BIOS:6610 – STATISTICAL METHODS IN CLINICAL TRIALS

Data Sharing / Standardization

Christopher S. Coffey
Department of Biostatistics
University of Iowa
Data Sharing:

“Responsible entity making data available via open or restricted access, and exchanged among parties”
Increased Data Sharing could facilitate:

- Independent re-analyses of trial results
- Addressing concerns about publication bias
- Characterizing trial outcomes by subgroups
- Considering additional questions beyond original trial hypotheses
- Carrying out meta-analyses for systematic review
- Facilitating hypothesis generation & areas that may warrant further exploration
DATA SHARING

But, data sharing can be a volatile topic

- **IOM (2013):** “A cultural change has occurred in which the conversation around data sharing has moved from whether it should happen to how it can be carried out”

A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group’s data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as “research parasites.”
Types of Shared Data:

- **Raw Data**
  - Observations about individual participants

- **Analysis Data Sets**
  - Data generated to serve as basis for later analyses
  - “Derived” variables
Potential Recipients of Shared Data:

- Researchers
- Study Participants
- Patient Community / Advocacy Groups
- Educators
- Funding Agencies
- Prospective Plaintiff / Attorneys for pending litigation
DATA SHARING

Investigator Controlled:

- Data sharing usually conducted at summary level
  - Peer-reviewed journal publications
  - Clinical trial registration sites
- Study team retains strict access to all data files
- Select study data may be made to individual researchers on a case by case basis
  - Access to data tightly controlled
Controlled Access Data Sharing:

- Data made publicly available, but access restricted to specific classes of users for specific purposes.
- Requires formal request/review to access data, providing clear rationale for proposed use.
- User must sign legally binding data sharing agreement before receiving data.
  - Include clauses to prohibit further sharing of de-identified/anonymized data & from attempting to identify individuals from data.
- Transferred via a secure method.
Open Access Data Sharing:

- Data made available at a defined time to any party seeking to use it
- Files completely open for access by anyone
- Requires no formal agreement to access/use data
DATA SHARING

Full Open ("Real Time") Data Sharing:

- Data made available as study is ongoing
- Files completely open for access by anyone
- May involve controlled or open access
DATA SHARING

Important to Balance:

➢ Patient Privacy

➢ Scientific Value of the Data
  • If personal data removed/de-identified and subject code identifiers cannot be linked back to specific individuals, then data are no longer considered to be personal data

Even simple methods can reduce utility of data.

Particularly problematic in some situations (e.g. rare adverse events), where rare events may be most informative scientific information!
Expert Determination Method:

- Involves a ‘person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable’:
  - Applying such principles and methods to determine risk is very small that information could be used alone or in conjunction with other available information by anticipated recipient to identify subjects
  - Documents methods and results of analyses that justify such determination
Complete De-Identification Possible?:

- 1997: Using known birth date, gender, and zip code, a computer expert able to identify records of Gov. William Weld from an allegedly anonymous database of Massachusetts state employee health insurance claims (Barth-Jones, 2015)

- 2013: Researchers able to identify “de-identified” male genomes through correlations with commercial genealogy databases (Science, 2013)
PPMI Experience – Full Open Access:

- **PPMI Sites**
  - **Clinical Core**
  - **Imaging Core**
  - **Biorepository**
  - **Bioinformatics Core**
  - **Bioanalytics Core**
  - **Genetics Core**

- **PPMI Steering Committee**

- **MJFF - Michael J. Fox Foundation for PD Research**

- **ISAB - Industry Scientific Advisory Board**

- **EAC - External Advisory Committee**

- **CSOC - Clinical Study Oversight Committee**

- **Statistics Core**

- **Biospecimen access**
FULL OPEN ACCESS

PPMI Experience – Full Open Access:

- Data Use Application
- Data Use Application Review
- Data Use Authorization
- Access Data
PPMI Experience – Lessons Learned:

- Privacy Concerns vs. Analysis Concerns
  - Need to restrict dates (possibly identifiable information)
  - Can prevent certain analyses from being conducted

- Need flexibility to add/modify database on the fly – Mechanism to upload ‘derived’ variables

- Need for database to be constantly informed by and interacting with integrated study team

- Need a publications/data sharing committee to keep track of who is accessing data, and what it is being used for
Data Sharing / Standardization - Hurdles:

Requires two very different sets of expertise:

- Data Coordinating Centers for multi-site trials
  - Tend to very good, but very “vanilla”
  - Main objective is to provide support to overall goals of study

- Bioinformatics focused groups
  - Tend to be a bit more ‘cutting edge’
  - Work more closely with the data – ‘data mining’

In general, groups will fall into one of these two categories – one not necessarily better than the other
Data Sharing / Standardization - Hurdles:

Few groups with expertise to address both issues.

- Experience suggests more likely to find groups suited for one type of data but not the other

- How to handle this?
  - Best solution might be to use multiple groups with different expertise → synergy
  - Need to keep an eye on end user
Data Sharing / Standardization - Hurdles:

- Regardless of complexity, a dataset is only helpful if it is useful.
- There is a misconception that public datasets can generally be obtained and immediately analyzed.
- In general, this is far from the truth.
- A data management resource to merge and create ‘analysis files’ to external users would save a great deal of time – And speed up reporting of important findings.
**SUMMARY**

Data Challenges:

---

**SPRINT Data Analysis Challenge**

**Nejm**

To explore the potential of clinical trial data sharing, the New England Journal of Medicine (NEJM) is hosting a challenge: use the data underlying a recent NEJM article to identify a novel clinical finding that advances medical science.

**SPRINT Data Analysis Challenge**

**New Available. Free On Demand Recording of the Web Event Sessions**

Browse the session catalog at events.nejm.org to view the recorded sessions and feel free to share the link with any colleagues or peers who would find the discussion of relevance.

**Thank you to all SPRINT Challenge participants. We appreciate your interest and the effort that went into each entry.**

**We are pleased to announce the SPRINT Challenge winners.**

View the top three entries and read the novel findings that were identified using the SPRINT trial data. >>Read more<<

Browse all the SPRINT Challenge submissions. >>Read more<<

**Thank you to our collaborators: the SPRINT Trial investigators and NHLBI.**

We also want to thank the SPRINT trial investigators who worked diligently to release the data underlying the NEJM article one year in advance and the staff at NHLBI’s Bioclincc dispository for all of their collaborative support.

**Let’s continue this important conversation.**

Join us for the free web event Aligning incentives for Sharing Clinical Trial Data on April 3–4, 2017. View the SPRINT Challenge winners’ presentations and hear what was learned from the SPRINT Challenge directly from the SPRINT PIs, data analysts, patient participants, and the NHLBI repository. >>Learn more<<

---

**2016 PPMI DATA CHALLENGE**

The Parkinson’s Progression Markers Initiative (PPMI) has generated a comprehensive, standardized, longitudinal set of clinical, biological and imaging data unique to the Parkinson’s disease (PD) field and ripe for novel and innovative exploration. In mid-2016, study sponsor The Michael J. Fox Foundation (MJFF) cast an open call to computational scientists, data scientists and neuroscientists to analyze PPMI data toward new insights into PD diagnosis and progression.

PPMI data is highly variable, with age of onset, rate of progression, and type and severity of symptoms different across the 5,000 worldwide living with the disease. Identifying models for prognosis and sub-typing would aid in subject selection for clinical studies and design of trials toward novel therapies.

PPMI data is uploaded in real time and accessible to qualified researchers. Learn more and download the data by visiting www.ppmi-info.org.

**WINNING SUBMISSIONS**

Read more on the analysis and findings of the winners who answered:

1. What factors at baseline predict clinical progression?
2. What are the sub-types of Parkinson’s disease?

**FORMAT**

The 2016 PPMI Data Challenge offered $50,000 in total prizes, awarding a $25,000 cash prize to the most comprehensive and applicable answer to each question. GE Healthcare, a PPMI industry partner, and MJFF each provided $25,000 for the prizes.

In an effort to assist analysts who may be unfamiliar with Parkinson’s disease, MJFF established a team of subject matter experts to collaborate with applicants to assist in understanding the rationale for the challenge questions and the nuances of the data set components.

Consistent with the spirit of PPMI, any algorithms, inventions or discoveries made using the PPMI database are subject to the PPMI Data Use Agreement. We expect these algorithms to be shared with MJFF and the research community. Organizations with existing background IP may maintain that IP while using the PPMI dataset as a tool for validation.