Data Provenance, Data Integrity, Scientific Rigor, and Science Culture Action Plan
Department of Neurobiology, Duke University School of Medicine

The outside world is looking at how we acquire, maintain, validate, and report our research findings. There is an epidemic of mistrust of science brought about by the combination of a few instances of serious, highly-publicized scientific misconduct, and many more examples of unreliable findings evidenced by "failures to replicate." These issues have been highly publicized, and remedies for our practice and culture deserve our attention.

Data provenance and integrity ensure that the knowledge we report is supported by the primary data, and the primary data are retained in a form that allows us to be certain of the veracity of our knowledge. Scientific rigor ensures the proper application of the scientific method using the highest standards in the field. Scientific rigor is essential to conduct of the scientific enterprise.

There are multiple reasons to take action. First, science is publicly funded, and its credibility in the public's eyes is vital to continuation and expansion of funding. We must do all we can to eliminate misconduct. Second, research builds on previous reports in the pursuit of accurate knowledge. We must be confident of the truth in prior publications to further the scientific enterprise. Third, proper research standards avoid wasting time or money following up on inaccurate or erroneous reports. We must keep the scientific enterprise moving forward. Finally, science drives translational and clinical research. We must conduct and report science properly to avoid exposing patients to harmful or ineffective therapies.

All investigators must remember the reality that scientific data are inherently messy, and as a corollary, data that are too clean may have been cleaned up. Investigators must keep the adage in mind: "If it seems too good to be true, then it probably is."

We should follow four general principles:

- Know where your data are.
- Know what has been done to acquire and modify your data.
- Make all efforts to ensure that data collection and analysis are at least unbiased, blinded if possible.
- Follow proper statistical procedures.

All Departmental staff should understand why these principles exist and should be incentivized to follow the principles.

Overview of Best Practices

1. Insofar as possible given the nature of the research, best practices in scientific rigor, including statistics, should be followed.

2. In recognition that no one size fits all, each laboratory should establish its own specific plan for scientific accountability and scientific rigor, according to established standards of its field, integrating industry or other perspectives when appropriate.

3. Record keeping should track all primary data and should provide a way to "audit" the data for each figure of each paper almost instantly.

4. All modifications of raw data should be performed on copies of the original data, if possible, and should be tracked, dated, and documented fully.

5. The laboratory head should avoid allowing his/her stresses about funding and publication to affect the attitudes, expectations, or behavior of the laboratory staff.

6. Scientific accountability and scientific rigor should be a frequent discussion between the laboratory head and the laboratory staff, to establish a sense of common purpose and a shared goal to discover the truth.

Best practices in experimental design

- Employ systematic random sampling for data collection, including selection of subjects, brain area, cells, or cell parts.
- Strive to eliminate bias in experimental procedures and analysis. If practical, experimenters should be blinded to treatment. The timing of experiments might be balanced to account for sources of bias over time (e.g., evolution of surgical skills, fatigue, change in personnel; test-order effects, circadian rhythms in experimental animals).
- Use positive and negative controls.
- Use replicate samples (including both technical and biologic replicates) for experimental groups, when appropriate.
- Use validated and/or well-characterized reagents (such as antibodies and pharmacological agents), or conduct full validation.
- Consider limitations of behavioral, animal, or cellular models including possible contributions of genetic background and gender.
- Find a proper balance between increasing numbers of animals for replication and the goals of "replacement, reduction, and refinement" in animal research.
- Obtain and study the raw data for any results provided by shared research cores.

Best practices in data analysis and statistics

- Consult with a bio-statistician both before and after data collection, if statistical analysis is needed.
- Determine sample size by pre-experiment power analyses, when possible. Identify stopping points *a priori* to avoid testing to a foregone conclusion.
- Conduct a thorough characterization of experimental effects.
- Repeat experiments within the laboratory to reduce likelihood of statistical flukes.
- Use quantitative analysis, when possible, to put bounds on the size of effects.
- Use care in pooling of data across experiments done at different times, multiple time points, or different experimental groups.
- Avoid data exclusion. If it is necessary, define and report objective procedures for dealing with attrition or other missing data and data exclusion. Unless there is a compelling, transparent reason to exclude data, include all runs of each experimental procedure. This applies to exclusion of individual points or complete data sets.
- Perform theoretically-correct analysis of data using appropriate statistics and sample sizes.
- Don't mislead with statistics. Take advantage of resources that provide professional statistical expertise (e.g., the Biostatistics consultation service)
    - Perform statistical tests to validate what is seen in the data, rather than to reveal effects that may be statistically significant but functionally non-significant.
    - Select appropriate statistical tests, including testing of statistical assumptions, such as adherence of data to a normal distribution.
    - Control for multiple comparisons.
    - Avoid "significance chasing" such as interpreting the data in different ways so that it passes the statistical test of significance, or analyzing different measures until finding one on which groups differ.

Best practices in data management

- The complete primary data should be retained, backed up, and protected against alterations.
- Alterations and modifications of the primary data should be performed on copies of the data whenever possible, and should be tracked, dated, and described.
- Data notebooks should be open for viewing.
- Digital archives should be properly organized and labeled so that each dataset has the equivalent of a "unique web address".
- Every figure of every paper should be cross-referenced with the location of the original data that contributed to the figure.

- We should develop mechanisms to monitor provenance for any data that comes from shared equipment or core facilities within the Department, if and when this becomes an issue.
- The level of information security should be appropriate for the material, especially for human subject protection and PHI.
- Data should be accessible readily to all data owners, and available to outside investigators if necessary.

Best practices in publication

- Report full details on methods and experimental design, including technical and biological replicates, methods for randomization and blinding, and self-replication efforts.
- Report complete results of all analyses done as part of an experiment, including statistical controls for multiple comparisons and identification of pre- and post-hoc analyses. Methods sections should be too long, rather than too short.
- Avoid "rushing" findings into publication without a full investigation and proper self-replication.
- Target appropriate venues for publication. Avoid pressure to publish in the most glamorous journal at the expense of following the best practices for experimental design, data analysis, statistics, and publication. If a paper requires a long methods section or many figures to document the science thoroughly, do not try to compress it into a short format, no matter how "important" the results seem. Publish well-controlled negative, "uninteresting," or "not novel" results in appropriate venues.
- Resist the emerging trend where the peer review process demands additional experiments on an abbreviated timeline, with the associated pressure for results to be interpreted to conform to previously-reached conclusions.

Creating a functional and proactive scientific culture

- There must be a culture of getting it right, with the expectation of open conversation and a lack of retribution for calling results or procedures into question either within the group, or to the lab head in confidence.
- All Department staff should know that they may bring any and all concerns to the attention of the Chair or the Department's Ombudsperson in confidence, without fear of retaliation or retribution. Staff should also be aware of the Duke Integrity Line to report concerns anonymously.
- Laboratory heads must minimize incentives or pressures (or the appearance thereof) that drive their staff to perform for reasons other than pursuit of truth. It is critical to avoid the real danger that staff will respond to the laboratory head's concerns about academic promotions, choice of publication venue, or competition with other labs.

- Issues of proper scientific conduct and scientific rigor should be discussed regularly with laboratory personnel, in both private and group meetings.
- Laboratory heads should be involved in laboratory procedures, should oversee some of the actual experimental work, and should "know" how things are done in their laboratory.
- Meetings with staff should include inspection of some primary data and discussion of detailed analysis procedures, as well as discussion of final publication-style figures.

Concrete steps to be taken by the Department

1. We will continue to talk about proper scientific conduct at all levels: faculty meetings, lab meetings, and courses.
2. We will strive to create a culture where we talk about the incentives for poor conduct openly, and try to address the pressures that create those incentives.
3. We will endeavor to create a culture where people are evaluated on the basis of what they have done rather than metrics that may be weakly correlated with accomplishment.
4. Each laboratory should develop a "Data Management Standard Operating Procedure (SOP)" that will provide specific guidelines for data acquisition, storage, and transparency. The SOP should cover 4 basic components of data management:
    a. How is the data collected and stored?
    b. How are notes taken and stored?
    c. How is analysis done, tracked and, if intermediate steps are saved, stored?
    d. How are figures made and linked to both the analysis steps and the original data?
5. The laboratory's Data Management SOP should be discussed with the Chair when it has been completed, and compliance measurement will be a topic in the annual 1-on-1 meeting.
6. All research staff in all laboratories must read the Department's Action Plan and the laboratory's Data Management SOP, and sign an affirmation that they have done so.
7. The Chair will serve as a "Data Integrity Liaison" to the School of Medicine. He will advise individuals or laboratories on all of the issues covered in this plan and will work with the school's designated official for scientific integrity.
8. Faculty who are expert in data analysis are available to advise students, postdocs, and faculty on how to analyze their data.
9. The Department's System Administrator, currently Steven Happel, will work with each laboratory to implement their chosen procedures for data storage, backup, and tracking.