictive mean matching and local residual draws
    - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4051964/
  - Details: from https://statisticalhorizons.com/predictive-mean-matching/
    1. For cases with no missing data, estimate a linear regression of x on z, producing a set of coefficients b.
    2. Make a random draw from the “posterior predictive distribution” of b, producing a new set of coefficients b*. Typically this would be a random draw from a multivariate normal distribution with mean b and the estimated covariance matrix of b (with an additional random draw for the residual variance). This step is necessary to produce sufficient variability in the imputed values, and is common to all “proper” methods for multiple imputation.
    3. Using b*, generate predicted values for x for all cases, both those with data missing on x and those with data present.
    4. For each case with missing x, identify a set of cases with observed x whose predicted values are close to the predicted value for the case with missing data.
    5. From among those close cases, randomly choose one and assign its observed value to substitute for the missing value.
    6. Repeat steps 2 through 5 for each completed data set.
- **MICE**
  - Aka fully conditional specification (FCS), sequential regression multivariate imputation, etc…
  - Suppose a dataset with variables x1, x2, and x3, and *n* observations with any data observed:
    - x1 and x2 contain some missingness

- Missing values assumed to be MAR
- Continuous variables assumed to be normal
- Aim: get a completed *n* x 4 multivariate sample – imputed dataset
- Steps taken:
  1. Initialize: fill in x1 and x2 by borrowing random observed values – resulting $x1*^{(0)}$ and $x2*^{(0)}$
  2. For each cycle:
     - a. Univariate imputation for x1, regressing the observed x1 on x3, x4, and $x2*^{(0)}$
     - b. Replace missing values in x1 to give $x2*^{(0)}$
     - c. Regress observed x2 on x3, x4, and $x1*^{(0)}$ (to impute $x2*^{(0)}$)
  3. Repeated until about 10 cycles completed (stabilize result such that imputed values unaffected by starting values) — imputed dataset
  4. Repeat whole procedure *m* times to give *m* imputed datasets

- Why?
  - Under MAR assumption, Missing imputation gives unbiased estimates, SE, and is efficient
  - Defends against the charge of “making up data”
  - Incredibly flexible: covers many different data structures and types
    - Use different imputation regression models for different data type
  - Useful but not the absolute best
  - Sophisticated robust statistical software widely available
- More FAQs about MICE
  - How many imputations to do?
    - *m* ≥ % of incomplete cases (e.g. 13% -> *m* = 15)
      - von Hippel 2018: https://journals.sagepub.com/doi/abs/10.1177/0049124117747303
      - White 2011: https://onlinelibrary.wiley.com/doi/10.1002/sim.4067
  - How much missing data?
    - Large amounts (may converge slowly)
    - Impact of missing data is greater
      - Departures from MAR will be very influential
  - What variables should go in an imputation model?
    - Analysis and imputation models should be “compatible”
    - As a minimum, the imputation model must contain all the variables in the analysis model
    - In particular, it must include outcome of the analysis model
    - Should consider non-linear associations with interactions
  - How do I include a survival outcome?
    - Suppose analysis model is:
      - $h_i(t) = h_0(t)\ exp(\beta_1 X_{1i} + \beta_2 X_{2i} + ..)$

- Need to impute incomplete binary/ continuous X_{1i}
- Correct imputation model is approximately a logistic/ linear regression of X_{1i} on D_i ,$H_0(T_i)$ and X_{2i} ,...
- Instead of $H_0(T_i)$, we use Nelson-Aalen estimator of H (T_i)
  - White & Royston (2009)
  - https://onlinelibrary.wiley.com/doi/10.1002/sim.3618

- Example - **Sudden Infant Death Symdrome (SIDS)**
  - Data available from a large Scottish database, linking biochemical, pregnancy, birth and death records for 214, 532 liveborn…
  - The problem of missing data
    - Potential confounders:
      - Maternal characteristics: age, height, deprivation category (1-7), smoking status (non, smoker, ex)
      - Pregnancy: gender, weight, gestation
    - Approximately 10% missing data in confounders, reducing the number of SIDS deaths to 99
      - Unlikely that missing data is related to outcomes
  - SIDS: Complete-case analysis
    - Outcome = SIDS, Exposure = AFP levels, Model = logistic regression
    - Unadjusted logistic model in all individuals
      - On full data
      - Unadjusted OR: 2.01
    - Unadjusted logistic model in complete-cases
      - Only on complete cases
      - Unadjusted OR: 1.94
    - Adjusted logistic model in complete cases, adjusted for confounders
      - Only on complete cases
      - Adjusted OR: 1.60
    - Adjusted OR on full data is calculated with MI
      - Objective
        - To impute missing confounder values using all other baseline covariates and SIDS outcome:
          - Assuming data are MAR
          - Enable us to include all SIDS case – no loss of efficiency
      - Adjusted OR: 1.54
- **Key points from Part 2**
  - Complete-case analysis not always doomed…
    - When imputing outcome only, imputation unnecessary unless you are using auxiliary variables in imputation model
  - MICE is very flexible – but I caution against it's "default black-box approach use", as many user-decisions are required for optimization (e.g. conditional imputation)
  - Must include outcome of interest in imputation model

- Should check imputed values seem reasonable
    - Although they may not necessary represent the observed data due to MAR assumptions
- Many extensions
    - Multi-level data,
    - Allowing deviation from MAR to MNAR

Q&A: