Demonstrating Practical Use of Data Share and Secondary Analyses

Presented by:
Dee Blumberg, PhD
Abigail G. Matthews, PhD

Learning Objectives

• Review the structure of the NIDA Data Share website.
• Understand the information provided, how to navigate through the website, and obtain all the relevant information
• Demonstrate the technical procedures for using SAS or ASCII data sets for conducting secondary analyses.

Polling Question

Have you been to the Data Share website?

☐ Yes
☐ No
Introduction to Data Share

What is Data Share?
• Repository of publically available studies funded by NIDA
• CTN trials: ~34, plus 2 follow up studies
• Non-CTN trials: ~ 15 studies from the Division of Therapeutics and Medical Consequences (DTMC)

Over 80 countries have downloaded data from Data Share

United States • India • UK • Canada • China • Spain • France • Malaysia • Bangladesh Turkey • Egypt • Poland • Nepal • Brazil • Afghanistan • Mexico • South Korea • Australia • Canary Islands • Croatia • Denmark • Ecuador • Netherlands • Pakistan • Singapore • South Africa •

Goals of Data Share
• Optimize research productivity and use of resources
• Promote new research and secondary analyses
• Facilitate career development
Funding Opportunity for Data Share

Accelerating the Pace of Drug Abuse Research Using Existing Data (R01)

• “Of particular interest are analyses of data that have been harmonized and merged across multiple different datasets, such as those with PhenX measures... as well as data from NIDA Data Share for clinical trials...”
• Deadline October 5, 2016

Goals of Data Share

For example, it can be used for:
• Grant applications:
  – Preliminary data
  – Power/sample size calculations
• Secondary analyses:
  – Subgroup analyses
  – Mediation and moderation
• Secondary outcomes:
  – Harm reduction
  – Quality of Life

Data Share Policies

• Trial data posted 18 months after database lock or after primary manuscript published, whichever comes first
• All CTN studies are posted to Data Share
• Only de-identified data is included
• Only raw data are provided
Data Share Policies

• Truly open access:
  – No proposal necessary
  – Must agree to terms and conditions of use
• When publishing or presenting, must recognize the data source (NIDA Data Share)

Overview of Posted Studies

• Most studies are:
  – Multi-site randomized controlled trials
  – Primary outcome usually substance-use related
  – Duration: intervention and follow-up are short (3-6 months)
  – Sample size: usually 100-500
• Types of interventions:
  – Behavioral (e.g., CTN-0037- Stimulant Reduction Intervention using Dosed Exercise (STRIDE))
  – Medication (e.g., CTN-0001- Buprenorphine/Naloxone versus Clonidine for Inpatient Opiate Detoxification)
  – Both behavioral and medication (e.g., CTN-0030- Buprenorphine/Naloxone Treatment Plus Individual Drug Counseling for Opioid Analgesic Dependence)

Assessments

• Demographics
• Substance Use
• Mental Health
• Impulsivity and General Trait and Behavior Scales
• Interpersonal Relationships/Culture
• Physical/General Health
• Health Cognitions and QOL
• Clinical Measures
• Sexual Behavior/HIV
• Clinic Related Surveys
Available Information

- Study documentation:
  - Brief study description
  - Protocol
  - Link to primary manuscript (available from CTN Dissemination Library)
- Data documentation:
  - Annotated CRFs
  - Data dictionary
  - De-identification notes

Some examples

- Secondary analysis of individual studies
- Analysis of merged data from several CTN studies

Examples
A Research Example

- A researcher interested in the comorbidity, between drug addiction and depression...
Questions

Any questions so far before we move on to the next section ...

TIPS AND TRICKS

Documentation

- Read each protocol to understand:
  - Eligibility criteria
  - Visit timing/windows
  - Procedures for data collection
- Review annotated CRFs to:
  - Determine data elements
  - Evaluate codelists (commonality/harmonization)
  - Identify corresponding data files
- Read de-identification notes/nulled values docs
  - What information has been deleted*
  - What fields have been removed

* Contact study PI for access to raw data
Data Structure

- **CRF Level** – wide data files
  - One dataset per case report form
  - Each question/variable is a separate column
  - Requires substantial merging of different datasets
- **CDISC** – long and narrow data files
  - One row per question on CRF
  - Variable name/question described in one or more columns
  - All values/answers in one column
  - Minimizes amount of merging
  - Data from one case report form can be mapped to more than one dataset

---

### CRF-level data (CTN-0052)

<table>
<thead>
<tr>
<th>CRF</th>
<th>AGE</th>
<th>SEX</th>
<th>RACE</th>
<th>BLOODPRESSURE</th>
<th>BLOODPRESSURE RACED</th>
<th>TRACKIDENTITY</th>
<th>TRACKIDENTITY RACED</th>
<th>SIZE</th>
<th>SYSTEM</th>
<th>SPLIT</th>
<th>SPLIT RACED</th>
<th>START</th>
<th>END</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>M</td>
<td>1</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>55</td>
<td>M</td>
<td>1</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
<td>F</td>
<td>1</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
<td>M</td>
<td>1</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>70</td>
<td>F</td>
<td>1</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

---

### CDISC Data (CTN-0009)

<table>
<thead>
<tr>
<th>CRF</th>
<th>BLOODPRESSURE</th>
<th>BLOODPRESSURE RACED</th>
<th>TRACKIDENTITY</th>
<th>TRACKIDENTITY RACED</th>
<th>SIZE</th>
<th>SYSTEM</th>
<th>SPLIT</th>
<th>SPLIT RACED</th>
<th>START</th>
<th>END</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Data Formats

- **ASCII**
  - Text files
  - CSV – comma separated values
  - Can be read into SAS, SPSS, Excel, Stata, etc.

- **SAS**
  - Usually SAS transport file (.xpt)
    - Can be read into SPSS, SAS, DBMS Copy, and other programs
  - Sometimes standard SAS file (.sas7bdat)
    - Can be read into SPSS and SAS

Notes on De-identification

- Participant IDs are *not* those of the actual study
- Site is dropped
- All free text fields removed
- Dates converted to study day or days since informed consent

*If you find any of the above information – please inform DSC, IGNORE and DELETE it*

Helpful Tip

- Pay special attention to CDISC data
  - “Epoch” and “Visit Name” usually have to be combined or use study day
  - “Domain” is defined on each CRF
- These are raw data
  - Data cleaning not perfect – check for inconsistencies and outliers
  - Note how missing data is coded – varies from study to study and form to form
- Be very careful when merging datasets
- When publishing/presenting, make sure to reference primary manuscript
Secondary analyses have several pitfalls:
- Type I error (false positives)
- Underpowered
- Must be interpreted with caution
  - Cannot imply causality
  - Analyses are exploratory in nature and hypothesis-generating
- Analyses from Data Share will likely be post hoc – must acknowledge

Despite these pitfalls, secondary analyses are powerful tools and, as long as they are reported in a transparent manner, will contribute significantly to the literature.
CDISC Resources

• Official website:  
  http://www.cdisc.org/
• Standards:  
  https://www.cdisc.org/standards/foundational/sdtm
• SAS import:  

References

  http://ctndisseminationlibrary.org/display/1039.htm
• Funding opportunity: Accelerating the Pace of Drug Abuse Research Using Existing Data.  

Questions / Comments

Alternatively, questions can be directed to the presenter(s) by sending an email to CTNtraining@emmes.com.
A recording of this presentation will be available electronically.

Next Topic...

Preparation for Drug Management and Accountability in a CTN Clinical Trial
October 19, 2016
1:00 pm ET

THANK YOU FOR YOUR PARTICIPATION