Organizing Your Approach to a Data Analysis

Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
University of Washington
www.emersonstatistics.com

December 29, 2013

I. Before looking at the data
   A. Identify overall goal of the study
   B. Identify specific aims and how they relate to overall goal
      1. Identify the current state of scientific knowledge
      2. Identify the competing hypotheses that the study is designed to discriminate between
         – (Often dictated by available data)
   C. Refine scientific hypotheses into statistical hypotheses
      1. Identify type of question
         a. Prediction, estimation, or testing
         b. Identifying groups, quantifying distributions, or comparing distributions
      2. Where appropriate, specify statistical hypotheses in terms of a summary measure for the distribution of measurements
         a. Summary measure: e.g., mean, geometric mean, median, proportion or odds above a threshold, event rate, hazard
         b. Contrast across groups: difference, ratios
   D. Consider design of ideal experiment
      1. Ignore practical, ethical limitations in order to be able to later compare how close the actual situation is to the ideal
         a. Who / what would be the subjects?
            i. Inclusion criteria (ideal: eventual target of inference)
            ii. Exclusion criteria (ideal: exclusions necessary of desirable in experimental setting)
         b. What would be the intervention(s)?
         c. How would subjects be assigned to the intervention?
            i. Interventions that are systematically varied
            ii. Interventions that are controlled at a single level
            iii. Observational (convenience) sampling
         d. What would be the variables measured
            i. Outcome variables
            ii. Implementation of interventions
            iii. Additional covariates
   E. Available data
      1. Sampling scheme

   a. Retrospective vs prospective
   b. Observational vs intervention
   c. Inclusion, exclusion criteria
   d. How was sample size determined
      - Overall
      - Within any strata
2. Variables in the data set
   a. Names
   b. Relationship to real world quantities
   c. Conditions under which they were measured
   - e.g., frequency, timing relative to any interventions
   d. Units of measurement (limitations)
      - e.g., qualitative vs quantitative, continuous vs discrete, patterns of missing data
3. Categorization of variables according to meaning
   a. Demographic (age, sex, etc.)
   b. Baseline physiology (SBP, performance status)
   c. Baseline disease risk factors, prognosis
   d. Measures of treatment intervention
   e. Measures of ancillary clinical course during treatment (e.g., ancillary treatments, environmental conditions)
   f. Measures of treatment outcome
4. Categorization of variables according to use in analysis
   a. Response (outcome) variables
   b. Predictor variable of interest (variable identifying groups) (POI)
   c. Variables identifying subgroups to explore effect modification
   d. Potential confounders
      - Causally associated with response variable (in truth) independent of predictor of interest (in groups where POI is held constant)
      - Association with predictor of interest (in the sample)
      - Not in causal pathway of interest
   e. Variables which allow increased precision
      - Variables prognostic of response variable, but not associated with predictor of interest
      - Questions about effects within such groups can be answered with more precision than questions about effects in the larger population (e.g., adjusting for age)
   f. Surrogates for response
      - Variables in the causal pathway of interest
      - Variables measuring a later (or near contemporaneous) effect of the response
   g. Irrelevant

II. Statistical Analysis Plan (SAP defined prior to looking at data)

A. Primary specific aim
    1. Outcome measurement
        a. Clinical definition
        b. Protocol definition of measurement
        c. Statistical summary measure
    2. Predictor(s) of interest (primary grouping variable(s))
    3. Subgroups used for effect modification
        a. Contrasts of summary measures across subgroups
    4. Statistical hypotheses
        a. Superiority (and whether that corresponds to higher or lower summary measures)
        b. Noninferiority (and whether margin corresponds to higher or lower summary measures)
        c. Approximate equivalence
        d. Inferiority (and whether that corresponds to higher or lower summary measures)
        e. Two-sided differences
    5. Statistical burden of proof
        a. Level of significance for hypothesis testing
        b. Descriptive criteria and precision
B. Secondary and exploratory specific aims
    1. These might represent supportive analyses using alternative clinical outcomes
    2. These might represent analyses considering specific mechanisms of action
    3. These might represent analyses restricted to particular subgroups
    4. These might represent safety analyses in RCT
C. Identify analysis populations
    1. Primary efficacy: the cases that will be included to answer your primary question
        a. In superiority RCT this will be per randomization or intent to treat
        b. In noninferiority RCT this might be a "per protocol" analysis of data while on study treatment
    2. Secondary efficacy
        a. In RCT this might be modified intent to treat to exclude some subjects based on pre-randomization variables
        b. In RCT this might be a "per protocol" analysis to assess mechanisms of action
    3. Safety
        a. In RCT this is usually all subjects who have taken any amount of study drug, and includes time up to 30 days after discontinuation
D. Identify statistical analysis model
    1. Univariate estimation
    2. Two sample tests
        a. Variations re variance estimation, permutation tests, exact methods, etc.
        b. Stratified tests

3. Regression methods
    a. Summary measures
    b. Link function
    c. Modeling of predictor of interest (dichotomized, dummy variables, continuous, transformed)
    d. Adjustment for covariates (and how modeled as dichotomized, dummy variables, continuous, transformed)
    e. Modeling of interactions
    f. Regression parameters used to form statistic
    g. Statistic used for estimation, testing (Wald, score, likelihood ratio)

E. Handling of missing data

F. Tables and Figures

1. Description of sampling scheme actually attained
    a. Timeframes and sample sizes
    b. Frequency and timing of measurements
    c. Missing data patterns
        – Censoring distribution
        – Subject / investigator specified reasons for drop out
2. Description of subjects and baseline variables
    a. What
        – Means, standard deviations, minimum, maximum, median, quartiles of continuous variables
        – Frequencies of binary or categorical data (including important scientific categories of continuous variables)
    b. How
    – Prospective cohort studies: Columns corresponding to groups defined by POI
    – Retrospective case-control studies: Columns corresponding to groups defined by outcome
    – Exploratory cross-sectional studies: Either of the above depending whether focus is more on identifying risk factors for outcome or on identifying all outcomes from risk factors
3. Description of outcomes
4. Preliminary estimates and SE related to primary question
5. Tables / figures of inference (regression parameters, estimates, CI, tests)
6. Exploratory analysis results

III. Univariate descriptive statistics

A. Goals

1. Identify errors in the data
    a. Particularly unusual measurements (out of range)
    b. Unusual combinations of measurements
2. Verify your understanding of the measurements

3. Identify patterns of missing data
4. Identify exact population used in study (Materials and Methods)
5. Identify aspects of the data that may present technical statistical issues
    a. Ideal: allows easiest, most precise statistical inference with smaller sample sizes
        - equal information about all groups being investigated (? equal sample sizes)
        - measurements of response within each group distributed symmetrically with no 'long tails' (outliers)
        - no missing data
    b. Potential problems suggesting possibility of problematic scientific interpretation (problems which can not necessarily be solved with the available data)
        - missing data patterns
    c. Potential problems suggesting less generalizable statistical analysis (problems not necessarily indicated by the measures of statistical confidence)
        - 'Outliers' in distribution of grouping variables (predictors): i.e., low sample sizes in some groups that are far away from the rest of the data (e.g., trying to determine an age effect in a sample in which most are between 10 and 20 years old, but one subject is 80)
    d. Potential technical problems suggesting possibility of less precise inference (problems that will tend to lower our reported level of statistical precision)
        - 'Outliers' in distribution of response
        - Too little variation in the distribution of the grouping variables (e.g, trying to determine an age effect from a sample in which everyone is between 20 and 21 years old)
        - Too much association among the different grouping variables (e.g., trying to determine an age effect when all the young subjects are male and all the old subjects are female)
    e. Potential technical problems which suggest we might need to use more complicated statistical methods
        - Repeated measurements on the same sampling unit (correlated response)
        - When comparing means: unequal variability across groups being compared
        - When comparing time to events: lack of proportional hazards
        - When adjusting for covariates: nonlinear effects; interactions

C. Order of investigation
    1. Potential confounders
    2. Predictor of interest
    3. Response

D. Tools
    1. Frequency tables
    2. Mean, median, standard deviation, etc.
    3. Box plots, histograms

IV. Bivariate and trivariate descriptive statistics

A. Goals

1. Identify confounding relationships
    a. Associations between other variables and predictor of interest
    b. Associations between other variables and response
2. Identify important predictors of response
    a. Univariate effects
    b. Effect modification (interactions)
3. Identify surrogates of response
4. Characterize form of functional relationships (linear, etc.)

B. Ideal (because easiest for the statistician)

1. Predictor of interest has no association with any other predictors
2. Only a few variables are markedly associated with response
3. All associations look like a straight line relationship
4. No interactions (effect modification)

C. Order of investigation

1. Relationships among other predictors
2. Relationships between predictor of interest and other predictors
3. Relationships between response and other predictors
4. Relationships between predictor of interest and response overall
5. Relationships between predictor of interest and response within subgroups

D. Tools

1. Contingency tables
2. Stratified means, medians, standard deviations, etc.
3. Stratified box plots, histograms, etc.
4. Scatterplots
5. Stratified scatterplots
6. Correlations

V. Defining a suitable context for modeling

A. Goals

1. Choosing appropriate form for response variables
    a. Selection of measure of response
        – Transformations of available data
    b. Summary measure to use as basis for statistical model
2. Selection of groups to be investigated / compared
        – Form for predictor of interest
        – Identification and form of interactions (effect modification)
        – Identification and form of potential confounders to be modeled
        – Identification and form of precision variables to be modeled
3. Choosing analysis method (type of regression)

B. Methods
    1. Ideal: Statistical model dictated entirely by scientific question (before looking at the data)
    2. Practical: Model building (but may lead to problematic inference)
        a. Educated guess for first models
        b. Fit models
        c. Evaluate validity of necessary assumptions

VI. Model Building to Address Primary Question

A. Goals (in order of importance)
    1. Selection of variables to address scientific questions (main effects and interactions)
    2. Selection of variables to minimize bias (address confounding)
    3. Selection of variables to maximize precision
    4. Selection of models which are easiest to implement (usually: have the least technical requirements on the distribution of response)

B. Methods
    1. Addressing scientific question: Thinking about the problem
    2. Addressing confounding: Adding or removing variables and observing effect on other regression parameters relative to findings in bivariate description of data (many difficult issues here)
    3. Addressing precision: Determining which variables tend to predict response (many difficult issues here)
    4. Evaluate extent to which data meets technical requirements of statistical procedures

VII. Exploratory Analyses for Hypothesis Generation

A. Modeling of exact form of predictor-response relationship (e.g., dose-response)

B. Identification of other predictors of response

C. Subgroup analyses: Compare effect of predictor of interest on response within subgroups (effect modification)

VIII. Reporting Results and Interpretation

A. Scientific Background and Hypotheses

B. Materials and Methods
    1. Sampling scheme
    2. Most basic descriptive statistics

C. Results (more objective first)
    1. Descriptive statistics
    2. Results of analyses about primary question
        a. Estimates of effect
            – Point estimates (single best estimate)
            - Interval estimates (range of estimates indicating precision)
        b. Decisions about hypotheses
            – Binary decision (yes or no)
            – Measure of statistical confidence in precision

3. Results of analyses about prespecified secondary questions or questions which demonstrate consistency (or lack of same) across alternative approaches
4. Results of analyses about questions that arose during analysis and that the vast majority of readers would agree could and should be answered by the data

D. Discussion (subjective, including particularly data-driven analyses)

1. Elaboration on ways that these analyses address the overall goal of the study
2. Results of the most speculative analyses of the data