RCR Instruction on Data Acquisition, Analysis, Ownership, and Retention: Can RCR Instruction Cover the Gaps?
Open Questions

• Why is RCR instruction on data acquisition, analysis, ownership (and sharing) and retention so important?
  – Data “management”
• What are the gaps for students and how do we approach RCR training on data at CSU?
• What are the gaps for faculty?
• Are there some potential solutions, approaches or strategies for filling in the gaps?
Defining Data

• How do we define data?
• Data are a collection of facts, measurements, or observations...used to make inferences about the world in which we live.
Examples of Data

- Not a complete list...
  - Material created in a wet laboratory, such as an organic compound, electrophoresis gel or a DNA sequence
  - Information obtained in social-science research, such as a filled-out questionnaire, videotapes, and photographs
  - Microscope slides, cell lines, climate patterns, soil samples, astronomical measurements, and spectrographic analyses
  - Custom software or hardware and specialized methods
  - Observations & interpretations made in the field (or lab)
  - Mathematical proofs, computer simulations
  - Other creative works
Why Data?

• Key component to RCR instruction [*Compliance*]
  – Mandated “core competency”

• Unifying topic that transcends all fields of study [*Education*]
  – Every student has “data”, even if they don’t call it that
  – Every student can identify with the need for “data” to achieve their goals
    • Graduation, publications, jobs
    • Justification for existence – building of professional reputation
  – Every student can relate to concepts of reliability and underlying need for good “record keeping”
Why Data?

- Heavily publicized fabrication of data affecting human health

The committee concludes that the body of epidemiological evidence favors rejection of a causal relationship between the MMR vaccine and autism. The committee also concludes that the body of epidemiological evidence favors rejection of a causal relationship between thimerosal-containing vaccines and autism. The committee further finds that potential biological mechanisms for vaccine-induced autism that have been generated to date are theoretical only.

IOM report on Immunization Safety Review, 2004
It all started with bad science. The now-debunked theory that autism is caused by the common immunizations nearly all children receive beginning in infancy began with a fabricated piece of research, a 1998 study published—and later retracted—in the journal Lancet. In 2010, Great Britain stripped Andrew Wakefield, the lead author of the study, of his medical license. An investigation had deemed his research an elaborate fraud.
Nevertheless, most new discoveries will continue to stem from hypothesis generating research with low or very low pre-study odds. We should then acknowledge that statistical significance testing in the report of a single study gives only a partial picture...

-Ioannidis, 2005
Scientists view...?

“Did you really have to show the error bars?”
Reliability & Reproducibility

• “Power Failure: why small sample size undermines the reliability of neuroscience”, Button et al., 2013
• “Challenges in Translating Academic Research into Therapeutic Advancement,” Matos et al., 2013 (epilepsy)
• “Reproducibility,” McNutt, 2014
• “NIH plans to enhance reproducibility,” Collins & Tabak, 2014
• “Reproducibility: Fraud is not the big problem,” – Gunn, 2014
Outcome

![Bar chart with captions]

- **A**: 25
- **B**: 35
- **C**: 50
- **D**: 60
- **E**: 85

Caption: "Sorry, we just can't trust you..."
• Issues with data reliability have brought external pressure on the scientific community
• From Congress
  – Presidential Council of Advisors on Science and Technology (PCAST) – “Improving Scientific Reproducibility in an Age of International Competition and Big Data”, 2014
    http://www.tvworldwide.com/events/pcast/140131/
• From the popular press and “watch dog” websites/blogs
  – The Economist - “Unreliable research: Trouble at the Lab”, 2013
  – NYT– “New truths that only one can see”, 2014
  – RetractionWatch.com
Of course, many outcomes are not readily captured, including creativity, risk taking, public advocacy and leadership. This kind of hypocrisy recalls the aphorism of “valuing what is measured” instead of “measuring what is valued.”

http://membercentral.aaas.org/blogs/driving-force/hypocrisy-accountability
The two opposite and contrary forces of data

Reproducible, robust, translatable to bedside, rigid, immutable, non-optimized, boring (preclinical or clinical study)

Dynamic, agile, discovery, exploration, optimization, creative, outside-the-box, anti-dogmatic (pre pre-clinical study)
Where’s the disconnect? (or How do we connect in RCR instruction?)

• If RCR instruction on data management is a “core competency,” why do scientists continue to have these huge (and public) gaffs related to their data?
• Where are the gaps for trainees?
• Can RCR instruction actually address these gaps, or are we doomed to failure?
Evidence of Failure

- Students self reporting
  - Sources of prior knowledge of relevant topics (CHEM601)
  - Data “pre-test” (GRAD544)
- Student examples of record keeping
  - Actual NIH/NSF supported research results
“Which of the following topics have you discussed in a class, with peers, in a research group meeting or with your research advisor or other faculty member?” (Check all that apply)

- Methods for proper record keeping
- Principles for responsible use of animal subjects
- Principles for responsible use of human subjects
- Importance of honestly reporting what you find
- Criteria for authorship
- Risks of conflicts of interest
- The peer review process
- Responsibility and strategies for action after having witnessed research misconduct
Where do students learn about data handling?

- Importance of honestly reporting what you find
- Methods for proper record keeping

- Some differences exist between populations
- Overall: >80% report discussions in at least 1 venue
Nearly 20% of student authors have never had a conversation about authorship criteria.

Less than 2/3 of student authors have discussed authorship criteria with a faculty member.

Can they reasonably be expected to defend the data in response to allegations of FFP or irreproducibility?
(More) Evidence of Failure

• <50% of the trainees have a good working knowledge of plagiarism or FFP (plagiarism* or F/F**)

• They are particularly handicapped in understanding specific issues around data reproducibility

• GRAD544 Data “pre-test” results
  – Extremely cross-disciplinary course
  – >500 trainees to date (grad students, postdocs, undergrads)
  – 20 different departments

Who owns it?

If you perform an experiment and record results in a notebook or on a spreadsheet, do you then "own" the data?

- Yes, I won 100% of it (10)
- Yes, but I only own 50% because my advisor also owns it (4)
- No, my funding agency owns it (15)
- No, CSU owns it (44)
- I am not sure who owns it (27)
The responsibility for deciding what and when to publish or share data with others is held by:
- You, the data collector (8)
- The PI (66)
- The institution (2)
- The funding agency (12)
- I am not sure (12)
Do you know how many times any particular experiment must be repeated in order to feel satisfied that the results are trustworthy?

- Yes (10)
- Yes, but each experiment is different (55)
- I think so, but I am not really sure (7)
- No (15)
- I do not think this is relevant to my field (12)
Are faculty any better?

- Anecdotally, faculty involved with MiS at every stage indicated that they were not trained in RCR or data collection.
- 75% CSU junior faculty*, in a mentoring workshop, indicated that they had received no RCR instruction on data acquisition.
- 42%* indicate that they are not providing specific instruction on data acquisition, storage, retention to their trainees.

Yet, we assume an apprenticeship approach to acquiring this knowledge: fundamental flaw?

* Broccardo et al., unpublished observations (n=32)
Yet, we (ORI/NSF) continue to require training only of students/trainees, but not faculty: fundamental flaw?
Gaps in Current RCR Instructional Materials

• **ORI: Introduction to the Responsible Conduct of Research (2007)**
  – 10 pages on data management practices
  – 3 pages on data collection
    • 2 Paragraphs on “Appropriate Methods”
    • 2 Paragraphs on “Attention to Detail”
    • 1 Paragraph on permissions, “Authorized”
    • 2 Paragraphs on “Recording”
      – Hard copy documentation
      – Electronic documentation
Gaps in Current RCR Instructional Materials

• **NAS On Being a Scientist**
  - 4 pages on “The Treatment of Data”
    • Fraudulent image manipulation
    • Poor experimental design or careless measurements
    • Documentation of methods
    • Accurate recording
    • Sharing data
“Six Common Failings” (Martinson) related to how we should teach RCR

1. Poor experimental design
2. Poor reagents
3. Poor analysis
4. Failure to reject hypothesis after observing discordant, valid experimental results
5. Deliberate bias in selecting positive rather than negative results to report, publish, cite, and fund
6. Failure to ask and follow through on “Why is this result NOT what I expected?”
RCR Texts & Mentors Are Insufficient

- Existing reference material offers broad, non-specific information
  - Students have difficulty applying to actual “data”
- We suggest that RCR texts, as most of us employ them, are insufficient.
  - They teach concepts and norms at a high level, but lack the specific details that are needed day-to-day
- Reliance on the “apprenticeship” model flawed
  - Mentors have failed to provide sufficient foundational understanding
  - Trainees become mentors themselves
Strategies & Approaches (Solutions?)

- What could/should RCR courses provide?
- National strategies
- Examples of “Record Keeping” homework assignments
- Two in class exercises on data
  - What to do with the data?
  - Is it ok to manipulate my data?
National Strategies

• National strategies
  – Publishing houses, NIH initiatives, Science exchange, etc.
  – Publishing vs. funding
    • Check list, impact factor, incentives for PIs to publish
    • Changing the biosketch for grants
  – Data repositories

Bringing it down to a fundamental aspect: data & record keeping
Meeting GLP: Good Laboratory Practice

A  Attributable (who made the entry)
L  Legible
C  Contemporaneous/Complete
O  Original
A  Accurate

This represents a change in pre preclinical science
Data Entry

- Raw data should be recorded and retained in indexed laboratory notebooks with permanent binding and numbered pages or in a dedicated electronic notebook (*original*)
- Writing should be done in permanent ink and legibly (*legible*)
- Recording should be done as data are collected (*contemporaneously*)
- Each page should be initialed and dated (*attributable*)
- Supervisors should review [and sign off on notebooks to signify their completeness and accuracy] (*accurate*)
Data Entry

- Material should be logged chronologically.
- Methodologies should be documented.
- Deviations from a standard technique should be explained.
- A second notebook should be kept for data, such as photographs, etc.
- A few blank pages should be kept at the front of a bound book for tables of contents.
- Data should be noted directly in notebooks without putting it on scraps of paper or relying on memory.
Data Entry

• Lot numbers should be recorded.
• Equipment calibrations should be recorded.
• All raw data should be included – include bad data.
• Data interpretations should be carefully written.
• Correspondence and notes about conversations related to experiments should be kept.
Data Corrections & Amendments

• Errors, additions, and modifications should be identified by crossing out the original data with a single line (do not obscure the initial data) and initialing, dating and providing a reason for the change.
• Missing or obscured data/pages are often interpreted as intentional obfuscation of data
  • Absence is interpreted as guilt …
How do we make this real to students?

Make them report out on their own record keeping
Student Narrative

“Since our lab is not required to keep a lab notebook, none of the researchers in the lab actually have a lab notebook. We record physiology data using Chart and then back it up with an excel spreadsheet. Our other data such as body mass, heart mass, muscle mass is just written on a piece of paper then entered into a spread sheet and the paper is put in a pile. Also, our biochemical data is written on a piece of paper or in the case of Westerns, a picture is taken. HPLC data is probably the best recorded methods we have. A journal is kept in MS Word as well as hand written with printed chromatograms. An example of our usual record keeping is below.”
Page 3: This is a continuation of the procedure from the previous page (numbers are cut off). The following problems can be found:

- I believe on this page we were doing a separation in a conical vial. I do not state what we are trying to accomplish by adding NaOH. I probably didn't know at the time.
- I did cross out the remaining empty space and dated the end of the procedure maybe I should have initialed where I marked out the empty space.
- I think a sketch would have helped make the procedure easier to follow.

Good: reaction scheme, page numbered, dated, amounts used given in table, theoretical yield reported, column gradient reported. Bad: TLC drawn in pencil, no mechanism shown, not enough observations, not all times recorded, TLC not labeled, no observations of pure NMR, no references, no copy of NMR, no information on column length.
Stats Basics: Every student should be able to discuss:

• The relationship between sample size, power, confidence and significance
• What is the fallacy of small $n$ values?
• How do you estimate variation based on the literature or pilot experiments to derive the appropriate sample size?
• How do data selection and statistical tests contribute to biased interpretation?

How do we make (at least some of) this real to students?

Give them some data and ask them to make decisions
Data Exercise

• We developed a very simple, straightforward “case study” to delve into students’ ability to apply their “core knowledge” of data “management” to a specific set of data
• Adapted from Ethics of Emerging Technologies (Budinger & Budinger)
Electromagnetic Data

- Your team has decided it wants to determine if there is a correlation between an increase in environmental electromagnetic noise and the incidence of adult leukemia.
- Your team gathers data regarding the incidence of adult leukemia over the past century, using hospital records.
- You take a look at the notebook(s) and find…
### Raw data

<table>
<thead>
<tr>
<th>Year</th>
<th>Incidence (#/mil)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1910</td>
<td>2</td>
</tr>
<tr>
<td>1920</td>
<td>2</td>
</tr>
<tr>
<td>1930</td>
<td>4</td>
</tr>
<tr>
<td>1940</td>
<td>4</td>
</tr>
<tr>
<td>1950</td>
<td>3</td>
</tr>
<tr>
<td>1960</td>
<td>3</td>
</tr>
<tr>
<td>1970</td>
<td>5</td>
</tr>
<tr>
<td>1980</td>
<td>4</td>
</tr>
<tr>
<td>1990</td>
<td>10</td>
</tr>
<tr>
<td>2005</td>
<td>5</td>
</tr>
</tbody>
</table>

**What conclusions can be made from the raw data?**

A. There is an upwards trend in the incidence of leukemia

B. The incidence in leukemia has not changed significantly

C. There are too few data points to make any conclusions

**Could any additional conclusions be made from looking at the descriptive statistics?**
Descriptive statistics

<table>
<thead>
<tr>
<th>Column1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4.2</td>
</tr>
<tr>
<td>Standard Error</td>
<td>0.72724747</td>
</tr>
<tr>
<td>Median</td>
<td>4</td>
</tr>
<tr>
<td>Mode</td>
<td>4</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.29975844</td>
</tr>
<tr>
<td>Sample Variance</td>
<td>5.28888889</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>4.89474361</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.95672725</td>
</tr>
<tr>
<td>Range</td>
<td>8</td>
</tr>
<tr>
<td>Minimum</td>
<td>2</td>
</tr>
<tr>
<td>Maximum</td>
<td>10</td>
</tr>
<tr>
<td>Sum</td>
<td>42</td>
</tr>
<tr>
<td>Count</td>
<td>10</td>
</tr>
</tbody>
</table>

Does the dataset look “OK”?
A. Stats look ok
B. Stats are worrisome
C. Not enough info to determine

One team member decides to perform an analysis of the raw data and make a preliminary conclusion about the incidence of leukemia
Team Member A concludes that there has been an increase in the incidence of leukemia over the past 90 years.

Team Member A further reports that over the same period of time, the background noise increased 100-fold.

Ready to call the NY Times?
A. Yes, this is pretty cool
B. No, it is not an interesting finding
C. This is potentially interesting but need more info
D. These data are raising a red flag for the team!
Team Member A reports that there is a significant difference in the incidence over the second half of the century versus the first half. True?

A. Yes
B. No

Team Member B tells Team Member A that it looks promising but the team (collectively) is still not quite convinced...
Team Member A decides to “clean up” the data so that the R-value is much stronger.

How was this done?

\[ y = -52.269 + 0.028604x \quad R = 0.7696 \]

Anova: Single Factor

<table>
<thead>
<tr>
<th>Groups</th>
<th>Count</th>
<th>Sum</th>
<th>Average</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Column 1</td>
<td>5</td>
<td>15</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Column 2</td>
<td>4</td>
<td>17</td>
<td>4.25</td>
<td>0.91666667</td>
</tr>
</tbody>
</table>

ANOVA

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>P-value</th>
<th>F crit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Group</td>
<td>3.47222222</td>
<td>1</td>
<td>3.47222222</td>
<td>3.60082305</td>
<td>0.09955917</td>
<td>5.59145974</td>
</tr>
<tr>
<td>Within Groups</td>
<td>6.75</td>
<td>7</td>
<td>0.96428571</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>10.2222222</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Column 1            |         |      |          |          |         |        |
| Mean                | 3.55555556 |
| Standard Error      | 0.37679611 |
| Median              | 4        |
| Mode                | 4        |
| Standard Deviation  | 1.13038833 |
| Sample Variance     | 1.27777778 |
| Kurtosis            | -1.17148258 |
| Skewness            | -0.1758314 |
| Range               | 3        |
| Minimum             | 2        |
| Maximum             | 5        |
| Sum                 | 32       |
| Count               | 9        |
Team Member A dropped the outlier

The argument is that the outlier was greater than 2 SD from the mean, so the 1990 data point was dropped. Is that reasonable?

A. Yes
B. No

Any special considerations about doing that?

<table>
<thead>
<tr>
<th>Year</th>
<th>Incidence (#/mil)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1910</td>
<td>2</td>
</tr>
<tr>
<td>1920</td>
<td>2</td>
</tr>
<tr>
<td>1930</td>
<td>4</td>
</tr>
<tr>
<td>1940</td>
<td>4</td>
</tr>
<tr>
<td>1950</td>
<td>3</td>
</tr>
<tr>
<td>1960</td>
<td>3</td>
</tr>
<tr>
<td>1970</td>
<td>5</td>
</tr>
<tr>
<td>1980</td>
<td>4</td>
</tr>
<tr>
<td>1990</td>
<td>10</td>
</tr>
<tr>
<td>2005</td>
<td>5</td>
</tr>
</tbody>
</table>
Thought Questions
For Your Team

What is the conclusion that your team may make about these data?

A. There is a rise in leukemia that is proportional to an increase in background electromagnetic noise
B. There is a rise in leukemia but it is not correlated to the rise in background noise
C. There is no evidence for a rise in the incidence of leukemia
D. There is insufficient data to make any conclusion
E. Other

D (48)
E (14)
C (7)
B (21)
A (10)
More Questions

Are there any circumstances whereby the outlier can be justifiably deleted?
A. Yes  
B. No

Are the results suitable for release to the NYT?
A. Yes  
B. No

Are they suitable for publication in Nature?
A. Yes  
B. No
Does it matter who funded the research? NIH? Verizon Wireless?
A. Yes
B. No

Does the suitability for publication depend on the ability of the public to understand the limitations of the data?
A. Yes
B. No

Does the team stick together on this one?
A. Yes
B. No
More Questions

1. Did the team member mishandle the dataset?
   A. Yes
   B. No

2. Are Team Member A’s actions consistent with research misconduct?
   A. Yes
   B. No

3. If your team thinks so, does it need to be reported to someone?
   A. Yes
   B. No
Key Teaching Points

• Notice key themes and RCR competencies that are addressed in this exercise
  – Importance of experimental design
    • Do we need more data? What if there’s no way to get more data?
  – (Appropriate) Statistical analysis of data is key
    • Are there outliers? What’s significant? Which analyses to do?
  – Team work/collaboration
  – When do you publish data?
    • Risk vs. reward/incentives
  – Conflicts of interest
  – Misconduct and whistleblowing
  – Social responsibility of scientists
  – AND PROBABLY EVEN MORE…
What constitutes FF for a specific type of data?

• Methodology:
  – Use a very common, specific type of data (images) and show how they can be manipulated
  – Discussion of whether the “manipulation” constitutes fabrications or falsification

• Image manipulation widespread and highly publicized
  – Can run the gamut from materials (nanoparticles) to biological samples (cells and bacteria) to historical events (photographs)
Misrepresentation of immunogold data

Rossner M, Yamada K. JCB 2004;166:11-15
Misrepresentation of image data

Rossner M, Yamada K. JCB 2004;166:11-15
Just Plain Bad Photoshopping Work

- Images from article published in Nano Lett in 2013
  - Chopstick Nanorods: Tuning the Angle between Pairs with High Yield
• Reading & discussion assignments:
  – “Why Most Published Research Findings are False”, Ioannidis, 2005
  – “Power Failure: why small sample size undermines the reliability of neuroscience”, Button et al., 2013

• Relating “data” to individual students’ actual work
  – Record keeping, notebooks, actual data sets

• Active learning on a dataset; in-class activities with iClickers

• And it is still not enough…
• It is a fallacy that our current faculty are well-trained in data management

• We cannot expect faculty in departmental courses to catch up to this water-shed moment without providing them with new instructional materials and approaches

• Federal agencies must address this deficit from a proactive respect, rather than as MiS
Old vs. New

- There is a wide gap between what was expected of RCR instruction before 2013 and what will be expected going forward
  - Stronger emphasis on data as key element to RCR instruction
  - Reproducibility may “win” over innovation

- Better information will be needed to both get grants and to publish works, so the stakes are very high
- RCR instructors have a new challenge, but the rewards likely include better translation of basic research to cures and treatments
• Kathy Partin, Ph.D.
• Carolyn Broccardo, Ph.D.; RICRO office
• Provost Rick Miranda
• CHEM601 and GRAD544 Students