BIOS 6312: Modern Regression Analysis

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 13: The Process of a Data Analysis

Version: 03/11/2021

Before we start

Preface to notes:

- This is a course in methodology, though we have relied heavily on real-world data sets in order to illustrate:
  - Implementation.
  - Interpretation.
- We've largely stayed away from the process of analyzing a data set from start to finish (there are other courses that suit this purpose).
- However, going through some strategies for a real-world data analysis together is a great way to bring together many of the course concepts.

Outline

Topics:

- Background: Prostate cancer and PTEN expression
- Clinical questions
- Data set
- Statistical challenges and considerations
- Descriptive statistics (dos and don’ts)
- Modeling

Outline

Topics:

- **Background: Prostate cancer and PTEN expression**
- Clinical questions
- Data set
- Statistical challenges and considerations
- Descriptive statistics (dos and don'ts)
- Modeling

Background

Discussion point: Epidemiology

- When working in a particular disease area, it is very helpful to have a working understanding of the epidemiology pertaining to that disease.
  - Orients clinical question.
  - Can influence study design.
  - Can set inclusion-exclusion criteria.

Background

Prostate cancer: Epidemiology

- Second most frequent cancer diagnosis among men.
- Over 1.2 million incident cases reported worldwide in 2018.
- Age strongly related to incidence and mortality rates.
- Fifth leading cause of death worldwide.
- Often asymptomatic at the early stage.
- Biochemical recurrence a strong predictor of mortality.

Background

Discussion point: Underlying biology

- When working in a particular disease area, it is very helpful to have a working understanding of some of the underlying biology (even if the understanding is somewhat superficial, the goal is to understand it to a degree that you could at least explain it to a scientifically savvy researcher).
    - Precursor to possible statistical challenges.
    - Bridge between high-level clinical questions and specific statistical measures of association.
    - If a clinician needs to understand statistics to some degree, then it's only fair for the biostatistician to understand the biology to some degree.

Background

Prostate cancer: Morphology

- Gleason pattern characterization developed in 1960s.
  - **Pattern 1**: Cancerous prostate closely resembles normal tissue. Glands are small, well-formed.
  - **Pattern 2**: Glands still well formed, but larger and with more tissue between them.
  - **Pattern 3**: Glands identifiable, but darker cells; cells beginning to invade the surrounding tissue.
  - **Pattern 4**: Few recognizable glands; poorly differentiated.
  - **Pattern 5**: Very few or no recognizable glands; often appear as sheets of cells throughout the surrounding tissue.
- `http://pathology.jhu.edu/prostatecancer/newgradingsystem.cfm`

Background

Prostate cancer: Morphology

- Gleason patterns can have sub-patterns.
- Within Gleason pattern 4:
  - **Cribriform sub-pattern**: sheets of epithelial cells punctuated by empty/gland like spaces.
  - **Poorly formed sub-pattern**: includes glands with no/rare lumens, elongated compressed glands, and elongated nests.
  - **Glomeruloid sub-pattern**: dilated glands containing intraluminal cribriform structures with a single point of attachment.
- `https://en.wikipedia.org/wiki/Gleason_grading_system`

Background

Prostate cancer: Intraductal carcinoma

- To make matters even *more* confusing, there is another pattern (not defined within the Gleason pattern) known as intraductal carcinoma, a pattern that is indistinguishable from invasive cribriform without further staining.
- Many studies do not differentiate between invasive cribriform and intraductal carcinoma, but there are conflicting results regarding the role of each in predicting poor outcomes.

Background

Prostate cancer: PTEN expression

- Phosphatase and tensin homolog (PTEN) is a protein that is encoded by the PTEN gene (enzyme acts as a tumor suppressor, regulating cell division).
- Mutations of the PTEN gene are a step in the development of prostate cancer.
- Loss of PTEN expression poor prognostic factor.
  - Loss can be homozygous or heterozygous.
- PTEN loss can occur within each of the Gleason patterns and/or sub-patterns discussed.
  - For instance, a patient can have heterozygous PTEN loss within their cribriform sub-pattern, homozygous PTEN loss within their Pattern 3 component, and have PTEN remain fully intact within their Pattern 5 component.

Outline

Topics:

- *Background: Prostate cancer and PTEN expression*
- **Clinical questions**
- Data set
- Statistical challenges and considerations
- Descriptive statistics (dos and don'ts)
- Modeling

Clinical questions

Discussion point: Forming questions

- In the ideal scenario, questions are sufficiently refined before the data are collected.
  - Most common in randomized controlled trials.
- Not always reality. Sometimes, investigators prioritize access to covariate-rich observational databases and then work with a team of biostatisticians and epidemiologists to help refine clinical questions.
  - Should be refined before data are analyzed, at the very least.
- Investigative goals can often be quite general, particularly at the earlier stages of research. One should be able to state clinical question(s) both in general terms and in unambiguous statistically well defined language.

Clinical questions

Prostate cancer: Study goals

1. To evaluate the degree to which different morphology patterns are associated with poorer outcomes.
2. To compare the frequency of PTEN loss in each pattern.
3. To understand the role of PTEN loss in predicting biochemical recurrence.

Outline

Topics:

- *Background: Prostate cancer and PTEN expression*
- *Clinical questions*
- **Data set**
- Statistical challenges and considerations
- Descriptive statistics (dos and don'ts)
- Modeling

Data

Discussion point: Data management

- Work with a clean copy of the data with a clear corresponding codebook including variable names, descriptions, coding (categorical) and/or units (continuous), other comments (including references as applicable).
  - You will thank yourself for putting together a clear codebook and clean data set. The process that takes you from raw data to clean data must be completely reproducible.
- Must distinguish between values that are circumstantially missing (e.g., age) and variables that do not apply (e.g., nodes testing positive for subjects having none removed).
- The following mean nothing to Stata or R:
  - Missingness codes such as 777, 888, 999.
  - Color coding in excel spreadsheets.
  - Extra notes at the bottom of an analytic spreadsheet.

Data

Prostate cancer data: Study

- Observational study of $N = 164$ Vanderbilt patients with prostate cancer, confirmed my total prostatectomy.
  - # of recurrences: 77
  - # of deaths: 2
- Possible lymphadenectomy at time of surgery as well (removal of surrounding lymph nodes).
  - Underlying biology: removal or more lymph nodes is understood to be associated with better outcomes, even if those nodes do not test positive for cancer.
- Baseline demographics, clinicopathological characteristics, morphology breakdown at time of surgery.
- Records of biochemical recurrence and death obtained from a mixture of prospective follow-up and retrospective examination of electronic medical records.

Data

Prostate cancer data: Baseline demographics (time of surgery)

- Age (years).
- Race (white or other).
- Preoperative PSA (prostate-specific antigen).
  - Underlying biology: PSA is a biomarker that serves as a correlate of cancer.

Data

Prostate cancer data: Surgery

- Indicator of lymphadenectomy.
- Indicator of lymph nodes testing positive.
- Indicator of positive margins.
- Tumor volume.

Data

Prostate cancer data: Morphology

- Overall Gleason breakdown, differntiating between intraductal carcinoma and cribriform (two ways):
    - Dichotomized (indicator of each pattern's presence).
    - Granular (percentage of total tumor).
- Gleason pattern 4 sub-pattern breakdown (two ways):
    - Dichotomized (indicator of each sub-pattern's presence).
    - Granular (percentage of total pattern 4).
- In these data, all patients had Gleason pattern 4 and none had Gleason patterns 1 or 2.

Data

Prostate cancer data: Outcomes

- Time to death.
- Time to recurrence.

Outline

Topics:

- *Background: Prostate cancer and PTEN expression*
- *Clinical questions*
- *Data set*
- **Statistical challenges and considerations**
- Descriptive statistics (dos and don'ts)
- Modeling

Statistical considerations

Discussion point: Anticipating sticky points

- Understanding what variables were measured (and how they were measured) will allow you to anticipate possible challenges.
- Knowing likely challenges and how you're going to address them will help shape the structure of your analysis.

Statistical considerations

Missing data:

- You should anticipate some level of missing data in almost any study.
- Plan in advance to handle missing data in a principled way.
  - For instance, multiple imputation via chained equations with $M = 500$ iterations.
- Determine the variables that will go into the imputation procedure *a priori*.
- Try not to let anyone relegate the results from the imputation model to a secondary/exploratory analysis in order to favor a complete-case analysis. The most principled approach should be the primary one.

Statistical considerations

Confounding:

- These data are observational in nature; one should anticipate that there will be some level of confounding.
- You must rely on the insights of your clinical collaborators to identify possible confounders of the association of interest.
  - Age.
  - Preoperative PSA.
  - Lymphadenectomy data.
  - Positive margins.
  - Cancer stage (T2, T3a, T3b).

Statistical considerations

Death as a competing risk:

- Right off the bat, we know we're not going to be able to analyze the data using time to death as the outcome (only two deaths).
- Death serves as competing risk for biochemical recurrence.
- The most principled approach would likely to be to use a competing risks model (subdistribution hazard).

Statistical considerations

Granular morphology:

- If we consider the morphology pattern in a granular fashion, we break it down into percentages, in which case we know those variables will add up to 100. Putting them all into the model violates the non-collinearity assumption.
  - A similar issue arises by breaking down the the Gleason 4 morphology into its sub-patterns.
- How does a reference variable get selected to avoid redundancy?
- It is an ordeal to interpret parameters from these models.

Statistical considerations

PTEN expression within a pattern:

- PTEN expression within a pattern only applies to subjects to have that pattern.
- To illustrate the challenge, consider the following two comparisons:
  - A comparison of outcomes between subjects differing in whether they have PTEN expression within a glomeruloid sub-pattern.
  - A comparison of outcomes between subjects having the glomeruoid sub-pattern, but differing in their PTEN expression within their glomeruloid sub-pattern.
- While a seemingly subtle difference, the questions can lead you to different results and therefore different conclusions regarding the relative role of morphology and PTEN loss in predicting outcomes.

Statistical considerations

Predictors of PTEN:

- One subject can have multiple morphologies. Therefore, a model with PTEN expression as an outcome will need to account for repeated measures in some fashion.

Outline

Topics:

- *Background: Prostate cancer and PTEN expression*
- *Clinical questions*
- *Data set*
- *Statistical challenges and considerations*
- **Descriptive statistics (dos and don'ts)**
- Modeling

Descriptive statistics

Discussion point: Good ideas

- Generate overall Kaplan-Meier (or cumulative incidence, in this case) curves for the whole sample, and possibly stratified by predictors of interest.
- Create organized tables characterizing the absolute and relative frequencies of baseline characteristics.
- Make degree of missingness for each variable of interest clear.
  - Sometimes, the easiest way to do this is to categorize continuous variables, so that you can add a row that denotes the missingness.

Descriptive statistics

Discussion point: Things to avoid

- Stratify descriptive statistics by the indicator of an observed event that is subject to censoring — this tells you *nothing*.
  - Some will try to characterize it as a description of the study sample. This is not a valid argument. The point of describing the study sample is to generalize to some population in a meaningful way.
- Stratify descriptive statistics in a way that flips the ordering of what you've identified as the outcome and the predictor.
  - For example, do not identify PSA as a possible predictor of morphology in your analysis question, and then compare mean PSA between groups defined by their morphology.

Outline

Topics:

- *Background: Prostate cancer and PTEN expression*
- *Clinical questions*
- *Data set*
- *Statistical challenges and considerations*
- *Descriptive statistics (dos and don'ts)*
- **Modeling**

Model

Exercise:

- With all this in mind, propose a brief analysis plan to answer one or more of the study questions. We will reserve some time to share ideas.

Thank you!

Next steps:

- Armed with the materials of this course, you are now prepared to delve more deeply into specific topics, including:
  - Survival analysis.
  - Advanced regression.
  - Data analysis and consulting.
  - Longitudinal data.
- Take a well deserved break, and keep up the good work!