Academic Research Integrity: Practical applications in the lab

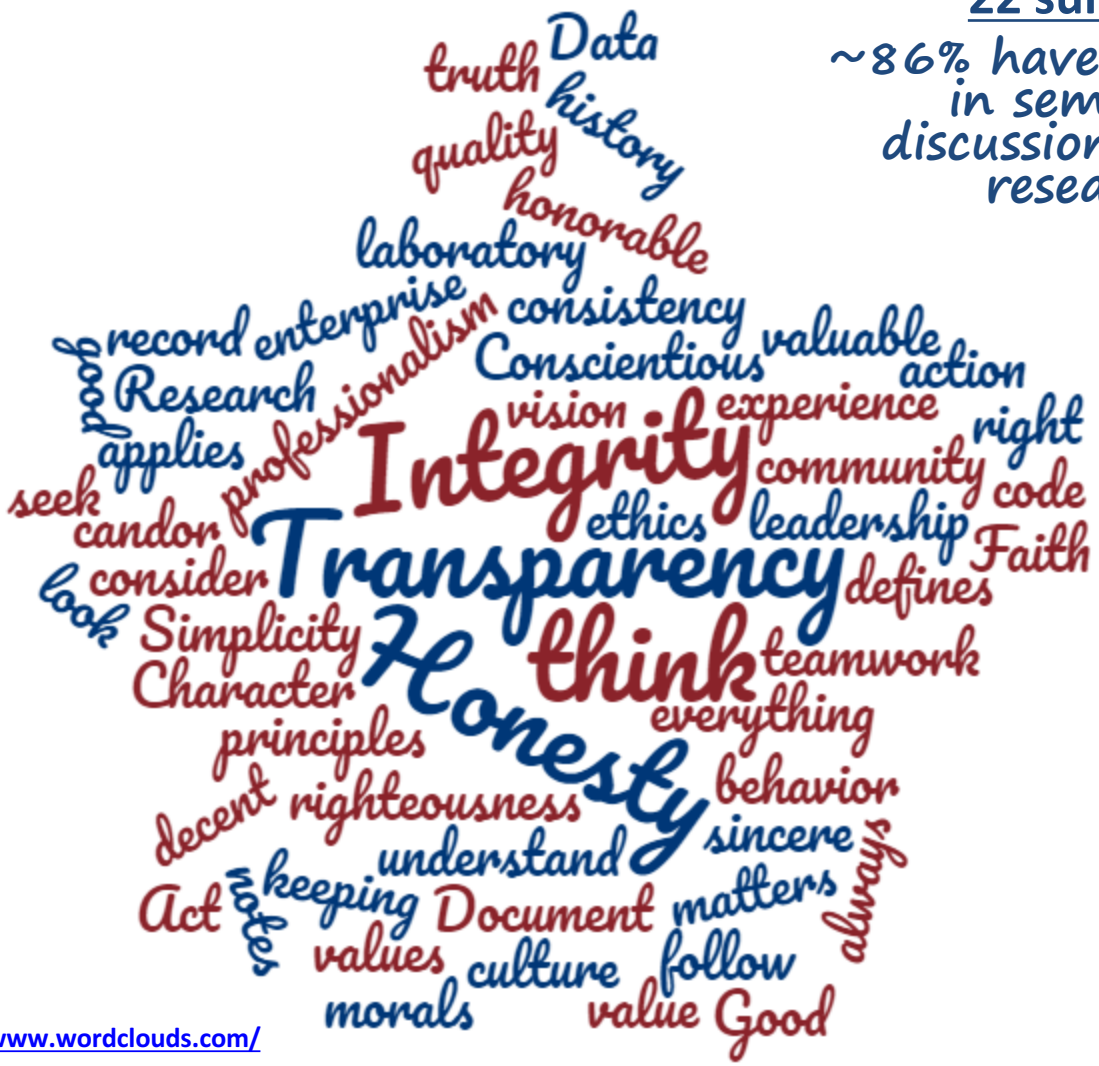
Mary C. Walsh, Ph.D.
Chief Scientific Investigator
Office for Academic and Research Integrity (ARI)
Harvard Medical School
mary_walsh@hms.harvard.edu

Daniel H. Wainstock, Ph.D.
Associate Director for Research Integrity
Office for Academic and Research Integrity (ARI)
Harvard Medical School
daniel_wainstock@hms.harvard.edu

What does the phrase "research integrity" mean to us?

22 survey responses

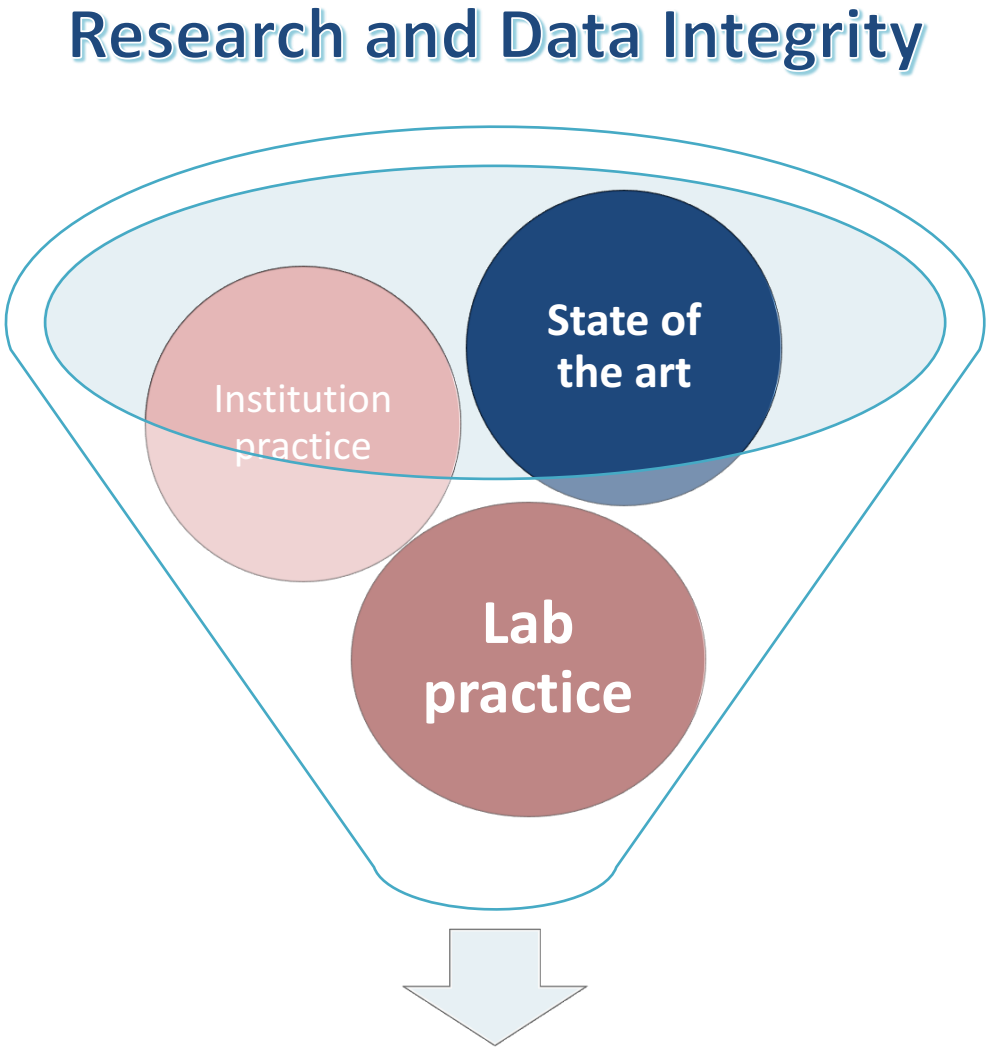
~86% have participated in seminars and/or discussions concerning research integrity



https://www.wordclouds.com/

What are we doing everyday in the lab that helps us "do it right"?

What helps us “do it right”?



Our approach to research and data integrity

What helps us "do it right"?

Research and Data Integrity

Think "field"

- *Federal grants: understand expectations of maintaining and sharing data*
- *Publishing data: what are the norms in the field(s)?*

Think "organizational"

- *Does my home institution have expectations around data maintenance?*
- *What are my resources to meet these expectations?*
- *How about my local Department or Center?*

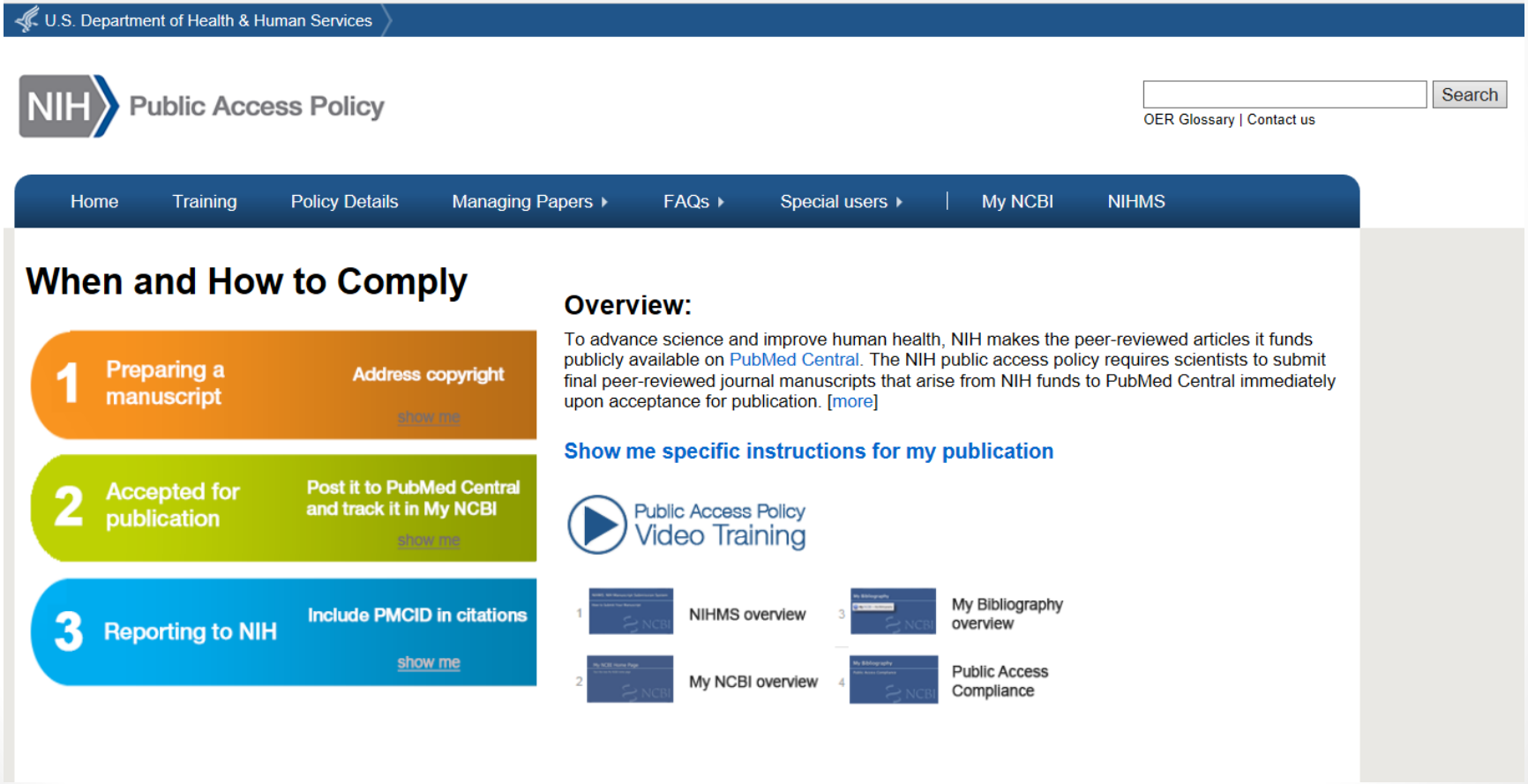
Think "local"

- *What does my lab (lab PI) expect of me?*
- *System for maintaining raw and analyzed data (local networks)*
- *System for tracking exact protocols and statistical analysis (notebooks/ELNs)*
- *Available hardware and software*
- *Track important biological metadata*

Field Resources

Research and Data Integrity

- **Requirements for Open Access and Data Sharing**
  - **Federal funding mandates* sharing! (NIH, NSF)**

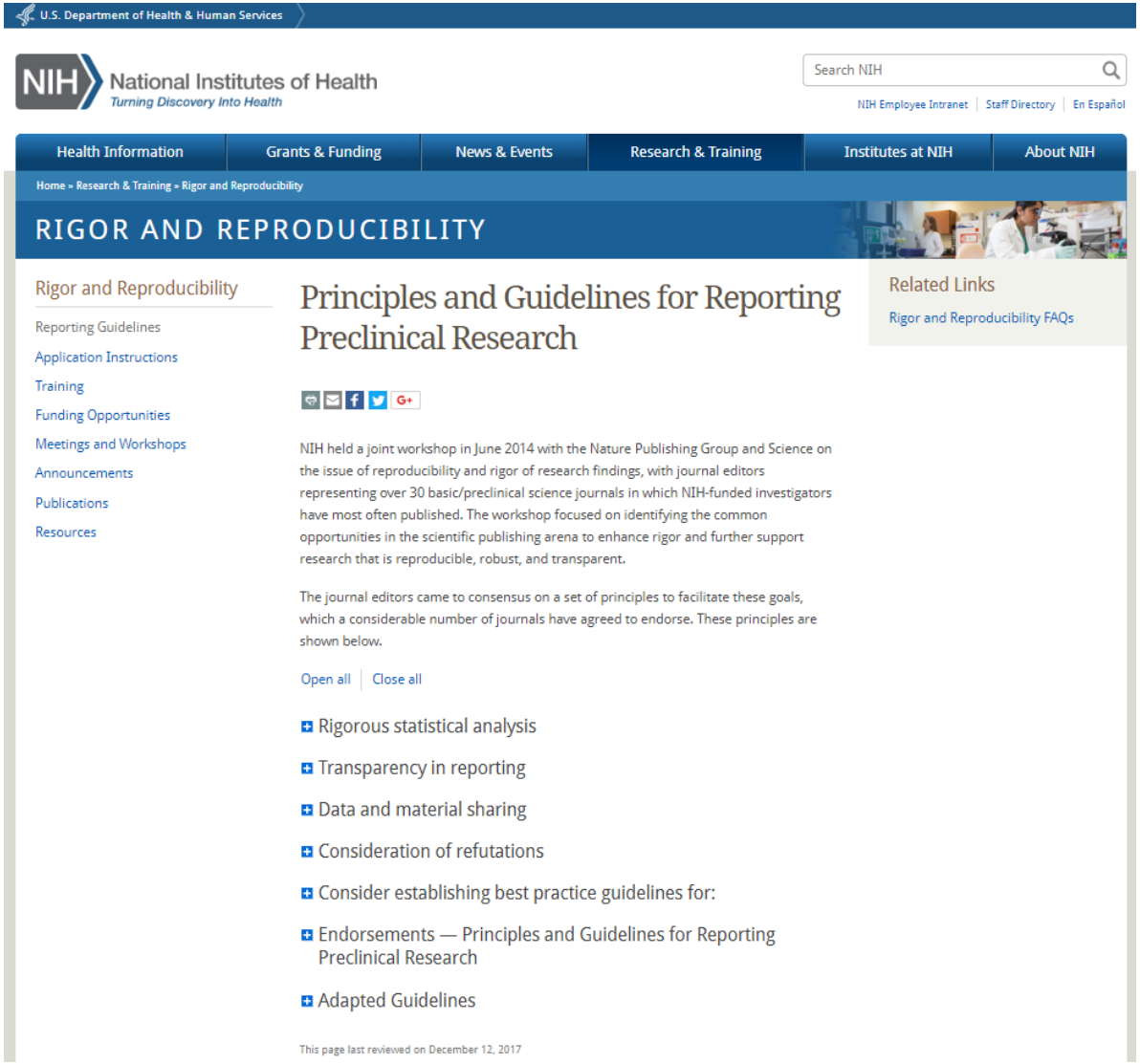


** while NIH Public Access is a global requirement Data Sharing has applicability guidelines*

Field Resources

Research and Data Integrity

- **NIH Rigor and Reproducibility Initiative**
  - **Consensus of thought leaders on scientific data (FAQs on Rigor and Transparency)**



Field Resources

Research and Data Integrity

- ***Reporting guidelines that may be required/applicable/helpful***

  The EQUATOR network (Enhancing the QUAlity and Transparency Of health Research)

  - ***ARRIVE** (Animal Research: Reporting of In Vivo Experiments)*
  - ***CONSORT** (Consolidated Standards of Reporting Trials)*
  - ***PRISMA** (Preferred Reporting Items for Systematic Reviews and Meta-Analyses)*

- ***Data management plans (DMPs)***

  - ***NISO Primer***
  - ***DMP Tool***

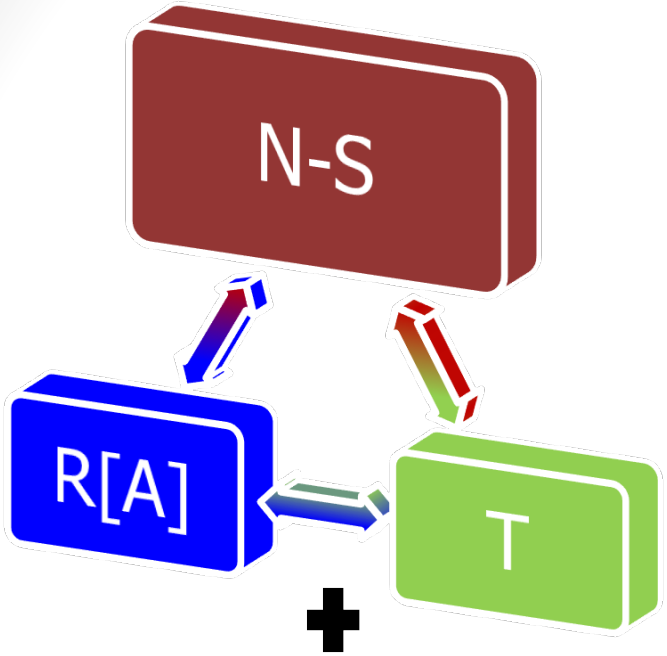
Field Resources

Research and Data Integrity

- ***What guidance do Associations/Societies provide?***
  - Society of Neuroscience (Journal of Neuroscience)
    - **Improving Your Science: Sample-size Planning, Pre-Registration, and Reproducible Data Analysis**
    - **Data science resources [Big Data: What you should know]**
    - **Webinar (*membership required*): Data Science Approaches for Neuroscientists**
  - AAAS **(Science)**
- ***What about journals?***
  - Nature **(Transparency and Reproducibility)**
  - Neuron (CellPress/Elsevier*)
    - **STAR methods**
    - **Mendeley***
      - ✓ **Importer and Desktop**
      - ✓ **Mendeley Data**

** Elsevier has partially supported other work independent of ARI Outreach initiatives*

- ***DMWG (Data Management Working Group)***
  - Links to fantastic resources supporting the Research Data lifecycle
    - **HMS Data Storage**
      - ✓ **OMERO**
    - **Repositories**
    - **ELNs**
- ***Harvard Medical School Research Computing***
  - Links to IT resources for HMS
    - **Storage and Training**
    - **Enterprise Licensed Software**
    - **HMS Software Wiki**
- ***Countway Library***
  - In addition to access to outstanding electronic journals and literature **(Hollis)**
    - **Research Services**
    - **Harvard Guidelines regarding public access and tools**
- ***Catalyst***
  - Great **research resources** (for basic and translational work) and **training opportunities**
    - **Biostatistics and Bioinformatics consulting**
    - **Programs and Classes available**
- ***edX***
  - Great free classes from experts on many topics
    - **Principles, Statistical and Computational Tools for Reproducible Science**
- ***Local Department Resources***
  - e.g., Department of Genetics has a **Computer Facility** in the NRB. Windows, Macintosh, and Linux operating systems on a wide variety of hardware. Installation/maintenance of desktop computers and servers; automated backups of servers; email for the Genetics Department; assistance with websites and programming.



RESULTS/RAW DATA

Data Integrity

Biomedical metadata

Lab Data Storage Structure and Lab Notebook (Notebooks and Storage)

- *ELNs*
- *Directory Structure*
- *File naming conventions*
  - *ReadMe files*
  - *Protocols (standards and, importantly, adaptation!) – Check out protocols.io*
  - *Statistical Analysis*

Reagents* [and Animals!]

- *antibodies*
- *cell lines*
- *reagent batching and lot#s*
- *animal work! (ARRIVE, IACUC, local facility, etc.)*
- *biological variables (e.g., sex as a biological variable [SABV])*

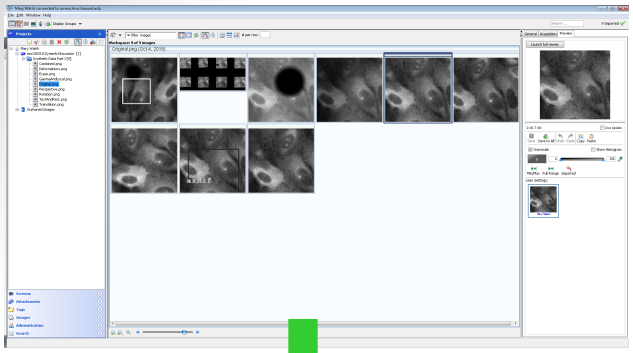
Hardware and Software (Tools)

- *Raw data production!*
- *understand and track specs*
- *HMS core facilities equipment that may have networked data storage on hardware*
- *lab equipment (networked or local data storage)*
- *paired software (proprietary or otherwise)*
- *Code? Check out GitHub, Jupyter and CodeOcean*

**NIH requires a reagent validation statement in all grant applications. Their Resource Chart is a helpful way to think about overall data organization.*

Local Resources and Practice

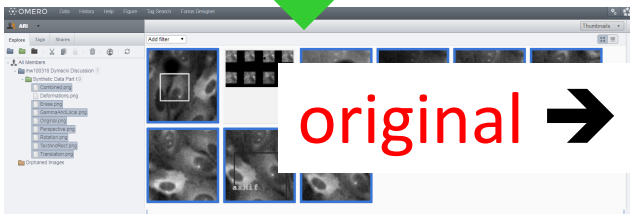
*original/analyzed/published data (a "synthetic" * example part I)*



ORIGINAL data acquisition - stays that way!

MCW9 OMERO
mw100318 Dymecki Discussion/
Synthetic Data Part I

.OIB file(s) "rotate.OIB"
Olympus FluoView FV1000



original ➔ analyzed ➔ published data

PUBLISHED

t al 2018)

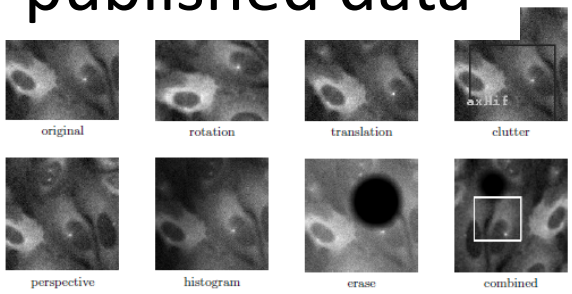
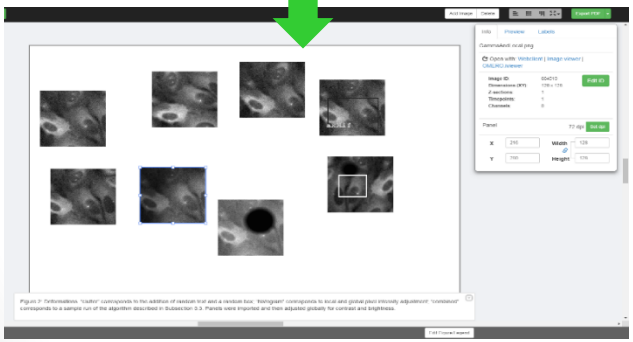


Figure 2: Deformations. "clutter" corresponds to the addition of random text and a random box; "histogram" corresponds to local and global pixel intensity adjustment; "combined" corresponds to a sample run of the algorithm described in Subsection 3.3.

COPY of original data into **mw100318 Dymecki Discussion/Synthetic Data Part I /ANALYSIS**

- .tif series created
- naming conventions/identification consistent



ANALYZED data in figure preparation software

- track all modifications/alterations (.jpeg)
- Lab presentation **mw100918.Dymecki discussion.ARI.pptx**
- Data Integrity manuscript *in preparation*

"Figure 2" Data. All modifications to original images (global) reported in Materials and Methods and Legend. When asked to make raw data available for editorial review all raw images (.oib) analyzed files (.tif) and the composite .jpeg figure with and without modification were provided

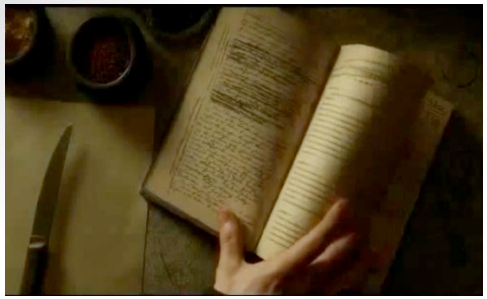
* *"L&A" disclaimer: any resemblance to actual persons, living or dead, or actual events is purely coincidental.*

November 2018
Academic Research Integrity
HMS ARI

Local Resources and Practice

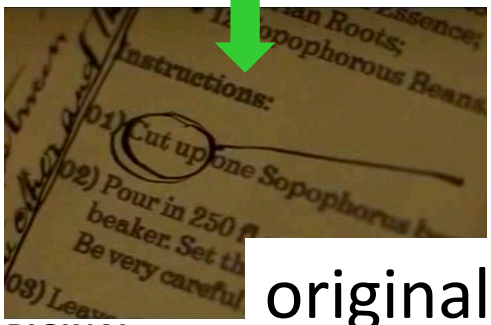
protocols/final protocols

(a "synthetic"* example part II)



Notebook 6 (MCW MBL KO work: 05.25.18 - 10.11.18)
pp.22-35
mw082118-092518
"Cytokine expression in isolated MBL KO vs WT microglia"

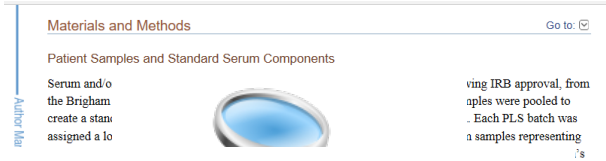
Snapshots taken by Mary C. Walsh from:
Draught of Living Death - Harry Potter and the Half-Blood Prince
YouTube · 7,000+ views · 5/31/2013 · by Ace Goins



original ➔ analyzed ➔ published data

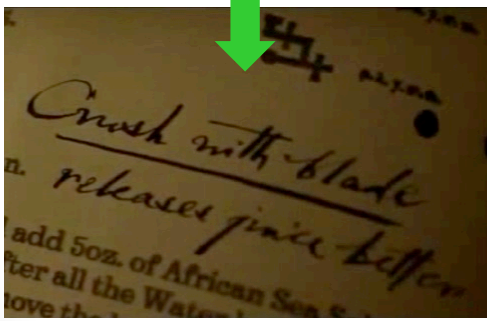
PUBLISHED

Materials and Methods section (**Welsh et al 2018**)



ORIGINAL

Notebook 6: pp.23-24, **mw082518**
Microglia Methods and Protocols (Springer 2013)
Microglial Activation: Measurement of Cytokines by Flow Cytometry (Chapter 9)



ANALYZED

Notebook 6: pp.24-35, **mw082518-mw092518**
Adaptation of Microglia Methods and Protocols (e.g., *intermittent vortexing 1hr on ice= 2x vortex, @15 min and @45 min for not longer that 10seconds/application*)

✓

Microglia were isolated and homogenized according to the protocols found in Microglia Methods and Protocols (Springer 2013, Chapter 9) Microglial Activation: Measurement of Cytokines by Flow Cytometry https://link-springer-com.ezp-prod1.hul.harvard.edu/content/pdf/10.1007%2F978-1-62703-520-0_9.pdf with the following notable adaptations: Optimal intermittent vortexing of initial isolation of homogenates was determined to be 2x vortex, @15 min and @45 min for not longer than 10 seconds/application during the 1 hr homogenate icing period.

* *"L&A" disclaimer: any resemblance to actual persons, living or dead, or actual events is purely coincidental.*

Local Resources and Practice

"statistics" for thought

(a "synthetic" example part III)*

ORIGINAL

Dymecki Discussion/ Synthetic Data Part III /mw082118/**Data Plan**

"Cytokine expression in isolated MBL KO vs WT microglia"

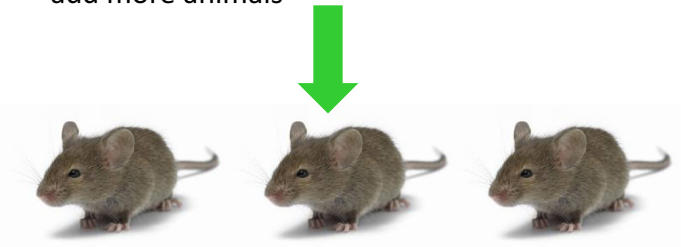
- **Variables**
- **Sample sizes**
- **Statistical tests**

ANALYSIS Part I

Dymecki Discussion/Synthetic Data Part III /mw090518/**ANALY**

- realize sample si
  seeing significant data on early samples)
- add more animals

original ➔ analyzed ➔ published data



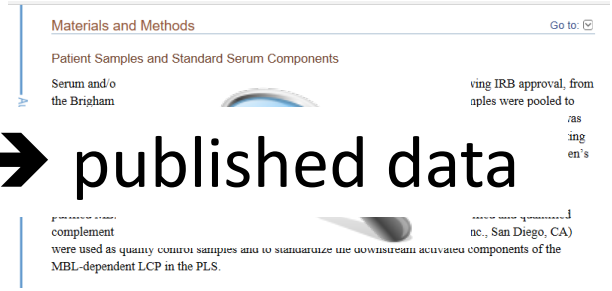
ANALYSIS Part II

Dymecki Discussion/Synthetic Data Part III /mw092218/**ANALYSIS**

- the original statistical tests proposed do not provide significance (even with additional animals)
- realize appropriate alternative test available, application results in significance

PUBLISHED

Materials and Methods section (**Welsh et al 2018**)


Materials and Methods

Go to:

Patient Samples and Standard Serum Components

Serum and/o ving IRB approval, from
the Brigham nples were pooled to
/as
ing
en's

complement nc., San Diego, CA)
were used as quality control samples and to standardize the downstream activated components of the
MBL-dependent LCP in the PLS.

For microglia cytokine expression experiments we provide all sample size, data exclusion and replication information as part of a detailed data management plan outlined at the initiation of experimentation and provided as part of the submitted "Life sciences study design". Any deviations from this plan are clearly detailed with accompanying rationale. Additional statistical assessments on data sets were included to provide a complete profile of hypothesis testing and resulting analysis rationale.

* *"L&A" disclaimer: any resemblance to actual persons, living or dead, or actual events is purely coincidental.*

NEWS

National Science Foundation: Science Hard

6/05/02 3:00pm • SEE MORE: SCIENCE & TECHNOLOGY

The scientists' assessment of a recent MIT paper on quantum physics.

"To be a scientist, you have to learn all this weird stuff, like how many molecules are in a proton," University of Chicago physicist Dr. Erno Heidegger said. "While it is true that I have become an acclaimed physicist and reaped great rewards from my career, one must not lose sight of the fact that these blessings came only after studying all of this completely impossible, egghead stuff for years."

https://www.theonion.com/national-science-foundation-science-hard-1819566466

"Genius is 2% (1%) inspiration and 98% (99%) perspiration."
-Thomas Edison

"Somewhere, something incredible is waiting to be known."
- (Carl Sagan) Sharon Begley

Thank You!

- Research and Data Integrity and the responsible conduct of research can have many meanings
- The steps we take every day, both large and small, can have an exponential impact
- Building careful data and metadata management into experimental design can speed up analysis, and ease trouble-shooting and replication
- When plans aren't enough, document changes
- Revisit the discussions such as the one we are having today (innovate)
- Share the wealth (disseminate)

Mary C. Walsh, Ph.D.
Chief Scientific Investigator
Office for Academic and Research Integrity (ARI)
Harvard Medical School
mary_walsh@hms.harvard.edu

Daniel H. Wainstock, Ph.D.
Associate Director for Research Integrity
Office for Academic and Research Integrity (ARI)
Harvard Medical School
daniel_wainstock@hms.harvard.edu

Acknowledgements

HMS Office for Academic and Research Integrity (ARI) Team
Gretchen Brodnicki, J.D.
Jennifer Ryan, J.D.
Mortimer Litt, M.D.
Blake Talbot, M.P.H.

HMS Data Management Working Group
Jessica Pierce

Columbia University Office of Research Compliance and Training
The Research and Data Integrity (ReaDI) Program
Michelle Benson, Ph.D.

Disclosures:
Maidstone Consulting Group, LLC

SURVEY FEEDBACK:

Elements (or resources) specific to research integrity, data integrity, transparency, and/or experimental rigor and reproducibility that you wanted to learn more about

- Electronic lab notebook organization and rigor, reproducibility, and note taking?
  - Check out slide 10
  - DMWG matrix on ELN (hyperlinked to the sites where you can get a sense of organization) and guides on Biomedical Metadata
  - Columbia ReaDI program (see slide 24, and Good Laboratory Notebook Practices)
  - Take a look at OSF (Open Science Framework) – search for "electronic laboratory notebook" some great examples
- Statistical tests?
  - Check out slide 10
  - UCLA idre: Choosing the Correct Statistical Test in SAS, STATA, SPSS and R and there are other examples out there…
  - Talk to your lab mentor and colleagues
  - Ask a statistician – Catalyst Biostatistical Consulting
  - Review your software guides (e.g., Stata handbook is actually great guide for applied statistics)
- GitHub and posting code?
  - Check out slide 10
  - GitHub, Jupyter and CodeOcean
- Reproducibility of computational research and working with large data?
  - Department of Biomedical Informatics
  - Harvard Chan Bioinformatics Core
  - Big Data to Knowledge (BD2K)
    - Best Practices for Biomedical Research Data Management
    - Big Data Science with the BD2K-LINCS (Library of Integrated Network-based Cellular Signature) Data coordination and Integration Center
  - Field specific field examples (e.g., check out slide 8 for neuroscience example - **Data science resources** [**Big Data: What you should know**])

SURVEY FEEDBACK:

Elements (or resources) specific to research integrity, data integrity, transparency, and/or experimental rigor and reproducibility that you wanted to learn more about

- Research misconduct that would be an accident or through ignorance, not anything deliberate?
  - Honest error is not research misconduct
  - Corrections (and retractions) are an appropriate element of typical scientific discourse
  - If you want to learn more about the standards for misconduct decisions (42 CFR § 93.104)
- Have there been cases of research misconduct that wouldn't be obvious to those of us who have taken courses on research misconduct?
  - Depends on the audience and the case
  - Check out RetractionWatch to get a sense of examples of possible cases and community discourse around process and outcome
- Statistics for misconduct?
  - ORI Website – Case Summaries
- How can we support or encourage our entire lab to evaluate (and possibly update) data management practices?
  - Discussion with your mentor(s) regarding local practice
  - Discussions at lab meetings regarding local practice (invite ARI to come if helpful!)
- At what point is reproducibility considered significant?
  - From day 1. Consider: If you and only you can do a thing once, how does it move the knowledge base forward? And keep in mind: you are your most frequent collaborator! So do yourself a solid and invest the sweat equity sooner rather than later.

Resources - slides

Free Online Word Cloud Generator https://www.wordclouds.com/

NIH Public Access Policy https://publicaccess.nih.gov/
NIH Data Sharing Policies and Implementation Guidance
https://grants.nih.gov/policy/sharing.htm
https://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm#app
NIH Data Sharing Policies Table
https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html
NSF Public Access Repository (NSF-PAR)
https://www.nsf.gov/news/special_reports/public_access/about_repository.jsp

NIH Rigor and Reproducibility https://www.nih.gov/research-training/rigor-reproducibility/principles-guidelines-reporting-preclinical-research
NIH Rigor and Reproducibility FAQs https://grants.nih.gov/reproducibility/faqs.htm

The EQUATOR network https://www.equator-network.org/
DMP Tool https://dmptool.org/
National Information Standards Organization. Research Data Management: A Primer
https://www.niso.org/publications/primer-research-data-management

Resources - slides

Society for Neuroscience http://www.sfn.org/
SfN Professional Development (Neuroscience 2017):
Improving Your Science: Sample-size Planning, Pre-Registration, and Reproducible Data Analysis
https://www.sfn.org/sitecore/content/Home/OMP/Articles/Professional-Development/2018/Improving-Your-Science-Sample-Size-Planning-Pre-Registration-and-Reproducible-Data-Analysis
SfN Scientific Research: Data Science Resources
https://www.sfn.org/sitecore/content/Home/OMP/Articles/Scientific-Research/2016/Data-Science-Resources
AAAS https://www.aaas.org/
Nature https://www.nature.com/authors/policies/reporting.pdf
CellPress STAR★METHODS https://www.cell.com/star-authors-guide
Mendeley Reference Tools
https://www.mendeley.com/reference-management/web-importer#id_1
https://www.mendeley.com/download-desktop/
Mendeley Data https://data.mendeley.com/

The Data Management Working Group
https://datamanagement.hms.harvard.edu/hms-data-management-working-group
HMS Research Computing https://rc.hms.harvard.edu/
Harvard Medical School Software Resources (HMS wiki) https://wiki.med.harvard.edu/Software/
Countway Library https://www.countway.harvard.edu/
HMS Catalyst https://catalyst.harvard.edu/
edX https://www.edx.org/
Department of Genetics http://genetics.hms.harvard.edu/

Resources - slides (con't)

DMWG ELN Matrix
https://docs.google.com/spreadsheets/d/1ar8fgwagOh30E31EAPL-Gorwn_g6XNf81g3VDQnQ_I8/edit#gid=0
DMWG (Best Practices Menu) https://datamanagement.hms.harvard.edu/
UCLA *idre*: What Statistical Test? https://stats.idre.ucla.edu/other/mult-pkg/whatstat/
Protocols.io https://www.protocols.io/
AntYbuddY https://www.antybuddy.com/
International Cell Line Authentication Committee http://iclac.org/databases/cross-contaminations/
Global Biological Standards Institute https://www.gbsi.org/
Animal Research: Reporting of *In Vivo* Experiments (ARRIVE)
https://www.nc3rs.org.uk/arrive-guidelines
HMS Core Facilities https://corefacilities.hms.harvard.edu/
GitHub https://github.com/
Jupyter http://jupyter.org/
Code Ocean https://codeocean.com/
Authentication of Key Biological and/or Chemical Resources
https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-068.html

The Book Designer
https://www.thebookdesigner.com/2010/01/6-copyright-page-disclaimers-and-giving-credit/
Nature Publishing Image Integrity and Standards https://www.nature.com/authors/policies/image.html
Microglia Methods and Protocols (Springer 2013)
https://experiments-springernature-com.ezp-prod1.hul.harvard.edu/springer-protocols-closure

Resources - slides (con't)

Mouse pic
http://4.bp.blogspot.com/-k9rfuvH0IFI/Tqq6LtQxfJI/AAAAAAAAALQ/6duHJ4osYVQ/s1600/Mus_Musculus_Unibe.jpeg
Stats pic
http://www.aph.gov.au/~/media/05%20About%20Parliament/54%20Parliamentary%20Depts/544%20Parliamentary%20Library/Flagpost/Statistics/125806739.jpg?as=1&w=250

The Onion (Science & Technology) 2002
National Science Foundation: Science Hard
https://www.theonion.com/national-science-foundation-science-hard-1819566466
1977 August 15, Newsweek, Volume 90, Seeking Other Worlds (Profile of Carl Sagan), Start Page 46, Quote Page 53, Newsweek, Inc., New York. (Verified on microfilm)
https://quoteinvestigator.com/2013/03/18/incredible/
1898 April, The Ladies' Home Journal, The Anecdotal Side of Edison, Subsection: His Estimate of Genius, Start Page 7, Quote Page 8, Column 2, Curtis Publishing Company, Philadelphia. (ProQuest American Periodicals)
https://quoteinvestigator.com/2012/12/14/genius-ratio/#note-5018-3

Additional resources

Columbia ReaDI Program
https://research.columbia.edu/ReaDI-Program

- Good Laboratory Notebook Practices
- Best Practices for Data Management when Using Instrumentation
- Reproducibility Resources and Guidelines by Topic
- Resources by Discipline: Neuroscience

Penelope AI
https://www.penelope.ai/

OSF (Open Science Framework)
https://osf.io/
"**TOP**" [**T**ransparency and **O**penness **P**romotion] Guidelines
https://osf.io/ud578/?_ga=1.211230620.829898984.1435325845

International Society for Pharmaceutical Engineering: GMP Resources
https://ispe.org/initiatives/regulatory-resources/gmp

Rigor and Reproducibility in NIH Applications: Resource Chart
https://grants.nih.gov/grants/RigorandReproducibilityChart508.pdf
Rigor and Reproducibility in NIH: Training
https://www.nih.gov/research-training/rigor-reproducibility/training

NIH Research Integrity and RCR
https://grants.nih.gov/policy/research_integrity/index.htm