Who Owns Your Data?: Ethical Issues in Data Management

David B. Resnik, JD, PhD
Bioethicist, NIEHS/NIH

This presentation does not represent the views of the NIEHS, NIH or US government. This presentation discusses some legal information but is not offered as legal advice.

What is Data Management?

- Collecting and recording data
- Auditing/editing/cleaning data
- Storing an archiving data
- Analyzing data
- Interpreting data
- Sharing data
- Ownership of data

What are Data?

- Data are recorded information used to develop or test human knowledge
- Data vs. observations
- Data vs. other recorded information used in research, e.g. protocols, SOPs, correspondence, etc.
- Data vs. research materials, e.g. cell lines, reagents, transgenic animals, biological samples, etc.
- Primary (original) data vs. secondary (processed) data.

Changing Nature of Data

- Old model: lab notebooks, field notebook
- New: many forms of recording data, e.g. computer disks, CDs, audiotapes, digital images, photos, gels, notes, printouts, etc.
- Old: single researcher or small group (lab)
- New: many researchers, groups, collaboration between different labs, disciplines, institutions, sectors (private vs. public).
- These changes mean that data management is now very complex and fraught with ethical and legal peril, especially when it comes to issues of access to data and ownership of data.

Case Study: The Baltimore Affair

- Investigations by MIT, Tufts, Office of Research Integrity, and a Congressional Committee.
- At one point, federal agents seized the records.
- Front pages of NY Times.
- A DHHS appeals panel found that Imanishi-Kari was not guilty in 1996.
- She admitted only to poor record keeping.
- Baltimore described the episode as a witch hunt. Though not implicated in the scandal, he resigned his position as President of Rockefeller University.
- MORAL: KEEP GOOD RESEARCH RECORDS

Reasons to Keep Good Research Records

- To defend yourself (or your boss or colleague) against an allegation of unethical or illegal conduct
- To comply with various institutional and federal regulations, e.g. FDA, HIPAA
- To protect intellectual property
- To publish your own work
- To allow others to reproduce your results
- To be honest, ethical, responsible (bad record keeping can lead to bad results)
Good Record Keeping Practices

Research Records should be:
- Thorough and complete
- Legible
- Well organized
- Stored in a safe place
- Destroyed if required in some cases (e.g. audiotapes from interviews or focus groups)
- Backed up; archived
- Accessible to people who need access
- Signed, dated, in some situations witnessed

Other good practices
- Supervisors should review research records kept by subordinates on a regular basis
- Any changes made should be noted—don’t erase or white out
- Research projects, especially large ones, should have a record keeping plan
- Care should be taken to keep up with changes in technology so that old data can still be accessed
- Researchers should have a lab notebook or similar book for each project that can be used to organize and reference data
- Record not just data, but ideas, rationales, future plans, problems, and other pieces of useful information useful

Auditing/Editing/Cleaning data

- It is important to audit data to check for errors.
- E.g. an Excel spreadsheet to the original data or other source.
- Sometimes data are recorded improperly.
- When errors are discovered, it will be necessary to make corrections to the data in a way that is honest and ethical.

Examples:
- Someone used inconsistent notation when recording data in a spreadsheet;
- An instrument was incorrectly calibrated;
- It is difficult to hear clearly what someone is saying on a tape you are transcribing;
- Someone appears to have misinterpreted the instructions for a survey;
- Dilemmas: keep it as is, fix it, or throw it out?

Storing/Archiving Data

- Data should be retained for enough time to allow for others to answer questions that arise from the research, such as verification of results, intellectual property, or clinical or practical implications.
- 5-7 years minimum; longer in many cases.
- Keep originals, if possible.
- Many institutions archive some data.

Retention dilemmas

Is there enough space to store research records?
- Is there enough money?
- What about keeping up with outmoded technologies?
- What’s the best method of archiving data?
- Should all data be archived?
- What responsibilities do institutions or research sponsors have to provide support for storing research records?
- What should happen when someone leaves the institution? Can they take originals?
Case Study: Millikan’s oil drops

- Robert Millikan won the Nobel Prize in 1924 for measuring the charge on an electron.
- His method for measuring this charge involved dropping small, electrically-charged oil drops, between electromagnetically charged plates.
- His assistant, Harvey Fletcher, gave him the idea to use oil drop and did not receive adequate credit, but that is another story.
- In a 1913 paper on his experiments, Millikan presented 140 measurements but excluded 49 (26%). In the paper, he claimed that he had presented all the measurements, but his lab notebooks indicated that he did not. In his lab notebooks he graded the measurements according to his judgment of the quality of the result.

Data Analysis: excluding data

- Data analysis usually involves cleaning up the data and exclusion of “bad” results, such as statistical outliers and results due to human or technical error (e.g. dropped the test tube, can’t understand the interviewee’s speech, etc.).
- What is the difference between cleaning/excluding/editing data and falsifying data?
- How should you discuss data exclusion in presenting your research?

Data Outliers

![Data Outliers Diagram]

Selective Publication

<table>
<thead>
<tr>
<th>Study #</th>
<th>Results</th>
<th>P value</th>
<th>Published</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Positive</td>
<td>.04</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Negative</td>
<td>.08</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Negative</td>
<td>.11</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Positive</td>
<td>.05</td>
<td>Yes</td>
</tr>
<tr>
<td>5</td>
<td>Negative</td>
<td>.05</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>Positive</td>
<td>.04</td>
<td>Yes</td>
</tr>
<tr>
<td>7</td>
<td>Negative</td>
<td>.05</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>Positive</td>
<td>.03</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Data Analysis: missing data

- Sometimes data analysis involves dealing with missing data, i.e. data that was not recorded or edited properly.
- If some data points from a particular measurement are missing, it can affect the statistical significance of the whole analysis. How do you deal with this?
- Is it acceptable to impute data?

Total n = 100; age is missing: 7

<table>
<thead>
<tr>
<th>Pt.</th>
<th>Age</th>
<th>BP1</th>
<th>BP2</th>
<th>Yoga/wk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>135/85</td>
<td>119/75</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>142/90</td>
<td>132/82</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>38</td>
<td>137/92</td>
<td>136/93</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>139/91</td>
<td>120/80</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>51</td>
<td>145/95</td>
<td>136/90</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>148/93</td>
<td>144/94</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
<td>150/86</td>
<td>130/82</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>49</td>
<td>149/91</td>
<td>140/87</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>54</td>
<td>138/88</td>
<td>135/87</td>
<td>0</td>
</tr>
</tbody>
</table>
Federal definition of research misconduct

- Research misconduct is defined as fabrication, falsification, or plagiarism in proposing, performing, or reviewing research, or in reporting research results.
- Fabrication is making up data or results and recording or reporting them.
- Falsification is manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record.
- Plagiarism is the appropriation of another person's ideas, processes, results, or words without giving appropriate credit. Research misconduct does not include honest error or differences of opinion.
- From the Office of Science and Technology Policy.

Image manipulation

- Digital images are very important in many different research fields, such as cell biology, genetics, etc.
- Computer programs, such as Photoshop and Photosmart, make it easy to cut and paste, touch up, enhance, or manipulate digital images.
- So enhancement that does not misrepresent the results may be appropriate. But be careful!

Manipulation of images of gels


Image manipulation

- "No specific feature within an image may be enhanced, obscured, moved, removed, or introduced. The grouping of images from different parts of the same gel, or from different gels, fields, or exposures must be made explicit by the arrangement of the figure (e.g., using dividing lines) and in the text of the figure legend. Adjustments of brightness, contrast, or color balance are acceptable if they are applied to the whole image and as long as they do not obscure or eliminate any information present in the original. Nonlinear adjustments (e.g., changes to gamma settings) must be disclosed in the figure legend."
- Journal of Cell Biology

Improper use of statistics

- "There are three types of lies—lies, damned lies, and statistics." Mark Twain.
- Statistical methods only apply to problems defined by specific features of a population (normal distribution vs. skewed), sample (e.g. random or non-random), and variables (independent vs. not independent, quantitative vs. qualitative).
- Statistics software makes it very easy to inappropriately apply statistical methods.
- This can be due to ignorance of the statistics.
- If you don't understand the statistics involved in your research, consult an expert.

Improper use of statistics

- For example, suppose that you are studying the relationship between dose of a drug and excretion of a metabolite in the urine. The sample is small (25 subjects). You could use non-parametric regression or linear regression to analyze the data. You try it with both and find out that the linear regression gives you a better P-value, so you choose linear regression. You do not know whether excretion fits a normal distribution curve.
Statistics: p-value obsession

- Obsession with P-value (statistical significance) of 0.05 or less can lead people to manipulate statistical tests to lower the P-value to 0.05 level.
- There is nothing magical about 0.05; it is an arbitrary cut-off.

Combining categories

- In a study of the relationship between religiosity of adults and age, researchers analyzed the data using 10 age groupings (e.g. under 18-26, 27-34, etc.) and 5 age groupings (18-34, 35-51, 52-68, 69-85, 85 and above).
- They found a statistically significant age effect with 5 groupings but not 10.
- How should they report/discuss this?

Data mining

- Now that large datasets are available electronically, data mining is common.
- Beware of this technique. Data mining can demonstrate interesting relationships that deserve further study but it does not provide an independent test of an hypothesis.

Statistics: under-powered studies

- Beware of statistical errors (Type I) from underpowered studies: no evidence of an effect is not the same as evidence of no effect.

Statistics: graphs and charts

- Breaking news: Allen has reversed his fortune and now leads Briggs in the Polls!

Two polls of likely voters with a 3% margin of error

Statistics: using experts

- Smith, Jones, and Bartlett conduct an experiment involving the application of a new type of fertilizer to tomatoes. When they are finished with their work, they do not detect any obvious effect of the fertilizer. They ask a statistician, Lopez, to reanalyze their data prior to help them find an effect. What is wrong with this picture?
Interpreting data

- In data interpretation, one infers or proposes that scientific or practical significance or meaning or the data.
- Do the data prove or disprove a hypothesis or theory?
- Do the data help us to explain any interesting phenomena?
- Do the data have implications for medicine, public health, engineering, public policy, environmental management and other practical disciplines?

Scientists have a natural tendency to oversell their data/results.
- It’s human nature—we all tend to be enthusiastic about what we do and view it as important.
- Research sponsors, such as pharmaceutical companies, are especially prone to this. See Vioxx, etc.

Example

A pharmaceutical has sponsored clinical trials on a new medication to treat blood pressure. According to the data from the trials, the new medicine lowers blood pressure 5% more effectively than existing medicines when taken at the same dose. The medication has some adverse effects, such as the potential for liver or intestinal damage. Since the medicine is under patent, it will cost at least 10 times as much as some current medications used to treat blood pressure. Scientists, whose research has been sponsored by the company, argue that the new medication is a scientific and medical breakthrough, and they urge doctors to switch their hypertension patients to the new medicine.

Skepticism

- Defend your results but don’t fall prey to self-deception.
- Be your own worst critic.
- Be aware of the potential biases that affect your judgment.
- Consider other interpretations of the data.

Sharing Data

- Openness—the sharing of ideas, data, methods, and results—is a key ethical principle of research.
- Openness is essential for collaboration, criticism, confirmation, and replication of results.
- Many funding organizations, such as the NIH and NSF, require researchers to disseminate their results and share data (following publication).
- In the US, private citizens can gain access related to government-sponsored research through the Freedom of Information Act.

Sharing data

Even though data sharing is important, there are some legitimate reasons to not share data:
- To protect ongoing research from premature disclosure. You may not be ready to go public with your work; your reputation may be on the line. Charles Darwin conceived of his theory of natural selection while on the HMS Beagle but waited over 20 years to publish his Origin of Species.
- To protect priority—if you share your work before publication, someone else may “scoop you.” Fear of being scooped may lead some to rush to publication. Darwin published when he learned that Alfred Wallace had the same idea and might steal his thunder.
Sharing data

- To protect intellectual property. In the US, if you disclose data related to an invention prior to filing a patent application, the disclosure can invalidate the patent.
- To be patentable, an invention must be "novel," which means that it has not been disclosed previously through publication or use.
- In the US, one should file a patent before publishing a paper describing the invention.
- You may still be able to publish parts of the invention without invalidating the patent, but be careful… Talk to your Technology Transfer Office about this.

Ownership of data

- Years ago, there were many disputes over data ownership.
- Typical case: research leaves one institution and takes data with him to the next one. Researcher develops a valuable invention from the data.
- The institutions fight over ownership of the data and the patent.

- Most institutions now have policies asserting ownership of research records and/or data, and.
- Institutional officials must be able to access data for oversight purposes.
- When you leave the institution, you can take a copy of the data, not the original data, except in unusual circumstances.

Current controversies: databases

- Old model: single hypothesis, single data set, single publication. All data is published and sharing at the same time.
- New model in many fields (e.g. genomics, proteomics, neuroscience, epidemiology): large database, multiple hypotheses, multiple publications.
- Not all data may be published at the time the article is published. Supporting data may not be published.
- Some journals require deposit of supporting data on a public database.

- Dilemma: who has the right to publish articles from the supporting data?
- Academic researchers have expressed concerns about granting access to supporting data on the grounds that they don’t want other researchers to use the data for publications. They reason that since they produced the data, they should be able to mine it for publications.
- What’s fair? What’s best for science?
- If people can’t have some control over their databases, they may delay publication until they have a several publications ready to go, or they may use other methods to protect them (patents?)
Dr. Margaret Clint, a second year postdoctoral fellow in the neuroendocrinology laboratory, has just completed a series of experiments during the course of this work. Dr. Clint carried out binding assays for a receptor mutant three times. In two experiments, the data were very consistent and supported the working hypothesis that Dr. Clint and her mentor were evaluating. However, in a third independent experiment, several of the samples showed the opposite effects. Dr. Clint is supposed to present her data at the weekly meeting of her laboratory group and is now considering how to do so. In this analysis of the binding of hormone to the mutant receptor, should she average all three experiments? Should she average the two sets of data from the independent experiments that are consistent and exclude the third experiment? Or should she consider that the findings are representative of one of the experiments and state that the findings are representative of three independent determinations? What if the experiment had been repeated six times and two of the experiments showed opposite effects? In a parallel study, Dr. Clint investigated the hormonal response of several clonal cell lines transfected with receptor variants. In analyzing the data, Dr. Clint noted that a number of cell culture plates failed to respond to the hormonal stimulus and that there was considerable variability in the dose response relationship to the hormone. The data from one cell line, with each symbol representing the response of one culture plate, are provided in Figure 1.

**Problem:** it is now possible to identify particular sample of that person's genome. This technique was developed for forensic science, but could threaten privacy.

**More generally, threats to privacy will probably continue to be an issue with access to databases.**

**Turnover sharing to a private company, who will charge a fee and handle all the requests.**

**Patenting.**

Dr. Clint was also perplexed as to how to present the hormone response data shown in Figure 1. She consulted Dr. Joseph Atwood, a senior research fellow in the laboratory. Dr. Atwood responded, “Why don’t you clean up the data? Seriously, you may never get this paper published unless you do.” He then suggested that the four culture points failing to show a response (along the X-axis at approximately 10% response) be dropped because the cells were probably dead. He also pointed out that she might eliminate the top data point at the 45 minute interval as an outlier. She said, “Perhaps I should repeat a few of the experiments or try to improve the culture conditions.” “No,” said Dr. Atwood. “If you’re convinced of your results, why go through the time and expense of more repetitions?” “Somewhat dismayed, Dr. Clint thanked him and turned back to her work. What do you think about Dr. Atwood’s comments on publication practices and his suggestions for “cleaning up” the data? How should Dr. Clint go about determining which points to include and which to exclude in Figure 1? What other course(s) of action would you recommend to her? Dr. Atwood’s perception about improving the chances of publication by “cleaning up” the data is not uncommon. How might journal editors and research work toward correcting this perception?”

**Case 1**

Dr. Clint was also perplexed as to how to present the hormone response data shown in Figure 1. She consulted Dr. Joseph Atwood, a senior research fellow in the laboratory. Dr. Atwood responded, “Why don’t you clean up the data? Seriously, you may never get this paper published unless you do.” He then suggested that the four culture points failing to show a response (along the X-axis at approximately 10% response) be dropped because the cells were probably dead. He also pointed out that she might eliminate the top data point at the 45 minute interval as an outlier. She said, “Perhaps I should repeat a few of the experiments or try to improve the culture conditions.” “No,” said Dr. Atwood. “If you’re convinced of your results, why go through the time and expense of more repetitions?” “Somewhat dismayed, Dr. Clint thanked him and turned back to her work. What do you think about Dr. Atwood’s comments on publication practices and his suggestions for “cleaning up” the data? How should Dr. Clint go about determining which points to include and which to exclude in Figure 1? What other course(s) of action would you recommend to her? Dr. Atwood’s perception about improving the chances of publication by “cleaning up” the data is not uncommon. How might journal editors and research work toward correcting this perception?” One day, Dr. Clint was called into her mentor’s office and given a preliminary report of her findings for publication. Unfortunately, the abstract was due in one week. Is Dr. Clint ready to write an abstract? How should she present the data discussed above? What should Dr. Clint discuss with her mentor?

**Research tools**

- Collaborate: share data or tools and receive authorship in return. Problem: doesn’t this lead to inappropriate authorship?
- Patenting.
- Delay publication.
- Charge a fee for sharing.
- Turnover sharing to a private company, who will charge a fee and handle all the requests.
**Case 4**

Dr. Bob is a promising junior faculty member at Z University. His major clinical research project, funded by an NIH grant, is a prospective, longitudinal study of changes over time in plasma levels of protein X and their association with cardiovascular disease. Previous cross-sectional studies by others have suggested that protein X levels increase with age and are associated with increased risk of cardiovascular disease. Dr. Bob's is the first longitudinal study to address this issue. A successful study would be publishable in a high-impact journal and give a substantial boost to his achieving tenure. Dr. Miriam is a resident at the Z University Medical School. She is interested in a research career, and approaches Dr. Bob for advice. He offers her the opportunity to help analyze data from the first 3 time points of his protein X study. She eagerly accepts this offer as an opportunity to gain research experience and perhaps co-authorship on a high-impact paper. When is it appropriate for Dr. Miriam to discuss her authorship status with Dr. Bob? Should she raise the issue now, before agreeing to analyze the data, or wait until after the results are known?

---

**Case 4**

When Dr. Miriam reanalyzed the data, the hypothesis was confirmed. However, she was puzzled that correction of “a few mistaken data entries” would so substantially change the outcome of the analysis. She compared the “corrected” spreadsheet with the study’s case report forms and found that the majority of data entries had been changed, always in the direction consistent with the hypothesis. Is it appropriate for Dr. Miriam to check the new spreadsheet against the case report forms? Should she have accepted Dr. Bob’s corrections and confined herself to the re-analysis? Under what circumstances would one check a transcribed or secondary data set against the primary or source data?

---

**Case 4**

Dr. Miriam presented the data discrepancies to Dr. Bob and asked to see the original patient files. Dr. Bob brushed this off as unnecessary. He blamed the apparent discrepancies on his own inexperience with Excel and on his use of imputed data (i.e., deriving some data entries from a statistical model, rather than actual measurements). Concerned about the situation, Dr. Miriam began reviewing patient records on her own without telling anyone. To her dismay, she found that many data entries in the spreadsheet had been changed from their true values, that some data entries did not correspond to actual measurements, and that some patients had been infected by a virus. How should Dr. Miriam deal with this issue?
Dr. Miriam continued to work with Dr. Bob while she searched for a new mentor, but did not tell him of her findings. She did share her concerns with one of Dr. Bob’s former fellows and with a collaborating faculty member in his department. Dr. Bob learned of Dr. Miriam’s questioning of his scientific integrity and stopped working with her. In response, Dr. Miriam lodged a formal complaint of scientific misconduct against Dr. Bob with the university.

Should Dr. Miriam have shared her concerns with others without first talking with Dr. Bob or lodging a formal complaint? What other steps could she have taken before lodging a complaint? When would have been the best time to lodge a formal complaint of scientific misconduct?