Gatlinburg 2019 Statistical Methods Preconference: Reproducible Science Through Strong Data Management

Aaron Kaat, Ph.D.
Aaron.Kaat@Northwestern.edu
Northwestern University

Cristan Farmer, Ph.D.
Cristan.Farmer@NIH.gov
National Institute of Mental Health
Disclosures

Dr. Farmer is an employee of the Federal Government. The views expressed here are her own and do not necessarily represent the views of the NIH or the US Government. She has no disclosures.
Outline

• Introductions
• Open science
• Research data in the 21st century
• Fundamentals of good data management practice
• Fundamentals of good analysis (methods) management practice
• Example database: NICHD Clinical Trials Database
• Questions and discussion
Reproducibility and open science
## Definitions

<table>
<thead>
<tr>
<th>Same Data</th>
<th>Reproducibility</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar Data</td>
<td>Replicability</td>
<td>Generalizability</td>
</tr>
</tbody>
</table>

Different groups use different terms. Our presentation will use these definitions, but they are not consistent across all publications.

In a perfect world, science is self-correcting.

But given the file drawer problem, how do we know that our accumulated knowledge is not just Type I error?
The replication crisis

~1% of published (psychology) studies are replications

Fig. 1. Replication rate in the top 100 psychology journals. The solid line represents the percentage of publications (from 100 journals with the highest 2010 5-year impact factor) that used the term “replicat.” The dashed line reports the replication rate based on the percentage of articles using the term “replicat” that were actual replications. The bars represent the total number of articles published in that decade. The 2010s bar is truncated because data from only 2.5 years of the current decade were available.

Questionable research practices inhibit reproducibility

Fig. 1. Results of the Bayesian-truth-serum condition in the main study. For each of the 10 items, the graph shows the self-admission rate, prevalence estimate, prevalence estimate derived from the admission estimate (i.e., self-admission rate/admission estimate), and geometric mean of these three percentages (numbers above the bars). See Table 1 for the complete text of the items.

Johh, Loewenstein, & Prelec (2012) Psychological Science
Open science promotes reproducibility

**Code & Process Sharing**
- Prevents questionable research practices (intentional or otherwise)
- Requires good methods management
- Shouldn’t sub for thorough reporting

**Data Sharing**
- Allows for determination of errors, and for sensitivity analyses
- Most effective if combined with methods sharing
- Requires good data management

**Registration**
- Reduces HARKing and other questionable research practices
- Prevents file-drawer problem
- Requires clear plan!

---

*Note: Even if you do not formally register your project, it is useful to “register” with yourself beforehand (i.e., write out your exact hypotheses, data requirements, exact analytic plan beforehand and STICK TO IT!)*
Barriers to open science

Logistical
• Historical
  • Data not in electronic format
  • Computing expensive/not powerful
• Who checks all of this?
• No funding for efforts related to sharing/archiving
• Ethical considerations

Psychological
• Historically, “team science” not the norm
• Fear of:
  • Misinterpretation
  • Misuse of data
  • Getting scooped
  • Getting caught (intentional or otherwise)
Research Data in the 21st Century
Who owns research data?

(not just you!)
Journals and funding agencies now recognize that data (and methods) sharing promotes good science

• NIH requires data sharing if >$500K/year, or if ASD
• NSF requires that data be deposited in a repository
• Most journals require an explicit statement of whether/how data will be shared (RIDD, RASD, Neurology) or author guidelines encourage sharing (JADD, Autism Research, JACP, JCCP, JCCAP)
  • AJIDD requires that authors share upon reasonable request the data on which any conclusions are based
  • Policy not described: Autism, JIDR, J Developmental & Behavioral Pediatrics
Where can you store (and share) research data?

- NIMH Data Archive: data-archive.nimh.nih.gov
- The Dataverse Project: dataverse.org
- Mendeley Data (repository): data.Mendeley.com
- “Upon reasonable request” (such a pain!!)

- At the end of a project: prepare a “shareable” version of your data and syntax
Where can you access shared data?

- Data-Archive.NIMH.NIH.gov
- NDAR (autism)
- NDCT (clinical trials)
- RDoCdb (RDoC)
- ABCD (adolescent brain development)
- SFARI.org
- Childes.talkbank.org
# The Dublin Core

<table>
<thead>
<tr>
<th>Element #</th>
<th>Term</th>
<th>Element #</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Contributor</td>
<td>9</td>
<td>Publisher</td>
</tr>
<tr>
<td>2</td>
<td>Coverage</td>
<td>10</td>
<td>Relation</td>
</tr>
<tr>
<td>3</td>
<td>Creator</td>
<td>11</td>
<td>Rights</td>
</tr>
<tr>
<td>4</td>
<td>Date</td>
<td>12</td>
<td>Source</td>
</tr>
<tr>
<td>5</td>
<td>Description</td>
<td>13</td>
<td>Subject</td>
</tr>
<tr>
<td>6</td>
<td>Format</td>
<td>14</td>
<td>Title</td>
</tr>
<tr>
<td>7</td>
<td>Identifier</td>
<td>15</td>
<td>Type</td>
</tr>
<tr>
<td>8</td>
<td>Language</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://dublincore.org/documents/dces/
**CDISC Metadata Standards**

- **Domain Dataset Elements**
  - Dataset Name
  - Description
  - Location
  - Structure
  - Purpose
  - Key fields

- **Domain Variables Elements**
  - Variable Name
  - Variable Label
  - Type

- **Decodes/Formats**

- **Origin**
  - Current domain
  - Source Variable (raw data)
  - Derived Variable (i.e. scoring)
  - External domain

- **Role**
  - Key Variables
  - Selection Variables
  - Review Variables
  - Support Variables

- **Comments**
Data Standards

https://www.go-fair.org/

• Findable
• Accessible
• Interoperable
• Reusable
FAIR - Findable

• F1 - (Meta)Data are assigned a globally-unique and persistent identified
• F2 - Data are described with rich metadata
• F3 - Metadata clearly and explicitly include the identifier of the data it describes
• F4 - (Meta)Data are registered or indexed in a searchable resource
FAIR - Accessible

• A1 - (Meta)Data are retrievable by their identifier using a standardized communications protocol
  • A1.1 - The protocol is open, free, and universally implementable
  • A1.2 - The protocol allows for an authentication and authorization procedure, where necessary

• A2 - Metadata are accessible, even when the data are no longer available
FAIR - Interoperable

- I1 - (Meta)Data use a formal, accessible, shared, and broadly applicable language for knowledge representation
- I2 - (Meta)Data use vocabularies that follow FAIR principles
- I3 - (Meta)Data include qualified references to other (meta)data
FAIR - Reusable

• R1 - Metadata are richly described with a plurality of accurate and relevant attributes
  • R1.1 - (Meta)Data are released with a clear and accessible data usage license
  • R1.2 - (Meta)Data are associated with detailed provenance
  • R1.3 - (Meta)Data meet domain-relevant community standards
When do you start thinking about data?
Creating a Data Management Plan

• Some grants require a data management plan (DMP)
• There are templates for DMPs online. A great resource is the Inter-University Consortium for Political and Social Research (ICPSR)
• At the most basic level, a DMP must include:
  • What is being collected, and how?
  • Who is responsible for data management?
  • How are data being validated?
  • How are data being manipulated prior to use (i.e., scoring)?
  • Metadata
  • Legal considerations and Intellectual Property
  • Data Security
Benefits of a DMP

• Reduce redundancy in within-wave data collection
• Reduce burden across assessment waves
• Prevents staff objections (“that’s not my job!”)
• Forces you to think about what analyses will be possible and how they map onto your hypotheses.
• Ethical benefits
  • Legal considerations and intellectual property
  • Data security
Data Capture, Storage, and Analysis

- **Data Capture**
  - Examiner-completed clinical report forms (CRFs)
  - Electronic data capture, like REDCap
  - Test protocols that are manually entered into scoring software

- **Data Storage**
  - Hold onto the data while the study is ongoing.
  - Many people choose to do this in one or more spreadsheets
  - Large studies will use a database program, likely with a SQL structure

- **Data Analysis**
  - Everything you want to analyze in an analyzable format
Basics of Databases

• SQL Terms for Databases
  • Rows - data sets representing SINGLE items
  • Columns - labeled element of the row (also called an attribute)
  • Table - any set of rows sharing THE SAME attributes
  • View - any set of rows, potentially combined from multiple tables, formed in response to a query.

• Queries on a Table may be useful for data quality assurance
• Statistical analyses will likely occur on a View
Tables, Tables Everywhere

• Most tables within a relational database are partially or fully normalized—that is, they contain no redundant information with other tables in the database.

• Views from a database are often non-normalized (or de-normalized)—that is, they may contain redundant information on multiple rows.

Transitive dependence: CITY + STATE = ZIP
## Non-Normalized Views - Wide Format

<table>
<thead>
<tr>
<th>PIN</th>
<th>FirstName</th>
<th>LastName</th>
<th>DOB</th>
<th>Sex</th>
<th>Age</th>
<th>IQ_Test_Name</th>
<th>IQ</th>
<th>Age</th>
<th>IQ_Test_Name</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
<td>Jones</td>
<td>1/21/1999</td>
<td>M</td>
<td>18.8</td>
<td>StanfordBinet_5</td>
<td>84</td>
<td>19.9</td>
<td>StanfordBinet_5</td>
<td>82</td>
</tr>
<tr>
<td>2</td>
<td>Jane</td>
<td>Smith</td>
<td>6/4/2006</td>
<td>F</td>
<td>11.4</td>
<td>WISC_IV</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Karen</td>
<td>Doe</td>
<td>3/18/2002</td>
<td>F</td>
<td>15.6</td>
<td>WISC_IV</td>
<td>97</td>
<td>16.7</td>
<td>WISC_V</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>Bob</td>
<td>Doe</td>
<td>12/20/2001</td>
<td>M</td>
<td>17.0</td>
<td>WAIS_IV</td>
<td>71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Non-Normalized Views - Long Format

<table>
<thead>
<tr>
<th>PIN</th>
<th>Visit_Num</th>
<th>FirstName</th>
<th>LastName</th>
<th>DOB</th>
<th>Sex</th>
<th>AgeAtObs</th>
<th>IQ_Test_Name</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Joe</td>
<td>Jones</td>
<td>1/21/1999</td>
<td>M</td>
<td>18.8</td>
<td>StanfordBinet_5</td>
<td>84</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Joe</td>
<td>Jones</td>
<td>1/21/1999</td>
<td>M</td>
<td>19.9</td>
<td>StanfordBinet_5</td>
<td>82</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Jane</td>
<td>Smith</td>
<td>6/4/2006</td>
<td>F</td>
<td>11.4</td>
<td>WISC_IV</td>
<td>89</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Karen</td>
<td>Doe</td>
<td>3/18/2002</td>
<td>F</td>
<td>15.6</td>
<td>WISC_IV</td>
<td>97</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Karen</td>
<td>Doe</td>
<td>3/18/2002</td>
<td>F</td>
<td>16.7</td>
<td>WISC_V</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Bob</td>
<td>Doe</td>
<td>12/20/2001</td>
<td>M</td>
<td>17.0</td>
<td>WAIS_IV</td>
<td>71</td>
</tr>
</tbody>
</table>
# Tables, Tables Everywhere

<table>
<thead>
<tr>
<th>PIN</th>
<th>FirstName</th>
<th>LastName</th>
<th>DOB</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
<td>Jones</td>
<td>1/21/1999</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>Jane</td>
<td>Smith</td>
<td>6/4/2006</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>Karen</td>
<td>Doe</td>
<td>3/18/2002</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>Bob</td>
<td>Doe</td>
<td>12/20/2001</td>
<td>M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs_key</th>
<th>PIN</th>
<th>AgeAtObs</th>
<th>IQ_Test</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>18.8</td>
<td>4</td>
<td>84</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>19.9</td>
<td>4</td>
<td>82</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>11.4</td>
<td>2</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15.6</td>
<td>2</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>16.7</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>17.0</td>
<td>1</td>
<td>71</td>
</tr>
</tbody>
</table>
Matches

- One-to-One
- One-to-Many
- Many-to-Many (a no-no in SQL-land)
  - We might have a table with copyright information related to a specific questionnaires (good data provenance)
  - We might have another table that gives information about each assessment
  - Each assessment might have multiple questionnaires, so we need an intersection table to make that work
Keys

• Primary Key - unique identifier for each row in a table

• Foreign Key - a column with identifiers that refer to unique rows in other tables, but may be repeated within this table

• Alternative Key - if there is not a Primary Key in a table, an alternative key can be made through the combination of multiple columns
Going Back to Our Examples:

<table>
<thead>
<tr>
<th>PIN</th>
<th>FirstName</th>
<th>LastName</th>
<th>DOB</th>
<th>Sex</th>
<th>AgeAtObs</th>
<th>IQ_Test_Name</th>
<th>IQ</th>
<th>Visit 2</th>
<th>IQ_Test_Name</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
<td>Jones</td>
<td>1/21/1999</td>
<td>M</td>
<td>18.8</td>
<td>StanfordBinet_5</td>
<td>84</td>
<td>19.9</td>
<td>StanfordBinet_5</td>
<td>82</td>
</tr>
<tr>
<td>2</td>
<td>Jane</td>
<td>Smith</td>
<td>6/4/2006</td>
<td>F</td>
<td>11.4</td>
<td>WISC_IV</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Karen</td>
<td>Doe</td>
<td>3/18/2002</td>
<td>F</td>
<td>15.6</td>
<td>WISC_IV</td>
<td>97</td>
<td>16.7</td>
<td>WISC_V</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>Bob</td>
<td>Doe</td>
<td>12/20/2001</td>
<td>M</td>
<td>17.0</td>
<td>WAIS_IV</td>
<td>71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PIN</th>
<th>Visit_Num</th>
<th>FirstName</th>
<th>LastName</th>
<th>DOB</th>
<th>Sex</th>
<th>AgeAtObs</th>
<th>IQ_Test_Name</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Joe</td>
<td>Jones</td>
<td>1/21/1999</td>
<td>M</td>
<td>18.8</td>
<td>StanfordBinet_5</td>
<td>84</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Joe</td>
<td>Jones</td>
<td>1/21/1999</td>
<td>M</td>
<td>19.9</td>
<td>StanfordBinet_5</td>
<td>82</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Jane</td>
<td>Smith</td>
<td>6/4/2006</td>
<td>F</td>
<td>11.4</td>
<td>WISC_IV</td>
<td>89</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Karen</td>
<td>Doe</td>
<td>3/18/2002</td>
<td>F</td>
<td>15.6</td>
<td>WISC_IV</td>
<td>97</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Karen</td>
<td>Doe</td>
<td>3/18/2002</td>
<td>F</td>
<td>16.7</td>
<td>WISC_V</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>Bob</td>
<td>Doe</td>
<td>12/20/2001</td>
<td>M</td>
<td>17.0</td>
<td>WAIS_IV</td>
<td>71</td>
</tr>
</tbody>
</table>
## Going Back to Our Examples

<table>
<thead>
<tr>
<th>PIN</th>
<th>FirstName</th>
<th>LastName</th>
<th>DOB</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
<td>Jones</td>
<td>1/21/1999</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>Jane</td>
<td>Smith</td>
<td>6/4/2006</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>Karen</td>
<td>Doe</td>
<td>3/18/2002</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>Bob</td>
<td>Doe</td>
<td>12/20/2001</td>
<td>M</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IQ_Test</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>StanfordBinet_5</td>
</tr>
<tr>
<td>2</td>
<td>WISC_IV</td>
</tr>
<tr>
<td>3</td>
<td>WISC_V</td>
</tr>
<tr>
<td>4</td>
<td>WAIS_IV</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs_key</th>
<th>PIN</th>
<th>AgeAtObs</th>
<th>IQ_Test</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>18.8</td>
<td>4</td>
<td>84</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>19.9</td>
<td>4</td>
<td>82</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>11.4</td>
<td>2</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>15.6</td>
<td>2</td>
<td>97</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>16.7</td>
<td>3</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>17.0</td>
<td>1</td>
<td>71</td>
</tr>
</tbody>
</table>
How crazy can this get?

ETL Conventions for use with PEDSnet CDM v2.9 OMOP V5.2

The PEDSnet Common Data Model is an evolving specification, based in structure on the OMOP Common Data Model, but expanded to a PCORnet Common Data Model and the primary research cohorts established in PEDSnet.

Version 2.9 of the PEDSnet CDM reflects the ETL processes developed after several iterations of network development. As such, it proposes CDM.

This document provides the ETL processing assumptions and conventions developed by the PEDSnet data partners that should be used by business rules. This document will be modified as new situations are identified, incorrect business rules are identified and replaced, as new rules, and as the PEDSnet CDM continues to evolve.

Comments on this specification and ETL rules are welcome. Please send email to pedsnetdcc@email.chop.edu, or contact the PEDSnet pr http://www.pedsnet.info).

PEDSnet Data Standards and Interoperability Policies:

1. The PEDSnet data network will store data using structures compatible with the PEDSnet Common Data Model (PCDM).
2. The PEDSnet CDM v2.9 is based on the Observational Medical Outcomes Partnership (OMOP) data model, version 5.2.
3. A subset of data elements in the PCDM will be identified as principal data elements (PDEs). The PDEs will be used for population-level will be marked as Optional (ETL at site discretion) or Non-PDE (ETL required, but data need not be transmitted to DCC), and will not be site.
4. It is anticipated that PEDSnet institutions will make a good faith attempt to obtain as many of the data elements not marked as Optional
5. The data elements classified as PDEs and those included in the PCDM will be approved by the PEDSnet Executive Committee (compr principal investigator).
6. Concept IDs are taken from OMOP 5 vocabularies for PEDSnet CDM v2.9, using the complete (restricted) version that includes license...
Views for Analysis

• Databases are great for data storage, but they are not very good for data analysis.

• Instead we want to request and export a view from our database and analyze that.
  • Statistically manipulating an export is better than manipulating the database, insofar as you don’t want to change any of the original data.
  • Lock the analyzed dataset so you can always go back to it and don’t have to try to recreate what it would have looked like from the database.
Views for Analyses

<table>
<thead>
<tr>
<th>PIN</th>
<th>FirstName</th>
<th>LastName</th>
<th>DOB</th>
<th>Sex</th>
<th>AgeAtObs</th>
<th>IQ_Test_Name</th>
<th>IQ</th>
<th>AgeAtObs</th>
<th>IQ_Test_Name</th>
<th>IQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Joe</td>
<td>Jones</td>
<td>1/21/1999</td>
<td>M</td>
<td>18.8</td>
<td>StanfordBinet_5</td>
<td>84</td>
<td>19.9</td>
<td>StanfordBinet_5</td>
<td>82</td>
</tr>
<tr>
<td>2</td>
<td>Jane</td>
<td>Smith</td>
<td>6/4/2006</td>
<td>F</td>
<td>11.4</td>
<td>WISC_IV</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Karen</td>
<td>Doe</td>
<td>3/18/2002</td>
<td>F</td>
<td>15.6</td>
<td>WISC_IV</td>
<td>97</td>
<td>16.7</td>
<td>WISC_V</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>Bob</td>
<td>Doe</td>
<td>12/20/2001</td>
<td>M</td>
<td>17.0</td>
<td>WAIS_IV</td>
<td>71</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Quality Assurance

• Data Entry
  • Automated (e.g. electronic data capture with direct data transfer)
  • Manual

• Within-database QA
  • Put in ranges for admissible scores. This will flag invalid data points.
  • Put in a value (or values) for missing data and flag it as a missing code.
  • Some databases will include the option to recalculate derived variables to ensure they are accurate beyond just range checks.
  • Other options may be available depending on the database you use

• Outside QA
  • Post-processing of cross-table data
Fundamentals of good analysis (methods) management practice
When do you start thinking about analyses?
The Earlier the Better

Before Data
- Research Question
- Testable Hypotheses
- Data Collection Plan
- Analysis Plan

Getting Data
- Data Cleaning & QA
- Interim Analyses

After Data
- Final Analyses
Pre-Registration

• Pre-registration means knowing and stating what you are going to do in advance
  • Number and nature of subjects, stimulus materials, procedures, measures, rules for excluding data, plans for data analysis, predictions/hypotheses, etc.
  • Posted in a time-stamped, locked file in an online repository that can be accessed by editors and reviewers (and, ultimately, by readers)

• Clinical trials need to pre-register their analysis plans
• Psychological studies are starting to pre-register more frequently (e.g. through the Open Science Framework [https://osf.io/k5wns/](https://osf.io/k5wns/))
Locked Datasets

• We already talked about locking Views for analysis
• This is important even for interim analyses so you can always reproduce your own results
• New data may allow you to replicate your own findings
• Alternatively, you may archive your locked interim analyses dataset and maintain a cumulative dataset for final analyses.

<table>
<thead>
<tr>
<th>Same Data</th>
<th>Reproducibility</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar Data</td>
<td>Replicability</td>
<td>Generalizability</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Same Methods</th>
<th>Alternative Method(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reproducibility</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>Replicability</td>
<td>Generalizability</td>
</tr>
</tbody>
</table>
Statistical Analysis Plans
Annotate Everything

Why would you annotate statistical methods?

• For Others:
  • Journal requirements or other research teams
  • Sharing analyses within a lab
  • How derived variables were calculated (and why)

• For Yourself:
  • How derived variables were calculated (and why)
  • Interpretation of interim results that are then used as inputs to additional analyses
How do you annotate?

• Different statistical programs have different annotation methods

• Syntax files are easier to annotate than point-and-click methods

• **A detour**: R Markdown
How can you annotate other things?

/* Long comments (multi-line) in SAS, Stata, and C++ can be formatted using asterisks and slashes */

* One-line short SAS or Stata comments use just an asterisk

// One-line comments in Stata or C++ use double slashes

# R and Python comments follow a pound symbol
Can you annotate SPSS?

• SPSS syntax editor allows you to enter comments in 3 ways:
  • Use the COMMENT command ending with a period.
  • Start a line with a single asterisk *
  • Use the asterisk-slash method (one-line only though) with /* and */

• You can’t put comments in point-and-click SPSS

• You can “paste” point-and-click into a syntax file and add comments!
Example Database: NICHD Clinical Trials Database
Neurodevelopmental and Behavioral Phenotyping Service (NIMH Intramural Research Program, PI: Audrey Thurm)

- Neurodevelopmental evaluations for ~15 collaborations with various NIH institutes (NIMH, NICHD, NHGRI...)

- Large volume of data
  - Many types of assessments (questionnaires, standardized tests, computerized assessments)
  - Many types of protocols (different batteries for different populations)
  - Usually longitudinal assessment
### DATA MANAGEMENT PLAN

**Parties involved:** PI, Data manager, Stats  
**Responsible:** PI

- Which data will be collected?  
- How will the data be collected (paper form, electronic entry by patient, electronic entry by clinician, etc.)?  
- How will the data be entered into the database (keyed in, direct entry, uploaded, etc.)?  
- Who is responsible for data entry, and what is the timeline?  
- How will data entered into the database be validated?  
- Do the raw data require any type of manipulation prior to use (“scoring”; may include total scores, standard scores, etc.)?  
- How will the data be analyzed?

### FORM AND REPORT PREPARATION

**Parties involved:** PI, Data manager, Stats  
**Responsible:** DM

- Each measure specified in the plan is reviewed by PI (or responsible clinician) and assessed for redundancy. A given datum should only be obtained in one place (e.g., patients' sex indicated only on demographic form, not elsewhere).  
- DM creates paper forms that use a standardized template in order to reduce error.  
  - PI (or responsible clinician) provides original form and validates DM creation against it.  
  - PI (or responsible clinician) gives ultimate approval.  
- DM creates electronic forms that reflect the paper form (or other data source, if electronic/uploaded).  
  - DM ultimately obtains confirmation from PI (or responsible clinician) that form is valid.  
- DM confers with PI and statistician to anticipate reporting needs. At minimum, a report containing the scored primary outcome measures will be planned.  
  - DM creates and validates report.  
  - DM ultimately obtains confirmation from PI and Stats that the report is valid.  
  - Once real data are entered, this step will be revisited by DM (errors may become apparent).

### DATA COLLECTION AND ENTRY

**Parties involved:** Data manager, RAs  
**Responsible:** DM

- Data collection (electronic or paper) is immediately reviewed by study staff for completeness.  
- Data entry (or transfer) occurs immediately, in order to facilitate QC process.  
  - RA enters data.  
  - Second RA double-keys data and addresses any discrepancies.  
- DM monitors progress of data entry, per the timeline specified in the data management plan, and queries any issues that arise.

### ANALYSIS PLAN

**Parties involved:** PI, Stats  
**Responsible:** Stats

- Researcher brings idea to Stats, who help to flesh out plan and determine data requirements.  
- Stats communicates data request to DM.

### DATASET PREPARATION

**Parties involved:** Data manager, Stats  
**Responsible:** DM

- Based on the data request from stats, DM creates any necessary reports (or reports-of-reports, which combine necessary forms into one dataset).  
  - DM ultimately obtains confirmation from Stats that the report is valid.  
  - DM provides analyzable dataset to Stats, per specifications.  
  - Through analysis, Stats may identify problems with the report and will notify the DM to identify and address the error.

---

**Overview of the Data Management Process**
Step 1: Create forms for (clinical) data capture

- CTDB staff receive physical measure form, and NBPS staff indicate which items will be entered
  - Only most essential form of data are manually entered - nothing derived
  - Lookup tables from the measure manual are created by CTDB staff
  - E.g. domain raw score is entered and standard score, age equivalent, etc. is calculated by CTDB using lookup table
- Alternatively, a form may be created for direct data upload (in cases where data are exported from scoring software)
- This is usually done by the PI and the data manager (not the research assistants)
Step 2: Check, and re-check

- Ensure that all necessary items are captured
- Verify QC measures, like data ranges, required items, missing values, etc.
- Back-and-forth validation process ensuring that form is correct prior to (major) data entry
  - Changes to forms become increasingly difficult as more data are entered
  - This is usually done by the PI and the data manager (not the research assistants)
Step 3: Create reports

- CTDB utilizes IBM’s Cognos interface; researchers do not view the relational database itself.
- Reporting manipulates the raw data that you entered into the forms.
- Request is made by PI/data manager.
- Examples:
  - Calculate standardized scores from some lookup table.
  - Create novel variables (e.g., nonverbal ratio IQ).
  - Combine data from different forms and/or studies into a single report.
Step 4: Check, and re-check (again)

• SO MUCH CHECKING

• Software has done what it was asked, but there may be unexpected results. It is the human’s job to ferret those out.

• As with any QC issue, often not obvious until you’re using the data.
  • Because reports are manipulations of data which are not themselves changed or altered in any way, changing these as needs arise is less of an issue than for forms

• Initial checking done by data manager and PI, then by research assistants. Statistician also performs check when first using report.
Step 5: Data capture

• Research assistants:
  • Consult checklist during each visit to ensure all data are collected per protocol
  • Review all forms prior to the departure of the family to ensure that all data were collected per protocol
  • Enter data into scoring software where applicable (scoring by hand is avoided if possible)
Step 6: Data entry

• Research assistants:
  • Export data from scoring software (where applicable) and upload directly to CTDB, or
  • Enter (key) data into CTDB forms
    • A second RA re-keys each form and discrepancies are resolved - we view this as essential to ensure that data are valid
• Any changes made to entered data must be justified (i.e., change is allowed, but pop-up requires response)
• When data entry is complete, data are locked
  • Once data are locked, changes cannot be made by RA or investigator; any changes must be done by CTDB
Step 7: Access data

• When data are locked, those with sufficient permissions (PI, data manager, statistician) are able to download reports through website.

• Few manipulations are needed to make data usable for analysis, because they’ve all been done in reporting, which reduces room for error.

• Data are stored on large secure server that is backed up every evening - far more security than if stored on local machine or even local network.
Step 8: Share data

- If it is known prior to initiation that data will be shared, then reporting procedures will take this into account (i.e., give variables the same name as target database)
- Otherwise, CTDB creates maps of local variable names to database variable names (“harmonizing”).
- This is simply a new report! Planning makes this (relatively) painless, but a well-organized dataset means that any sharing request can be easily accommodated.
Acknowledgments

• Audrey Thurm, Ph.D. contributed to this work.
References

• Johh, Loewenstein, & Prelec (2012) *Psychological Science*.
• [https://www.go-fair.org/](https://www.go-fair.org/)
Resources

- https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/dmp/
- https://osf.io/k5wns/


- Please email us (see first slide) if you have questions!