Draft syllabus for BIOS 669, Working with Data in a Public Health Research Setting, spring 2014

669 WORKING WITH DATA IN A PUBLIC HEALTH RESEARCH SETTING (3). Prerequisite, BIOS 511, EPID 700, or permission of the instructor.

This course provides a conceptual foundation and practical training to students who will be working with data from clinical trials or other public health research studies. Topics include data issues in study design, collecting high quality data, using SAS and SQL to transform data into structures useful for analysis, producing typical reports, data closure and export, and working with big data. Spring.

This course is about data management, not study management, although how a study is set up and managed will affect its ability to do good data management. Its general target audience is people who will be statisticians, project managers, and analysts, and its specific target audience is Masters level biostatistics students in their second semester of graduate study. An ever-present theme will be the necessity for documentation. Compared to BIOS 511, it will be about applications, not tools.

BIOS 669 units (approximate # classes):

1. Course intro (1.5)
    a. What course is, what it is not
    b. Types of data used in course:
        i. Clinical trial
        ii. Observational study (unweighted)
        iii. Observational study (weighted)
        iv. Data not collected for research purposes
    c. Study organization/personnel roles/need for documentation
    d. Introductions to the METS study (i) and LONGSCAN study (ii) used for course examples; data type iii will be used in an NHANES exercise, and data type iv will involve data downloaded from the internet, possibly hospital quality data from Medicare (see a later unit)
2. Study start-up (2.5)
    a. Recruitment, screening, and enrollment
    b. Randomization plans (for clinical trials)
    c. Blinding – establishing and maintaining (for clinical trials)
    d. IDs – designing, creating, assigning, tracking
    e. Pilot studies
3. Data collection systems (3)
    a. Selection of an appropriate system for the study type, scale, setting, etc.
    b. Implementation issues
    c. Form design, including internationalization/translation issues and the importance of complete up-front edit specification within and between forms
    d. Data security
4. Data quality (3)

    a. Minimizing data transitions
    b. Data cleaning – detection, correction, documentation
    c. Data collection and structures to support data quality analyses (special IDs, replication, data splitting, etc.)
5. SQL as a data management and reporting tool (2)
6. Reshaping data for analytical purposes: Analysis data sets and variables (8)
    a. Combining data sets safely and correctly – key variables, renaming, aggregation, etc.
    b. Data structures needed for various reports and analyses – visualizing and achieving
        i. Types of counts and percentages, indicators, differences, etc.
        ii. Case-cohort, time to event (with censoring), interval censoring and observation times, etc.
    c. Designing, creating, and checking complex derived variables
        i. Scoring rules in the presence of missing values
        ii. Using coding systems such as ICD10 codes and WHO drug dictionary
        iii. Creating adverse event indicator variables by preferred term and MedDRA SOC (System Organ Class)
        iv. Definition of treatment-emergent adverse events, endpoints, and other critical variables based on a study's statistical analysis plan
    d. Naming conventions and storage schemes, especially for long-term studies
    e. Codebooks
7. Reports (3)
    a. Types, timing, design, production, checking
    b. Some standard tables – break into pieces, structure code to produce pieces, combine pieces to produce report data set, present data set
    c. Large reports such as DSMB and OSMB
    d. Include graphs if time
8. Data closure (2)
    a. Preparing limited access data sets
    b. Conforming to external standards for facilitated data sharing and reuse
    c. Surfacing data to end users through an application
9. Big data – concepts and practical tips (3)
    a. CMS data
    b. Genomics/SNP data
    c. Electronic health records
10. Pharma-related topics (1)
    a. CDISC as an external standard
    b. Clinical study reports and how they are written in industry for Phase I-IV studies: complete analysis plan before unblinding, table shells/mock before programming
11. Ethical issues – data integrity as the basis of sound research, avoiding "studies gone wrong" (1; include course wrap-up)

Have resources available (notes, references, videos) but do not cover in class:

1. SAS refresher (including exercises)
2. Data conversion
    a. Database to SAS data
    b. XLS to SAS data set
    c. SAS data set to XML
    d. XML to SAS data set
    e. Complex and/or large text files
3. PROC COMPARE
4. PROC TRANSPOSE
5. PROC APPEND
6. PROC PLAN
7. PROC REPORT
8. Metadata concepts and examples (SAS dictionary tables, status bytes, ODM XML)
9. Interpolation/extrapolation/imputation
10. JMP as a tool for data cleaning/exploration/understanding
11. Tips for working on a collaborative manuscript
12. Applied simulation (e.g. trial results)
13. Fuzzy matching
14. Implementing look-up tables in SAS
    f. Match merge
    g. Hash table
    h. Format with PUT
    i. SQL

Sample data for course notes and exercises: METS (trial), LONGSCAN (observational study)

Have one NHANES exercise for experience downloading, converting, and merging data, and using the provided instructions to do a weighted analysis.