Data Management and Resource Sharing

Rigor & Reproducibility Workshop

24 May 2022

Melissa Eitzen, MT(ASCP), MS, RQAP-GLP
Director, Regulatory Operations
UTMB Institutional Office of Regulated Nonclinical Studies
mmeitzen@utmb.edu

utmb Health

1

Data Management and Resource Sharing

Topics

- Principles & Guidelines
- Data Lifecycle
  - Data Quality & Integrity
- Case Study—Break out session

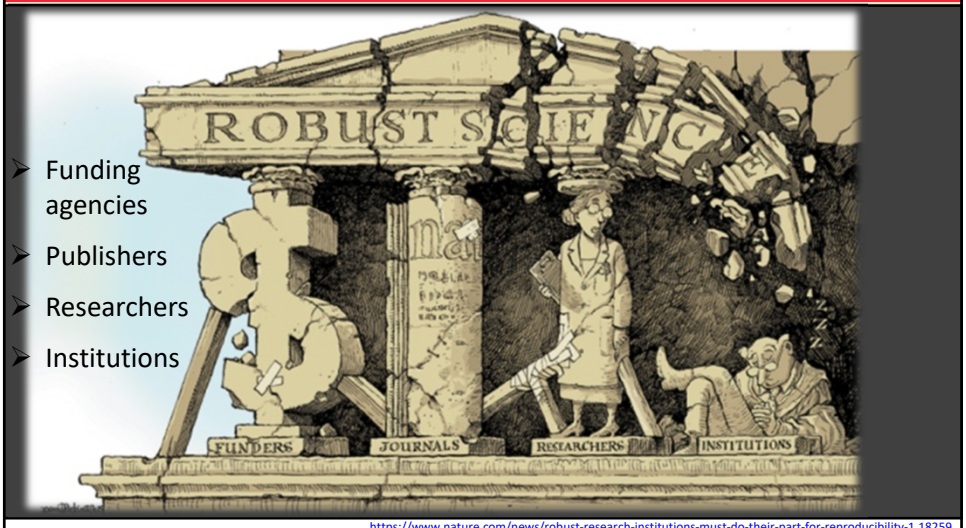
References provided on slides

utmb Health

2

Stakeholders of Robust Science

- Funding agencies
- Publishers
- Researchers
- Institutions



https://www.nature.com/news/robust-research-institutions-must-do-their-part-for-reproducibility-1.18259

utmb Health

3

3

NIH Public Workshop (2014)

- **Funding agencies**
- **Publishers**
- Researchers
- Institutions

- **Sponsors:** NIH + Nature Publishing Group + Science
- **Issue:** Reproducibility, Rigor of research findings
- **Attendees:** Journal editors (>30 basic/preclinical science journals where NIH-funded investigators publish)
- **Goals:** Identify common opportunities in the scientific publishing arena to ***enhance rigor and further support research that is reproducible, robust, and transparent***
- **Outcome:** set of principles to facilitate these goals, which a considerable number of journals have agreed to endorse

https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

utmb Health

4

4

NIH Principles and Guidelines

- **Funding agencies**
- Publishers
- Researchers
- Institutions

Principles and Guidelines for Reporting Preclinical Research:

- Rigorous statistical analysis
- Transparency in reporting
- Data and material sharing
- Consider establishing best practice guidelines for:
  - Images
  - Biological materials (antibodies, cell lines, etc.)
  - Animals
- Endorsements (journals, associations, societies)
- Adapted Guidelines

https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

utmb Health

5

5

Data and Material Sharing

- Funding agencies
- **Publishers**
- Researchers
- Institutions

- Require datasets be made available (where ethically appropriate) upon request
  - during manuscript review
  - upon publication
- Recommend datasets in public repositories, where available
- Encourage presentation of all other data values in machine readable format in the paper (or supplementary information)
- Require materials sharing after publication
- Encourage sharing of software
- Require a statement in the manuscript describing if software is available and how it can be obtained

https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research

utmb Health

6

6

Data and Material Sharing

- **Funding agencies**
- Publishers
- Researchers
- Institutions

Include a data sharing plan in research proposals seeking $500,000 or more in direct costs describing how final research data will be shared. Alternatively, the investigator is expected to explain why data sharing is not possible.

Release and share the data, as described in the approved application, no later than the acceptance for publication of the main findings from the final dataset.

On/After January 25, 2023:

Submit a Data Management and Sharing plan outlining how scientific data and any *accompanying metadata* will be managed and shared, taking into account any potential restrictions or limitations.

Comply with the Data Management and Sharing plan approved by the funding Institute or Center (IC).

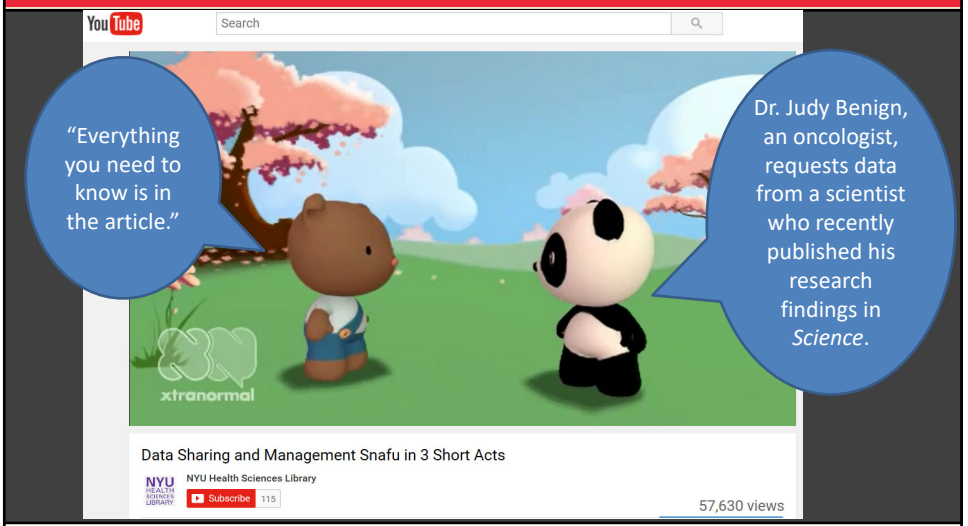
utmb Health

https://sharing.nih.gov/

7

7

Why is Data Management and Resource Sharing Important?



utmb Health

https://www.youtube.com/watch?v=N2zK3sAtr-4

8

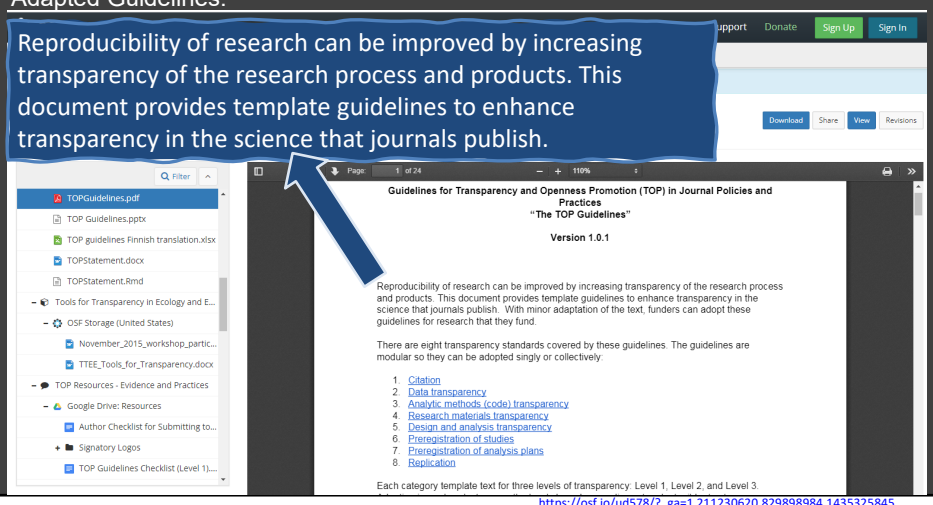
8

Guidelines for Transparency & Openness Promotion (TOP)

- Funding agencies
- **Publishers**
- Researchers
- Institutions

Adapted Guidelines:

Reproducibility of research can be improved by increasing transparency of the research process and products. This document provides template guidelines to enhance transparency in the science that journals publish.

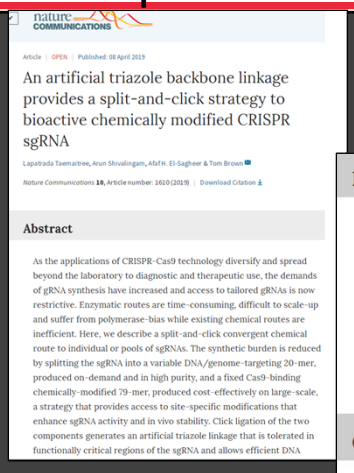


https://osf.io/ud578/?_ga=1.211230620.829898984.1435325845

9

Data Transparency—Example

- Funding agencies
- **Publishers**
- **Researchers**
- Institutions



Data availability

Sequencing data that support the findings of this study have been deposited in the NCBI Sequencing Read Archive with the accession code PRJNA512007. The source data underlying Figs. 1B–D, 2C–E, 3B–C, 4A–C; Supplementary Figs. 1, 2, 3B–C, 4, 5, 6, 7, 8, 9, 10; Supplementary Tables 1, 3, 4, 5, 6 and Supplementary Data 1 and 2 are provided in the Source Data file. All data for gels, graphs and mass spectrometry are provided as a Source Data file.

Code availability

Post-CIRCLE-seq data plotting code is available upon request.

10

Resource Sharing—NIH

- **Funding agencies**
- Publishers
- Researchers
- Institutions

NIH considers the sharing of unique research resources developed through NIH-sponsored research an important means to **enhance the value and further the advancement of research**.

When resources have been developed with NIH funds and the associated research findings published or provided to NIH, it is important that the **results be made readily available** for research purposes to qualified individuals within the scientific community.

https://grants.nih.gov/grants/peer/guidelines_general/Resource_sharing_plans.pdf

utmb Health

Resource Sharing—NIH

- **Funding agencies**
- Publishers
- Researchers
- Institutions

- Samples
- Reagents
- Model organism (e.g., transgenic mouse strain)
- Data

utmb Health

Data—Definition

Definition of Data

Data means recorded information, regardless of form or the media on which it may be recorded. The term includes computer software (computer programs, computer databases, and documentation thereof), and records of scientific or technical nature. The term does not include information incidental to award administration, such as financial, administrative, cost or pricing, or management information. In practice, scientific data include both intangible data (statistics, findings, conclusions, etc.) and tangible data. Tangible data include, but are not limited to notebooks, printouts, electronic storage, photographs, slides, negatives, films, scans, images, autoradiograms, electrophysiological recordings, gels, blots, spectra, cell lines, reagents, modified organisms, specimens, IRB consent forms, case report forms, drilling cores, collected organisms, and other materials that are relevant to the research project.

http://www.provost.pitt.edu/documents/RDM_Guidelines.pdf

utmb Health | University of Pittsburgh | Library System 13

13

Research Data—NIH Definition

NIH's 2003 Data Sharing Policy applies to the sharing of **final research data**.

Final research data* is recorded factual material commonly accepted in the scientific community as necessary to document and support research findings. For many scientific areas, final research data *includes both raw data and analyses conducted on the data*…

*Final research data = recorded factual material commonly accepted in the scientific community as necessary to validate research findings…

https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policy/research-covered-under-the-data-management-sharing-policy#before

utmb Health 14

14

Scientific Data—NIH Definition

The NIH Data Management & Sharing (DMS) Policy, effective January 25, 2023, applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of **scientific data**.

Scientific Data is defined as data commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications.

- Scientific data **includes** any data needed to validate and replicate research findings.

https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policy/research-covered-under-the-data-management-sharing-policy#after

utmb Health 15

15

Where Do We Begin?

utmb Health 16

16

Data Management and Resource Sharing

Topics

- Principles & Guidelines
- **Data Lifecycle**
  - **Data Quality & Integrity**
- Case Study—Break out session

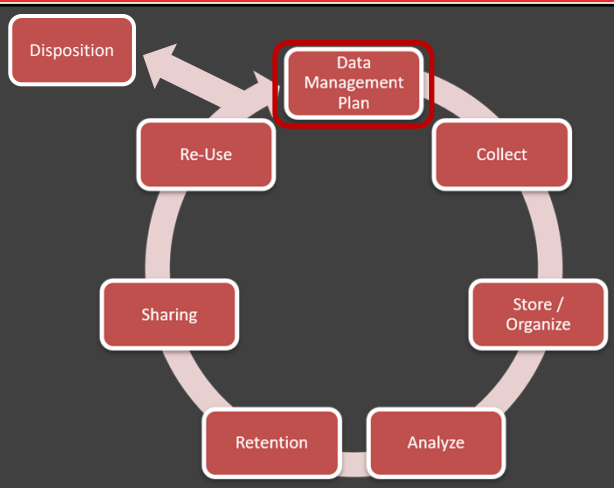
References provided on slides

utmb Health

17

17

Data Lifecycle



Research data management is a term that describes the organization, storage, preservation, and sharing of data collected and used in a research project.*

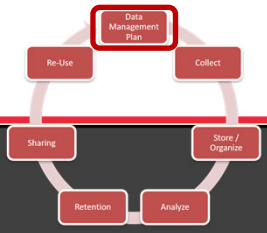
* https://pitt.libguides.com/managedata

utmb Health

18

18

Data Management



- Data is (are) a scholarly product
- Data are fragile and easily lost
- Growing research data requirements
- Good management helps prevent errors and increases the quality of your analysis
- Well-managed and accessible data allows others to validate and replicate findings
- ***Research data management*** facilitates sharing of research data and, when shared, data can lead to valuable discoveries by others outside of the original research team

utmb Health

https://pitt.libguides.com/managedata

19

19

Data Collection

Perform an inventory…

- Source (Raw) Data
- Types (observational, derived, etc.)
- Format (text, numeric, modeling, images, etc.)
- Quantity
- Regulatory (e.g., HIPAA)
- Proprietary
- Owner

20

20

Source Data (Original)

First capture of information

21

Images courtesy of Bing

21

Data Quality/Reproducibility Exercise

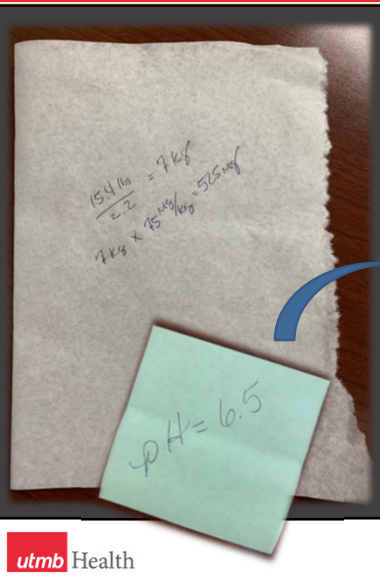
?

utmb Health

22

22

Source Data / Non-enduring



Scenario 1 (Worst case):

Scenario 2: Transcribe

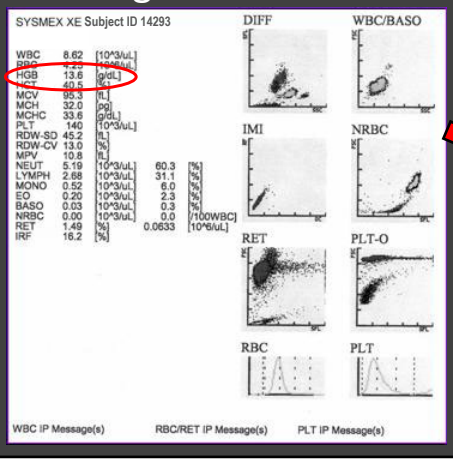
Scenario 3 (Lower Risk): Affix

Scenario 4: Eliminate bad practice

23

Transcription Errors

Hemoglobin Value



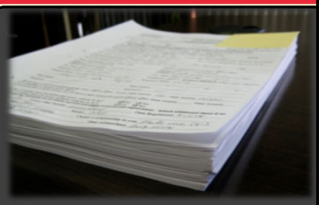
Animal	HGB
12938	12.2
14039	8.9
14293	3.6
14980	13.8
15209	12.5
15490	9.5
15560	14.0

Source: Google Images

24

ALCOA Principles

Applies to paper and/or electronic data



Data Quality

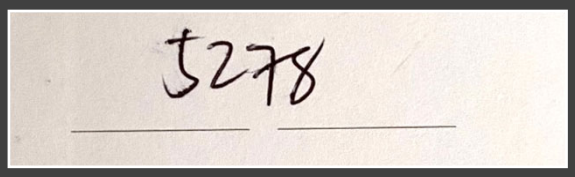
- **A**ttributable
- **L**egible
- **C**ontemporaneous
- **O**riginal
- **A**ccurate

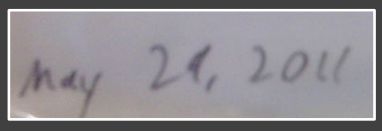
Data Integrity

- Complete, Consistent, Enduring, Readily Available

25

Illegible Data Entries





5/3/2022

26

Data and Data Integrity

"***Data*** are the foundation on which scientific, engineering, and medical knowledge is built."

~Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age, National Academy of Science, National Academy of Engineering, and Institute of Medicine; Preface, 2009

"***Data integrity*** is the degree to which data are complete, consistent, accurate, trustworthy and reliable and these characteristics of the data are maintained throughout the data life cycle."

~OECD Draft Advisory Document on GLP Data Integrity; 7 August 2020

utmb Health

27

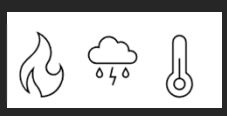
27

Organization and Storage / Retention

Things to think about prospectively…

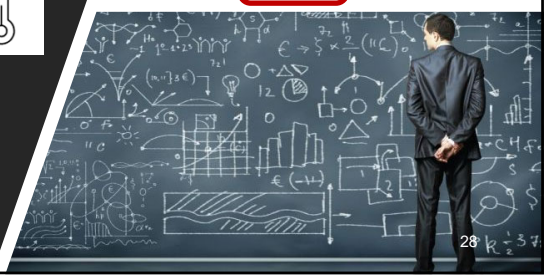
Conditions

- Location (physical / electronic)
- Transcription of source data
- Accessibility (limited)
- Security
- Change control
- Protection
- Migration



Contents

- Retain data and *methods* to allow for study reconstruction
- Retain protocols, reports
- Critical communication?

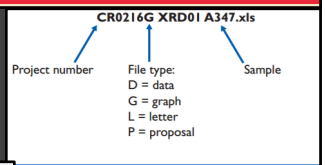


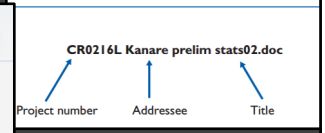
28

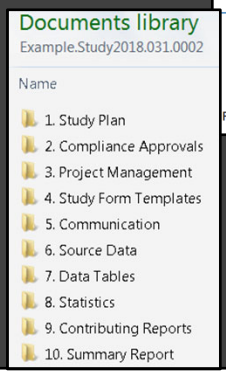
28

Managing Electronic Data

- Audit Trails / Meta Data
- Security / Encryption
- Software Compatibility
- Back-up
- Program Updates
  - Automatic
  - Impact to significant digits
- Data Migration
- Windows PC vs. MAC
- Checksums







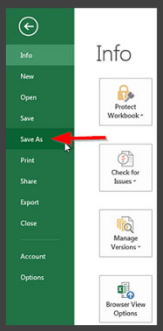
utmb Health

29

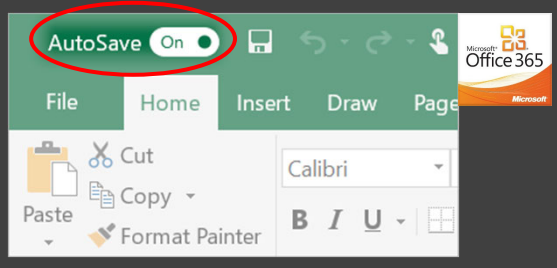
29

Risks to Electronic Data

Overwriting of information



- Save
- SaveAs
- AutoSave



utmb Health

30

30

Electronic Laboratory Notebooks

Pros

- Project organization
- Collaboration
- Custom forms/fields to assure all data are captured
- Procedure Checklists
- Time standardization
- Auto reminders
- Searchable
- Audit trail
- Data exportable

Cons

- Cost
- Sustainability ($)
- System administration
- Compatibility with other systems
- Software updates/data migration verification
- Discontinued (or support discontinued)

utmb Health

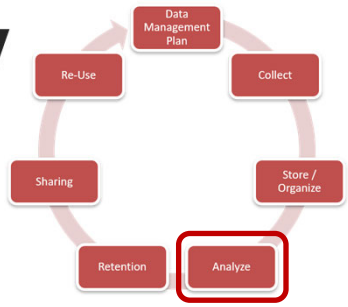
31

31

Data Analysis (Data Manipulation)

Prospective thinking…

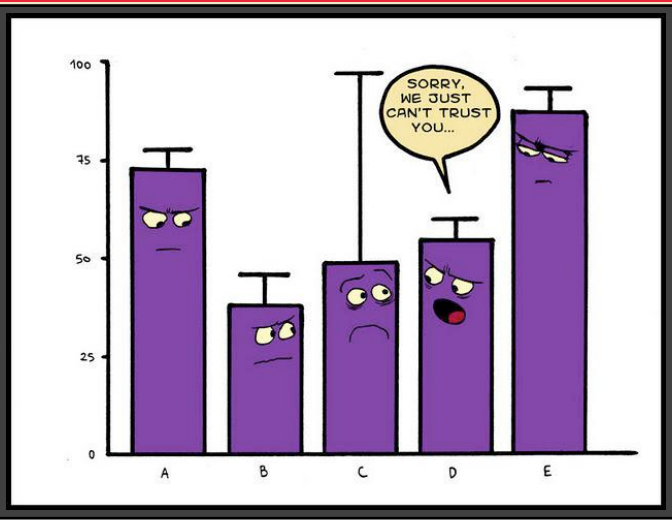
- Methods to reduce transcription errors
- Define inclusion / exclusion criteria
- Develop statistical plan (study plan)
- Retain *methods* to allow for study reconstruction



32

32

Data Manipulation

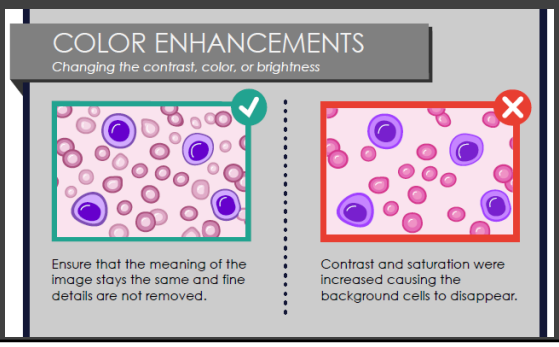


utmb Health

33

Image Manipulation

- Document all changes
- Retain unprocessed image
- Follow journal guidelines for permissible processing

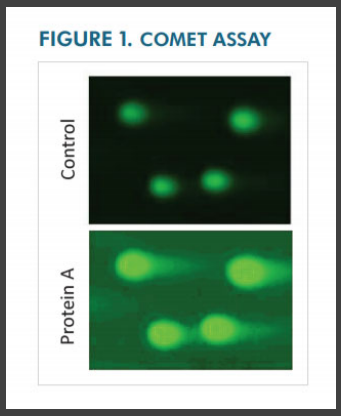


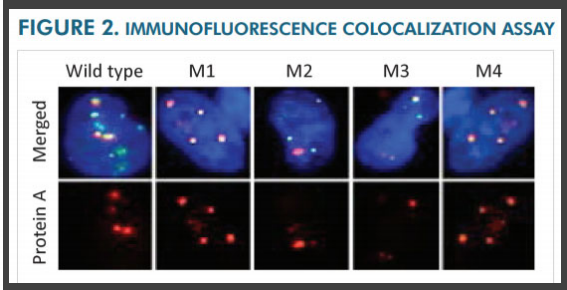
https://ori.hhs.gov/sites/default/files/2017-12/6_Image_Manipulation_scalable.pdf

utmb Health

34

Image Manipulation





https://ori.hhs.gov/infographics

utmb Health

35

Image Manipulation

DISCUSSION

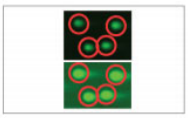


FIGURE 1. COMET ASSAY
The control image was cropped and relabeled as the image for Protein A. It was also intentionally lightened to make the "tails" appear longer.

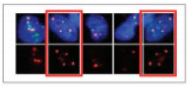


FIGURE 2. IMMUNOFLUORESCENCE COLOCALIZATION ASSAY
M1 and M4 are the same image but flipped vertically.

https://ori.hhs.gov/infographics

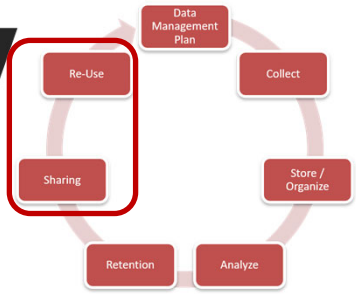
utmb Health

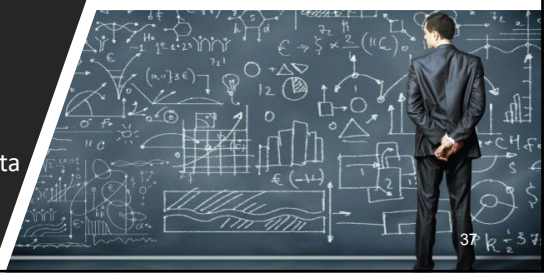
36

Mechanisms / Conditions for Sharing

Define conditions…

- Mechanisms & Format
  - Email
  - Online repositories
  - Supplemental to publication
  - Coding
  - Mixed media
- Restrictions (e.g., HIPAA), Conditions / Exclusions
- Sharing agreements / plans
- Schedule/timeline
- Acknowledgements of source data
- Instructions



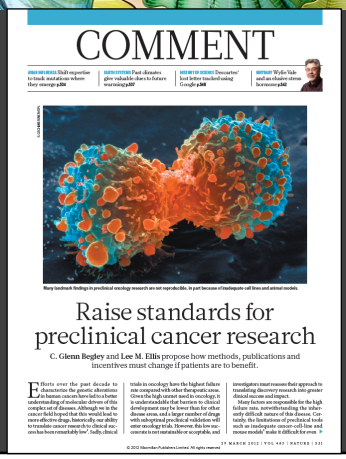


37

Data Management and Resource Sharing

Closing Thoughts…

- 53 landmark studies
- 6 confirmed (11%)
  - Controls
  - Reagents
  - Investigator bias
  - **Described complete data set**



COMMENT

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

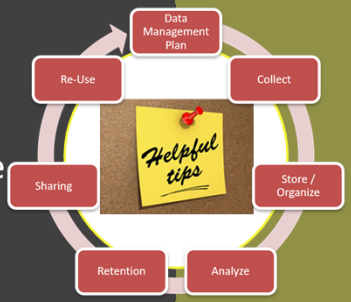
utmb Health

38

38

Data Management and Resource Sharing

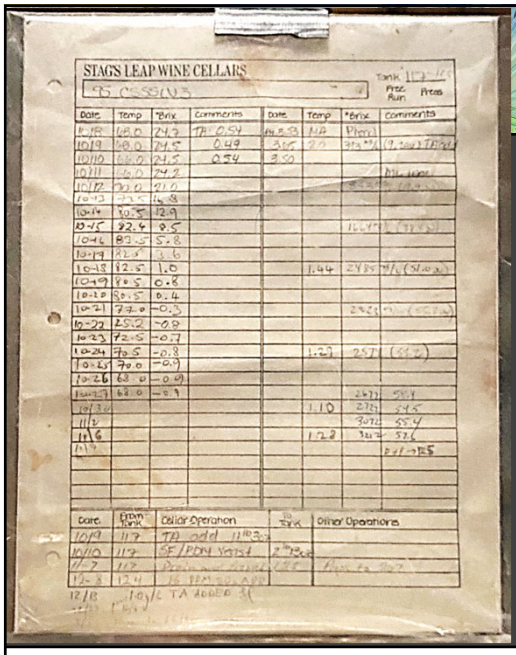
- Get organized!
- Data stewardship throughout the data lifecycle
- Data management plan
- Implement the ALCOA principles



utmb Health

39

39



Wine cellar, 1966

Photo by Ansel Adams, courtesy of University of California, Riverside

The wine cellar in the enology building at UC Davis held a wide variety of wines produced by faculty and students since the late 1930s. University researchers kept the wines for analysis and study, a resource that contributed to improvements in wine production.

Smithsonian National Museum of American History
Photos taken 23 April 2022 by M. Eitzen

utmb Health

40

40

Data Management and Resource Sharing

Topics

- Principles & Guidelines
- Data Lifecycle
  - Data Quality & Integrity
- **Case Study—Break out session**

References provided on slides

utmb Health

41

41

Case Study—Data Sharing

Identify options (i.e., conditions) for sharing data from a study with 500 human subjects being screened for sexually transmitted diseases.

utmb Health

42

42

Case Study—Data Sharing

The proposed research will include data from approximately 500 subjects being screened for three bacterial sexually transmitted diseases (STDs) at an inner-city STD clinic. The final dataset will include self-reported demographic and behavioral data from interviews with the subjects and laboratory data from urine specimens provided. Because the STDs being studied are reportable diseases, we will be collecting identifying information. Even though the final dataset will be stripped of identifiers prior to release for sharing, we believe that there remains the possibility of deductive disclosure of subjects with unusual characteristics.

Thus, we will make the data and associated documentation available to users only under a *data-sharing agreement* that provides for:
(1) a commitment to using the data only for research purposes and not to identify any individual participant;
(2) a commitment to securing the data using appropriate computer technology; and
(3) a commitment to destroying or returning the data after analyses are completed.

https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#ex

utmb Health

43

43

Photo: Suzanne & Walter Scott Chihuly Sanctuary, Univ. of Nebraska Medical Center

Thank you!

Melissa Eitzen, MT(ASCP), MS, RQAP-GLP
Director, Regulatory Operations
Institutional Office of Regulated Nonclinical Studies (ORNcS)
University of Texas Medical Branch at Galveston
2.810 Rebecca Sealy, Mail Stop 0184
301 University Boulevard
Galveston, TX 77555-0184
Email: mmeitzen@utmb.edu

utmb Health

44

44