Biostatistical Methods for Clinical Research III

Written Project

This will be a written report of a data analysis, similar to a brief, focused research manuscript but with more statistical detail than usual. Ideally, the project should be an original analysis of data relevant to your own research interests, potentially leading to a submitted manuscript. Address a single substantive issue or a few closely related issues; this **should not be** a comprehensive report of all findings from a study.

The report should:

- carefully define the key issues, questions, or hypotheses to be investigated
- describe the study design and the data that were collected
- explain the statistical methods used and why you chose them, including problems or difficulties and how you handled them
- provide valid and informative summaries of relevant results
- interpret the results appropriately, noting caveats
- discuss substantive conclusions and implications; this will often include noting what is different from the most relevant previous studies and explaining possible reasons, along with implications for further research

Text length should be **less than 2000 words**, with a total of **no more than four figures and/or tables**. Include a title and references, but no abstract. Details of non-statistical methods, such as laboratory or surgical techniques, should be omitted if possible, along with lengthy details of inclusion/exclusion criteria that are not essential for interpreting the results. Limit the background to the minimum necessary for understanding the key issues, avoiding extensive summaries of previous literature.

You will be assigned a mentor from the Biostatistics Division faculty to provide guidance for your project. Please make use of this resource.

Submission process: email electronic copy by 5PM on day of deadline to your advisor.

Project sessions will be scheduled with an advisor (not necessarily the one you have been working with).

Guidelines for Oral Presentation

Prepare a 25-minute talk. Each room will have a laptop available for presentation. It is ideal for you to arrive with your talk on a USB drive, or make arrangements to email it to your project advisor or Olivia DeLeon in advance.

Apportion the time approximately as follows

5 minutes on background
(the scientific question, available data)

5 minutes on statistical methods
(how you analyzed the data and why)

10 minutes on major results
(need not be as comprehensive as in the written report)

5 minutes for questions and reflections

The talk should have a methodological bent. You'll want to highlight any subtle of difficult choices you made in the analysis and explain why you made them. Nearly every project will present some issues that we haven't covered in class. Take the opportunity to explain to fellow students any novel issues you encountered and the techniques you used to address them.

Expectations for Written Project

Major plusses

- Appropriate use of difficult methods, such as multiple random effects, difficult time-varying covariates, cross-validation, propensity scores, or multiple imputation (especially if this makes a discernable difference compared to casewise deletion), etc.
- Unusually insightful interpretation of the statistical results, especially if sensibly synthesizing biological knowledge or results of prior studies.
- Devising non-obvious ways to focus on issues that aren't optimally addressed by typical approaches.
- Recognition of subtle but important possible limitations or biases.
- Exploration of alternative approaches, with excellent assessment of what differences or agreements show.

Minor plusses

- Interpretation of estimates and confidence limits.
- Appropriate use of intermediate methods, such as mixed effects models, ordinal logistic regression, simple time-varying covariates, bootstrapping, etc.
- Checking major assumptions of analyses.
- Appropriate handling of assumption violations.
- Exceptionally clear, informative graphs or figures.
- Correct interpretation of statistical results. For example, correct recognition of relevant confounding, mediation, or lack thereof.
- Sensible variable selection methods for building a multivariate model.
- Major plusses, except that they do not appear to be necessary or make much difference compared to simpler approaches, or they were not explained well.

Tiebreakers (good)

- Interesting question or findings.
- Clear writing.
- Helpful re-scaling of variable(s).
- Providing rationales for variable definitions.
- Minor plusses that have some problems with how they were carried out or explained, but are not completely off-target.

Tiebreakers (bad)

- Poor writing.
- Excessive or inadequate precision.
- Poorly scaled predictors, e.g., effect per year of age, per 1 CD4 cell/mm3.
- Failing to provide units for some results.
- Use of “significant” alone, instead of “statistically significant” or “important”.
- Poor summaries, such as mean +/- SD when data are severely skewed.
- No indication of what summaries are (e.g., mean or median).
- Use of SE when SD would be more appropriate, because purpose is to describe the population.

Minor problems

- Failure to check some important assumptions.
- Indirect analyses or summaries when more direct ones are possible. E.g., giving separate summaries of before and after values without summaries of changes, or separate summaries of rates in two groups without giving odds ratio or risk difference.
- Confidence intervals not provided for some key results.
- Definitions of important variables not clear enough.
- Inadequate details about followup and/or event ascertainment and definition.
- Use of possibly excessive number of predictors without any checking for problems.
- Poor variable selection methods for multivariate modeling, such as reliance only on univariate results.
- Unexplained discrepancies in numbers of subjects.
- Potentially important covariate(s) available but not evaluated, with no explanation for why not.
- Important assertions that are not supported by any results or details. For example, “the

effect of age was linear" or "the model shown in Table 2 fit the best", without any indication of how these issues were assessed and what was found.
- Inaccurate interpretation of key concepts like odds ratio, relative risk, relative hazard, etc.
- No indication of refusal rates and/or amounts of missing data, even though they are clearly present and potentially important. Or, noting substantial amounts of missing data but doing nothing to assess or at least acknowledge its possible impact.
- Weak, indirect phrasing of important results. For example, "we were not able to show X" instead of better descriptions of what was found such as, "our results suggest X but did not reach statistical significance", "our result provide evidence against X but are not conclusive", or "our results provide strong evidence against X".
- Describing "statistical significance" as if it were an actual property in the state of nature rather than a characterization of evidence about the state of nature.
- A major problem, but on a side issue.

Major problems
- A data set and question that does not permit very interesting analysis.
- Exclusive focus only on p-values when estimates and CI's could be obtained and examined.
- Failure to address obvious violations of model assumptions. E.g., using PH regression or logrank test when K-M curves show severe non-proportionality.
- Ignoring dependence in the data, e.g., clustering, or unpaired analysis of paired data.
- Major interpretations that exaggerate how conclusive the results are.
- Interpreting large effects as "no association", "no difference", etc. (usually because p>0.05). (This can instead be a minor problem if the effect in question is of little interest or only for checking model assumptions.)

Severe problems
- Inappropriate use of only cursory statistical analyses. For example, only correlations or t-tests when multivariate modeling is both possible and needed.
- The main questions or issues don't make sense or can't be discerned.
- Conclusions that contradict the results of the analyses.