Integrity of Research Data

Data Acquisition, Management, and Sharing

Definition of DATA

Recorded information; includes writings, films, sound recordings, pictorial reproductions, drawings, designs, or other graphic representations, procedural manuals, forms, diagrams, work flow charts, equipment descriptions, data files, data processing or computer programs (software), statistical records, and other research data

[*NIH Grants Policy Statement,*

"Rights in Data"]

Some Issues In Data Integrity

- Acquisition and record keeping
- Processing
- Ownership and control
- Retention and storage
- Access and sharing

Policies & Guidelines Concerning Data Integrity

- Federal (e.g., Public Health Service, Food and Drug Administration)
- Institutional (e.g., MIT)
- Laboratories or research groups

Data Acquisition and Record-Keeping

Record Keeping

Standards from FDA (standards for eventual approval of biologics, drugs, devices)

- Basic science: “Good Laboratory Practices” or GLP’s
- Clinical studies: “Good Clinical Practices” or GCP’s
- Standards for design, conduct, recording, reporting of research studies – data for approval of use must be of “highest quality and integrity” [FDA]

Record Keeping in Research Studies

FDA GLP:

- All data should be recorded promptly
- Entries must be dated and signed
- Any changes should not obscure original entry
- Reason for change must be indicated, signed, and dated

[*21 CFR PART 58*]

Laboratory Notebooks

What to record?

- Methods
- Results/observations
- Ideas
- Location of relevant computer & paper files
- Changes to protocol
- Notes from lab meetings
- Contributors at each stage

Laboratory Notebooks

How to record?

- As results are generated
- Chronologically
- With clear, readable entries

Laboratory Notebooks Do’s and Don’ts

Do

- Use bound notebook with sequentially numbered pages
- Write in black ink (resistant to moisture, solvents)
- Sign and date each entry
- Correct mistakes by slash, initialed and reason given

Laboratory Notebooks Do's and Don'ts

Do

- Include attachments by taping into notebook and putting initials/date across tape
- Save notebooks

Laboratory Notebooks Do’s and Don’ts

Don’t

- Erase
- Leave blank spaces without slash, signature, and date
- Tear out pages

Record Keeping in Clinical Studies

FDA GCP

- Maintain case history records
  - Case report forms
  - Supporting documents
- Ensure accuracy, completeness, legibility, and timeliness of data
- Make corrections to case history records so as not to obscure original entry (applies to both written and electronic corrections)

[*Federal Register, vol. 60, number 159*]

Record Keeping in Clinical Studies

- Case History Records should include
  - Basic subject information
  - Information that subject meets eligibility criteria
  - Information on treatment
  - Information from tests and examinations
    - Lab results
    - X-rays
    - Physical examinations, etc.
- Form of record can be printed, optical, or electronic

Record Keeping Using Computers

Computerized data collection systems must

- Allow data entry by authorized individuals only
- Control ability to delete or alter entries
- Provide audit trail
  - Who made change
  - When
  - Why
- Protect from tampering
- Ensure preservation of data (backup & recovery)
- Have system for generating print-outs

Data Processing

Data Processing

- Maximize following in data set:
  - Completeness
  - Accuracy
- Verify that
  - Data are entered correctly
  - Each variable falls within proper range
  - No missing values (or identify and handle cases with missing information)

Data Entry

- Use *double entry* systems
  - Enter observations into initial data set
  - Re-enter to validate
- Use direct transfer from equipment
- Check after entry

Out-of-Range Values

"Outlier" is observation that appears inconsistent with remainder of data set

- Detect outliers in a single variable
  - Are the minimum and maximum values for each variable what you expect?
  - Does the distribution have outlying tails?
- If possible, determine cause of outliers
- Decide how to deal with them before analysis

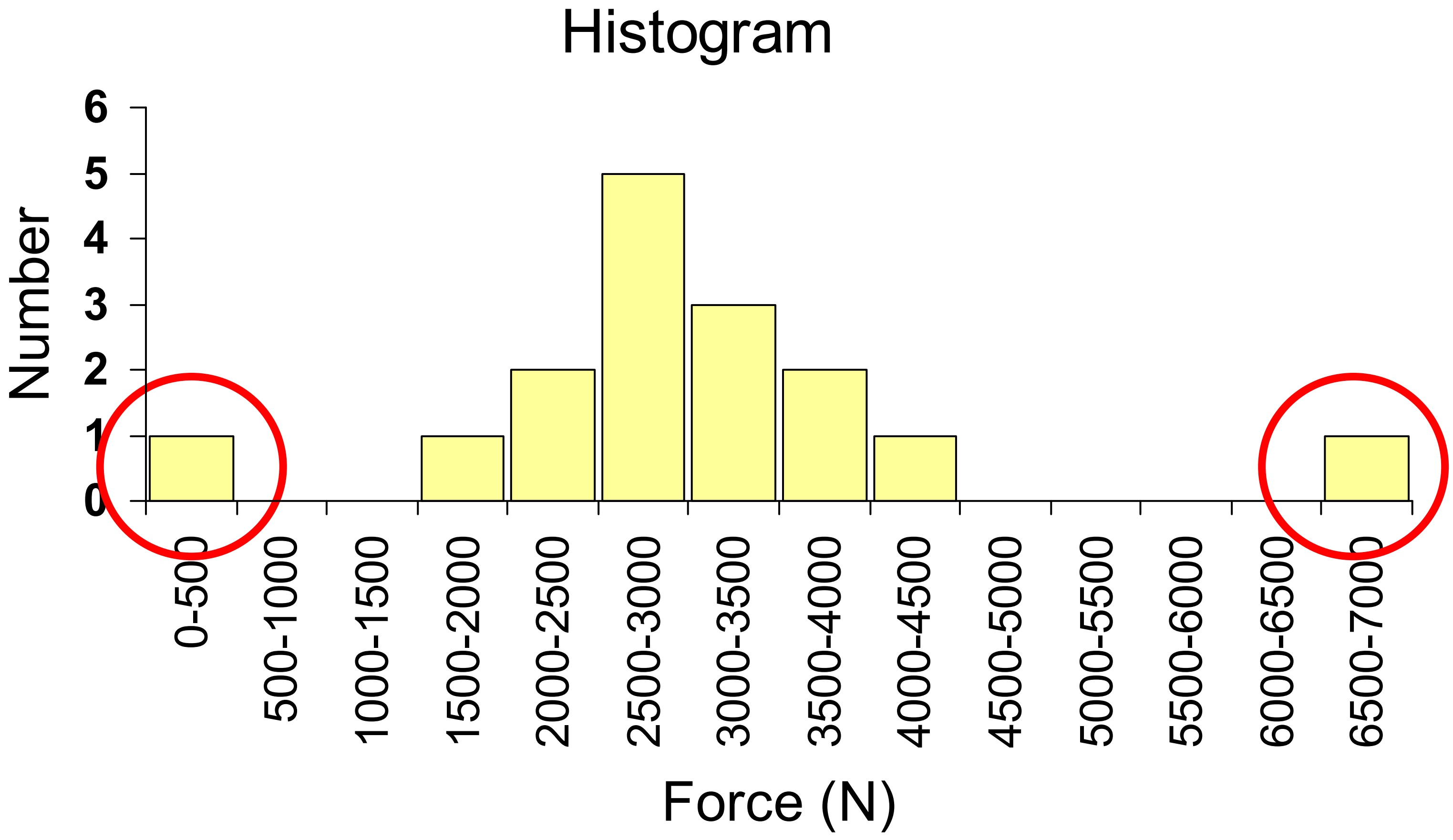
Possible Sources of Outliers

- Errors made in taking, recording, or entering data
- Case with extreme value is not part of population you intended to sample
- Extreme value is result of extreme (but real) biological or environmental variation

Example

Case	F (N)
1	20
2	3400
3	4500
4	2010
…	…
16	7000

Example


Histogram
Number
6
5
4
3
2
1
0
0-500
500-1000
1000-1500
1500-2000
2000-2500
2500-3000
3000-3500
3500-4000
4000-4500
4500-5000
5000-5500
5500-6000
6000-6500
6500-7000
Force (N)

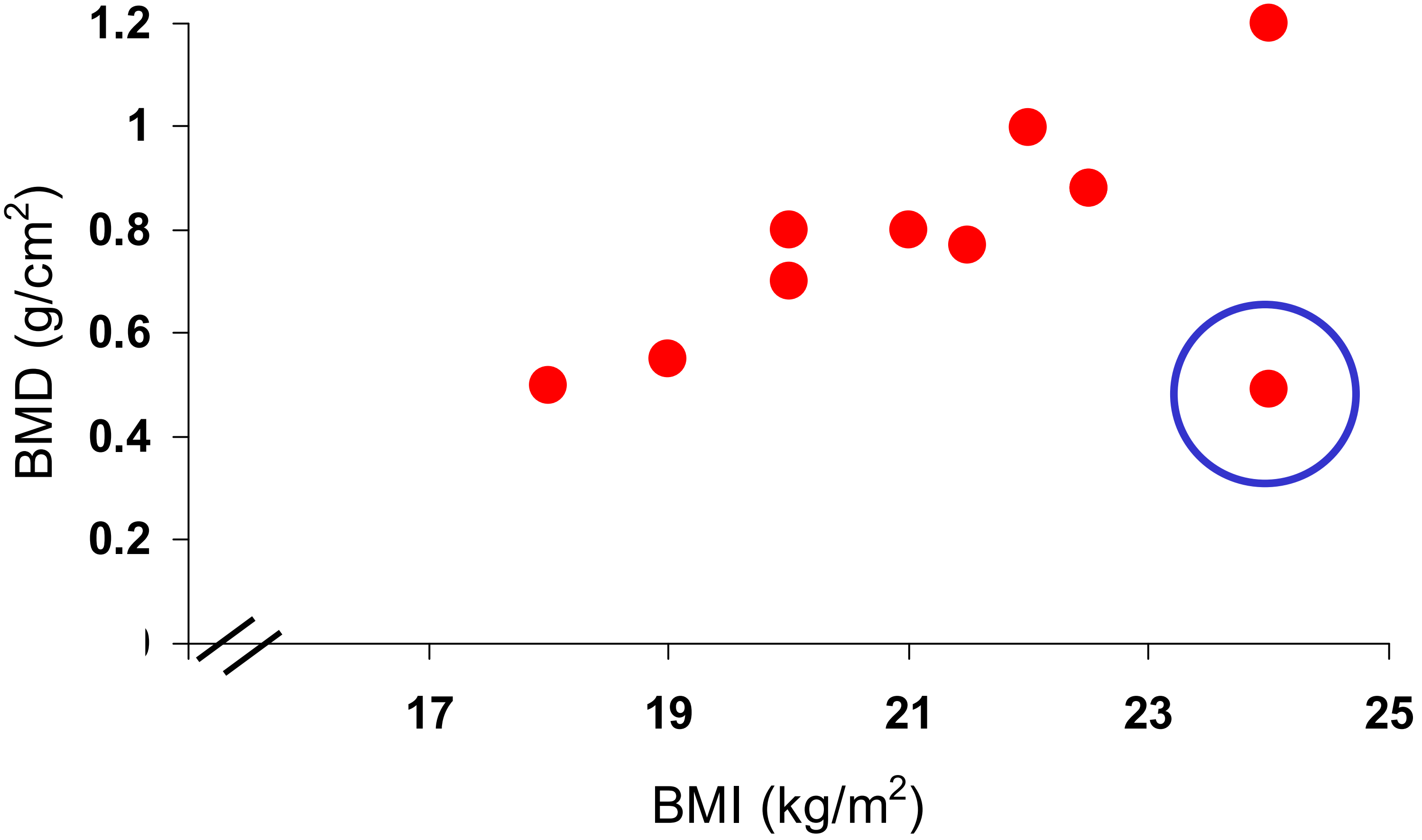
Example

- When checking original tracing of load cell output, find force for Case 1 = 2000 N → data entry error
- When checking original description of cases, find that Case 16 is male. Sample was intended to be female → case is not part of intended sample

Outliers in a Combination of Variables

- Unusual combination of two or more measurements
- Can use scatter plots to search among continuous variables for unusual patterns
- There are statistical multivariate methods to detect these outliers

Example of Multivariate Outlier



Outlying Data
What To Do

Strategy depends on source

- Inaccurate data entry
  - Replace with correct value
- Case is not from target population
  - Delete case if you know that you inadvertently sampled unit that is not from target population
  - Run analysis with and without case and compare results if you only suspect

Outlying Data
What To Do

- Extreme variation
  - Run analysis with and without case and compare results
  - Use techniques during the analysis phase that adjust for skewed data
  - Live with it and accept any distortion caused by outlier

Missing Values

- Identify and flag missing values
- Determine cause if possible
  - Examples: Patient drop-out, machine malfunction
- Quantify amount of missing information
- Check pattern -- should be random
- Decide how to deal with missing values before analysis

Missing Values -- What To Do

- Drop variable with missing value(s)
- Drop case with missing value(s)
- Accept unequal numbers of observation
- Be cautious with repeated-measures design

Bone Mineral Density (g/cm^2)

ID	Baseline	Year 1	Year 2
1	1.12	1.22	1.19
2	0.97		1.01
3	0.86	0.88	0.99

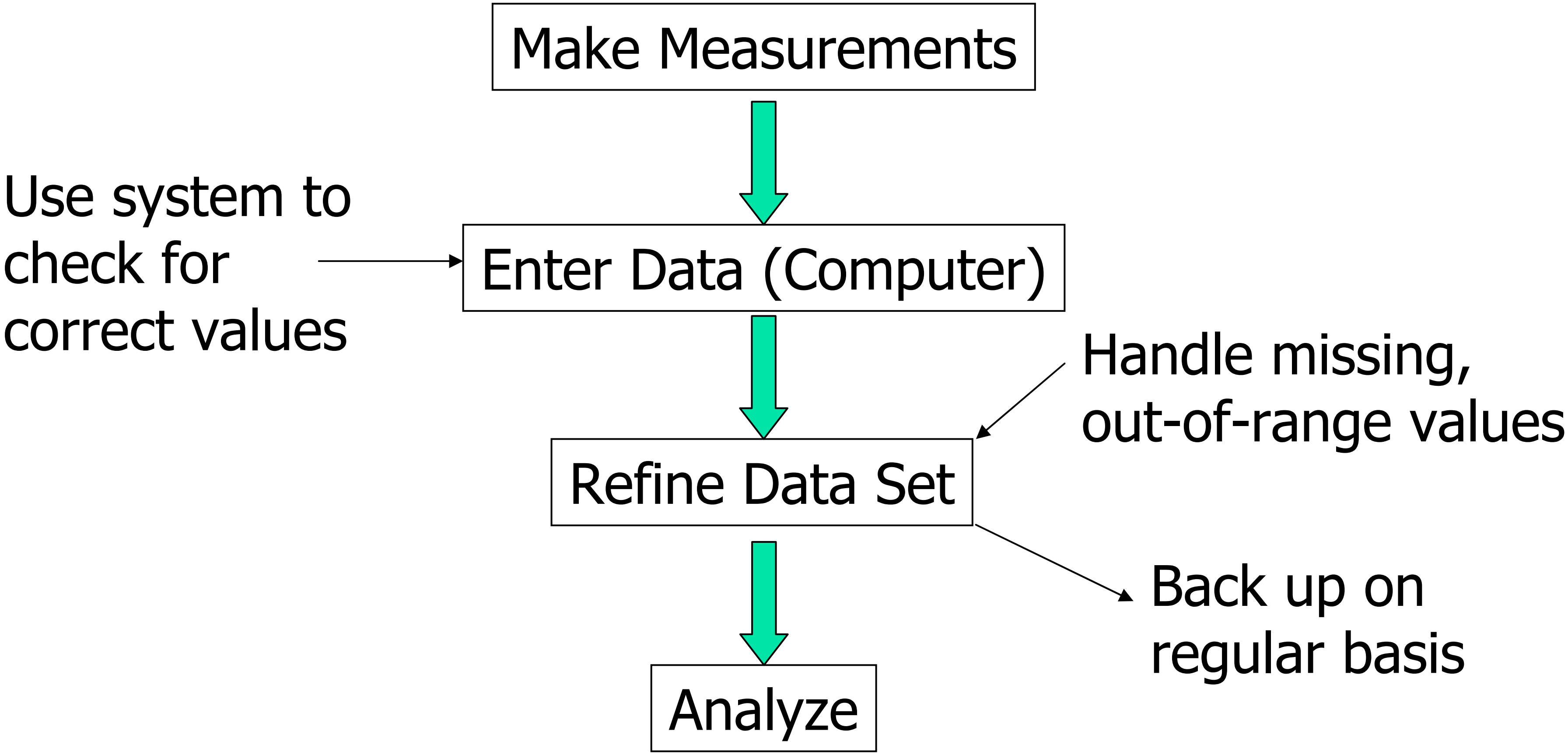
Missing Values -- What To Do

- Impute missing values (be careful)
  - Use average of non-missing values
    - Potential problem: variability of new data set may be underestimated
  - Use preceding non-missing value after sorting data into meaningful order
  - Use relationships with other variables to estimate missing value (do not use variables to be used in hypothesis testing)

Missing Values -- What To Do

- Treat missing data as data
  - Use when failure to have value may itself be predictor of outcome in your study
  - Make new “dummy” variable out of variable with missing values and use as another variable in analysis
    - Complete = 0
    - Missing = 1

Data Processing -- Summary



Make Measurements

Enter Data (Computer)

Refine Data Set

Analyze

Use system to check for correct values

Handle missing, out-of-range values

Back up on regular basis

Data Processing – Responsible Conduct

- Refining data set may require omission or modification
- Disclose basis for dropping or modifying data

Data Ownership and Custody

Data Ownership

Data ownership standards and policies depend on source of support

- Public Health Service (PHS): Data generated under PHS grant is owned by grantee institution
- Sponsored research: Ownership and control depend on research contract
- Unsupported: Typically institution retains ownership

Data Ownership – PHS Policy

“In general, grantees own the data generated by or resulting from a grant-supported project”

NIH Grants Policy Statement,
"Rights in Data"

Data Control - Sponsored Research

- For sponsored research contracts, control of data is often in “publication” paragraph
- Examples:
  - Institution is free to publish results of research after providing Sponsor with time to review for
    - Patentable material
    - Inadvertent disclosure of proprietary information

Data Custody

Custody or guardianship is typically in hands of PI or Laboratory (on behalf of institution)

Departure of Investigator

- Policies regarding departure depend on institution
- Examples:
  - If PI leaves institution
    - PI able to remove original data while allowing future access
    - PI takes copies
  - If co-investigator, post-doc, student leaves
    - Original data left with PI; copies provided
    - Original data left with PI; access allowed

Data Retention and Storage

How long must original data records be retained?

- PHS: 3 years after end of grant

  [45 CFR Part 74]
- FDA: 2 years after marketing approval

  [21 CFR Part 58]
- Institutions/laboratories: Varies from 2 years after publication to indefinitely

Data Retention and Storage

Where should original data records be stored?

- Often not stated explicitly
- Examples:
    - In laboratory
    - In research office of institution

Data Retention and Storage

Under what conditions should original data be stored?

- Ideal conditions
  - Temperature and humidity controlled
  - Protection against natural disaster or theft
  - Access controlled (no alterations) but easy to retrieve
- Reality
  - File drawer

Documentation of Storage

- Usually no set institutional policy
- Recommendation: Maintain tracking directory that records
  - Type of data (notebook, tape, x-ray, etc.) and identifying number
  - Project title
  - Investigator
  - Date data were recorded or notebook completed
  - Storage location

Data Sharing

Data Access and Sharing

Potentially opposing goals of data exchange:

- Right of investigators to get credit for data (through analysis, presentation at research meetings, publication)

versus

- Obligation of investigators to make data available to colleagues

Data Access and Sharing

Why do we need access?

- Openness and exchange of information are basic tenets of science
- Access may be required to investigate allegations of misconduct
- Access may be requested through Freedom of Information Act
- Access required for PHS-funded research

MIT Policy Statement on Data Sharing

“The prompt and open dissemination of the results of M.I.T. research and the free exchange of information among scholars are essential to the fulfillment of M.I.T.'s obligations as an institution committed to excellence in education and research”

Guide to the Ownership, Distribution and Commercial Development of M.I.T. Technology

NIH Statement on Sharing Research Data

- NIH expects timely release and sharing of final research data for use by other researchers
- NIH will require applicants to include a plan for data sharing or to state why data sharing is not possible. (applies to grants over certain amount)

NIH Statement: Why share?

- Extends NIH policy on sharing research resources
- Reinforces open scientific inquiry
- Encourages diversity of analysis and opinion
- Promotes new research
- Supports testing of new or alternative hypotheses and methods of analysis

Why share?

- Facilitates the education of new researchers
- Enables the exploration of topics not envisioned by the initial investigators
- Permits the creation of new data sets from combined data

What kind of information should be shared?

- Final research data necessary to validate research findings
- Does not include:
  - laboratory notebooks
  - partial data sets
  - preliminary analyses
  - drafts of scientific papers
  - plans for future research
  - communications with colleagues
  - physical objects, such as gels or lab specimens

To what kind of research does this apply?

- Data generated with support from NIH:
  - Basic research
  - Clinical studies
  - Surveys
  - Other types of research
- Especially important to share:
  - Unique data sets that cannot be readily replicated
  - Large, expensive data sets

Caveats for studies including human research participants

- Investigators need to be cautious with
  - Studies with very small samples
  - Studies collecting very sensitive data
- However, even these data can be shared if safeguards exist to ensure confidentiality and protect identify of subjects

What is meant by timely?

No later than acceptance for publication of main findings from the final data set

How to share data?

- Provide in publications
- Share under the investigator’s own auspices
- Place data sets in public archives
- Put data on a web site
- Place in restricted access data centers or data enclaves
- Other ways

Summary

Summary

- Policies on responsible data management are evolving
- Some current views:
    - Investigator's responsibility to ensure accurate and reliable data set, including
        - Primary sources (e.g., notebooks)
        - Refined data set

Summary

- Institutions tend to own data
- PI’s or laboratories tend to have custody of data
- Original data should be retained for ≥3 years
- Access is required for PHS-funded research (after publication) and may be required for other types of support
- Open dissemination of results is basic tenet of research community

Resources

- Committee on National Statistics, National Research Council, *Sharing Research Data,* National Academies Press, available at http://www.nap.edu/books/030903499X/html/index.html

Resources

- MIT TLO Policy Guide, available at http://web.mit.edu/tlo/www/guide.2.html
- National Aeronautics and Space Administration (NASA) Guidelines for Ensuring the Quality of Information, available at ftp://ftp.hq.nasa.gov/pub/pao/reports/2002/NASA_data_quality_guidelines.pdf
- NIH Data Sharing Information Main Page, available at http://grants2.nih.gov/grants/policy/data_sharing/