Introduction: Differentiating between Honest Discourse and Research Misconduct

(*When is it research misconduct versus honest scientific difference of opinion?*)

During just the past couple of years within the NIH Intramural Research Program, nine cases have required formal Inquiry committee examination for scientific misconduct, and four cases have already proceeded to full-scale Investigations. NIH staff engaged in this very stressful, time-consuming process included two tenured investigators, one tenure-track researcher, eight trainees, and a support staff member. Concerns raised about their research have included deliberate falsification of data, data manipulation, misrepresentation of findings, and authorship issues (which are not research misconduct). This is alarming.

Recently a finding of misconduct in the extramural community resulted in a 366-day Federal prison term for a scientist because his actions led to loss of government funds, obstruction of justice, and abuse of a position of trust. The sentenced scientist had the following explanation for his actions:

"First, I believed that because the research questions I had framed were legitimate and worthy of study, it was okay to misrepresent "minor" pieces of data to increase the odds that the grant would be awarded to UVM and the work I proposed could be done. Second, the structure at UVM created pressures which I should have, but was not able to, stand up to. Being an academic in a medical school setting, I saw my job and my laboratory as expendable if I were not able to produce. Many aspects of my laboratory, including salaries of the technicians and lab workers, depended on my ability to obtain grants for the university. I convinced myself that the responsibility I felt for these individuals, the stress associated with that responsibility, and my passion and personal ambition justified "cutting corners". Third, I cannot deny that I was also motivated by my own desire to advance as a respected scientist because I wanted to be recognized as an important contributor in a field I was committed to." Underlying this case was the issue of inappropriate data management, which was detected by one of the scientist's staff. He admitted to destruction of electronic evidence of his falsifications and fabrications, among other things.

Scientific misconduct is detrimental to all parties involved. Everyone in a lab has a responsibility to be informed and vigilant about appropriate data management to prevent instances of scientific misconduct. It is also important, however, to distinguish between misconduct, bad behavior, and honest differences in opinion.

Some of the following scenarios are based on actual misconduct cases. Choose to present 2-3 cases from 1-4 and one of the Research Reproducibility cases.

Case #1 – Handling of Images and Graphs

Case #2 – A Technically Challenging Method Collides With a Hot Topic

Case #3 – Handling of Clinical Data

Case #4 – Sources of Potential Bias and Data Sharing

Case #5 – Research Reproducibility I: Sample Composition and Reproducibility

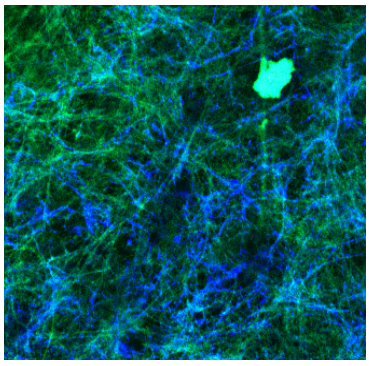
Case #6 – Research Reproducibility II: Prostate Cancer Serum Biomarker Study

Case #1 – Handling of Images and Graphs

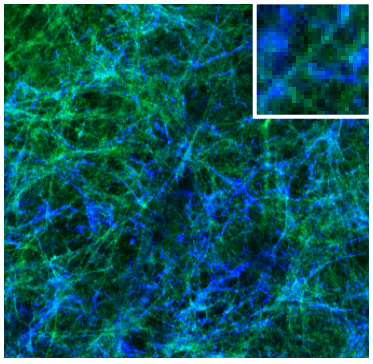
Dr. Gomez is preparing a manuscript for submission to a prominent journal and is trying to decide the best way to present her image and gel data. Other postdocs in the lab tell her that her results will have to look "clean" to be able to impress the editors and reviewers. She comes to you for advice about the following potential figures.

Imaging data

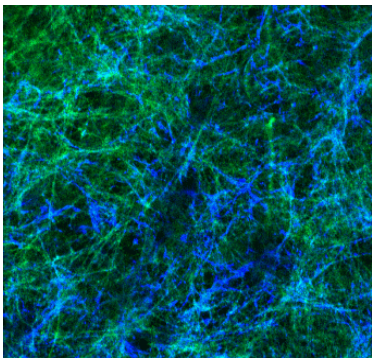
She complains that the best fluorescence images of her protein called "excitin" often have an unexplained bright blob of material that looks like junk and will be distracting to readers. She debates what to do, including covering it up with an inset, fixing the problem by masking the junk using the "clone" function in Photoshop, or by cropping the picture.



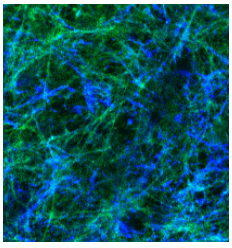
Original with "junk"



Covered up by inset



"Fixed" with Photoshop



Cropped out

What is your advice?

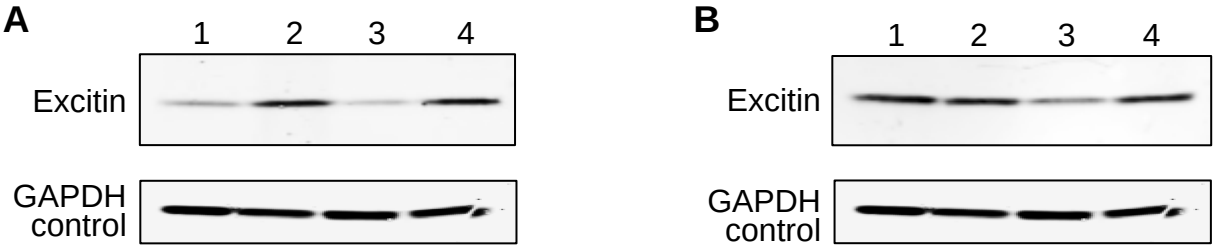
Does any approach constitute research misconduct?

What are the ethical boundaries of what data you show, and what is a "representative" image or other form of data?

Might something be missed by omitting "junk" from figures?

Gels and controls

Dr. Gomez receives the following gel images from Dr. Brown showing changes in excitin expression with different drug treatments.



Do you notice anything strange about panels A and B?

Is this permissible?

Is there any concern about showing a single part of a gel, i.e., only showing the band of interest?

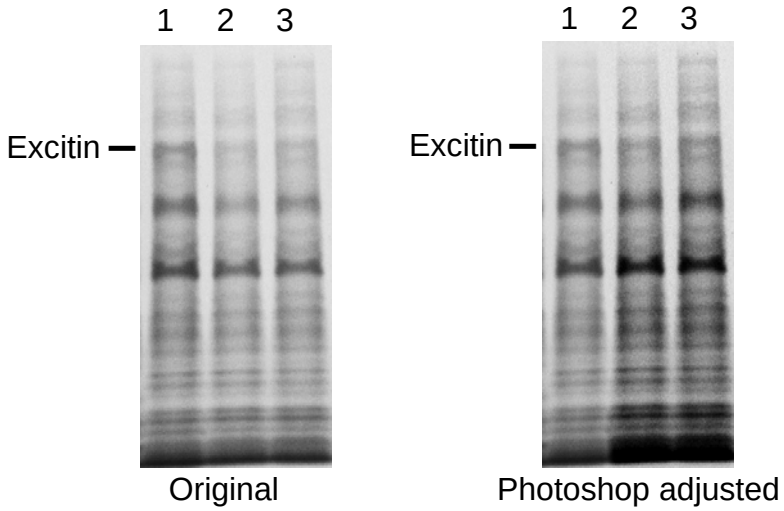
Dr. Gomez saw the same pattern of reduced excitin in lanes 1 and 3 shown in panel A in two experiments, but not in a third repeat (panel B).

Can she present just the results shown in panel A?

How should you deal with experiments that "work" sometimes but not always?

Gels – brightness/contrast

Her next question to you involves her gels, where she thinks she probably accidentally loaded less into lanes 2 and 3. Dr. Brown tells her that she should just adjust the darkness of these lanes to look equal. He says this is permissible because it is involves changing the darkness of the entire lane, not just one band.

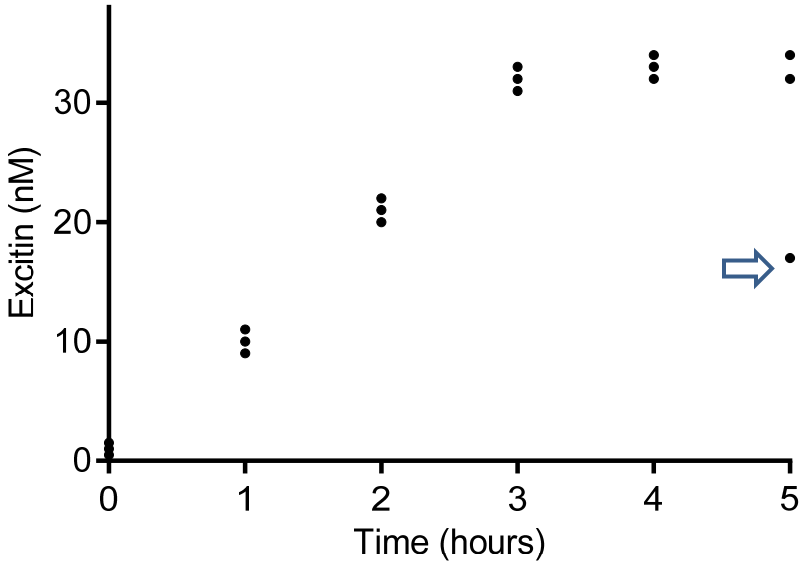


Is this change acceptable?

Why or why not?

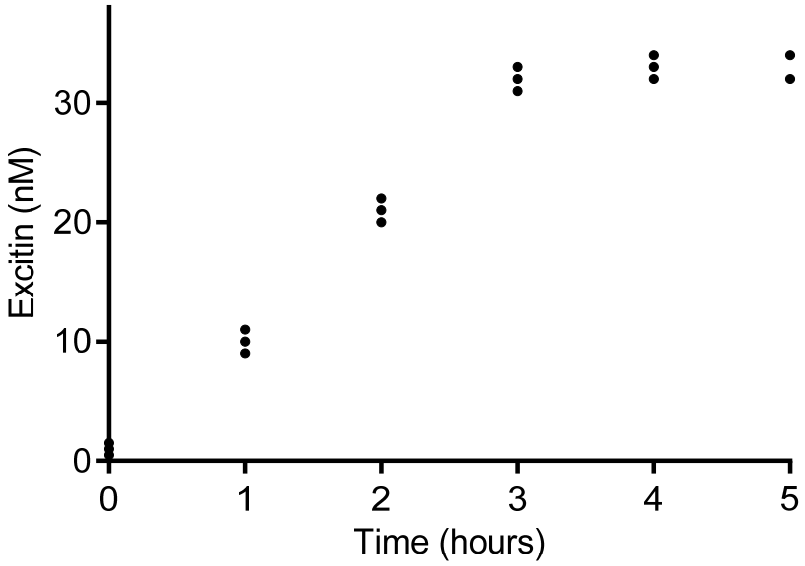
Data in graphs

Her collaborator, Dr. Blue, provides the following graph, saying that his supervisor advises that they should delete the single point that doesn't fit with the rest.



Can they omit the point because it is an obvious outlier?

What if Dr. Blue re-analyzes that specific data point with remaining sample and finds that the first analysis was in error – can he modify the figure to reflect the re-analysis?



They decide to just eliminate the outlier. Does anything else bother you about this graph?

If you were a lab colleague, what would you do?

What would you do if you were a journal editor, and a reviewer raised a concern about this figure?

Case #2 – A Technically Challenging Method Collides With a Hot Topic

(This case is based on an actual publication issue. The facts have been streamlined to highlight the ethical issues, rather than the science itself.)

"Cryo electron microscopy" or "Cryo-EM" is a method for determining macromolecular structures. The technique has been evolving for a couple of decades. Currently it is most suited for large particles such as ribosomes, proteasomes and viruses. Recent advances in sensitive electron detection, advanced computers for data collection, and application of estimates of expected structures have converged to produce a number of fascinating publications. Unfortunately, the controversial nature of some findings has touched off concern over journal peer-review practices. One recent publication of the structure of a viral surface glycoprotein complex has triggered an interesting set of articles about the method, its utilization for certain structural investigations, and the processes for data review by journal referees.

Part 1: Investigators determine the structure of the complex using cryo-EM and submit it to a journal. Peer reviewers, experts in cryo-EM, reject the paper on technical grounds. The authors then submit the same manuscript, virtually unchanged, to a second journal where – as luck would have it – one of the reviewers is the same expert who had seen the copy the first time. This time, he demands to see all of the original data before rendering his opinion. The authors supply some of the data but not all of it as he requested. Based on the information in the manuscript and the supplemental data to be published and upon the additional information provided by the authors, the reviewer again rejects the paper – this time providing extensive scientific and technical criticisms of how the authors' conclusions could be erroneous.

1. What do you think about the authors' decision to submit the same manuscript to a different journal without alteration? Should they have been required to inform the second journal that this paper had been rejected elsewhere?
2. Was the reviewer justified in requesting the complete data set? The reviewer agreed to maintain confidentiality of the data set; under these circumstances should the authors have agreed to provide the entire set?
3. Should this expert have refused to be a reviewer the second time around? Did he have an obligation to inform the editors that this paper had been submitted elsewhere?

Part 2: The authors submit the paper to journal number three, this time arranging for a selected set of reviewers. The paper is accepted and appears in print. Several experts in cryo-EM subsequently write letters to the journal that give substantial and technical criticisms of the paper and the application of the method as done by the authors. One critic [who turns out to have been the original reviewer] points out that incorrect application of appropriate controls can even allow investigators to deduce structures from the data where only random noise is actually present. In a response letter, the authors provide a detailed rebuttal to all of the criticisms and stand behind their findings.

4. What responsibilities do editors have to weigh standards of scholarship against the risk of losing out on publication of hot papers?
5. What other steps should the authors have taken in response to criticisms? Are they under any responsibility to withdraw the paper?
6. The experts did not attempt to repeat the experiments themselves. Rather, they performed re-analysis using the authors' published data. Should criticism be accompanied by attempts by peers to carry out the exact same experiments?

Part 3: Your journal club discusses all of these papers.

7. What lessons should your postdoctoral trainees learn from this series of events?
8. What obligations do authors, editors and reviewers have to ensure that adequate expertise is available when complex methodology is used and evaluated?

Case #3 – Handling of Clinical Data

Dr. Bob is a promising mid-career faculty member at Z University. His major clinical research project 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 suggested that protein X levels increase with age and are associated with increased risk of cardiovascular disease. A successful longitudinal study would be publishable in a high-impact journal and give a substantial boost to his achieving tenure.

Dr. Miriam, a resident at the Z University Medical School, approaches Dr. Bob for advice about a research career and he offers to let her 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?

Dr. Miriam performs a statistical analysis on a spreadsheet provided by Dr. Bob, but her results are not consistent with the hypothesis Dr. Bob wrote in his grant applications as she found no association of Protein X with cardiovascular risk. When Dr. Miriam presents her analysis to Dr. Bob, he is noncommittal and suggests that she has incorrectly analyzed the data. He says he will check her work and the next week, Dr. Bob returns the spreadsheet to Dr. Miriam, explaining that he has corrected a few mistaken data entries. He asks her to redo the analysis.

- Should Dr. Miriam ask for an explanation of the data corrections? Would it make a difference if he used his home computer or his work laptop for checking her work?

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 multiple 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 (note that she did not actually participate in the trial)? Should she have just confined herself to the reanalysis given that she was not named in the trial protocol as a participant? Under what circumstances and who should have access to check a transcribed or secondary data set against the primary or source data?

When Dr. Miriam presented the data discrepancies to Dr. Bob, he blamed the apparent discrepancies on his own ineptitude with Excel and on his use of data imputed from statistical modeling, rather than actual measurements. Concerned about the situation, Dr. Miriam began secretly reviewing patient records. She found that many data entries in the spreadsheet had been changed from their original values and that some patients recorded as participating in the study did not actually exist. Based on her analysis, she began to consider lodging a formal complaint of scientific misconduct against Dr. Bob.

- Is Dr. Bob's explanation of the data discrepancies justifiable? Would it be appropriate for Dr. Miriam to independently discuss the general principles of the case with a bio-statistician for further insight?
- Is it appropriate in this context for Dr. Miriam to access patient records? Should she first have shared her concerns with someone in authority and gotten permission? Does this situation represent scientific misconduct? If so, what type of misconduct is it?
- Should Dr. Miriam have immediately lodged a formal complaint upon finding data altered?
- 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?

Case #4 – Sources of Potential Bias and Data Sharing

Dr. Whitaker is the principal investigator for a retrospective, case-control study examining the relationship between cell phone use and three types of brain cancer (primary glioma, meningioma, or acoustic neurinoma) funded by the NIH. The study includes 1000 cases of brain cancer from the U.S., Canada, U.K., Germany, and France matched with 1000 controls from the same countries. The study asked both cases and controls to recall their cell phone use for a fifteen-year period and collected data on other risk factors, such as medical history, family cancer history, smoking, diet, and age. After analyzing the data, Dr. Whitaker found that cell phone use was associated with a 25% increased risk of brain cancer. Dr. Whitaker published his results in a top-tier epidemiology journal. One month after publication of the article, the editors of the journal informed Dr. Whitaker that they were planning on publishing a commentary critiquing the article's methodology. The editors want to give Dr. Whitaker an opportunity to respond to the letter in the same journal issue. The commentary cited a study published last year demonstrating systematic bias in recollection of cell phone use. The study showed that cases tend to overestimate their cell phone use, which would tend to bias research in favor of an association between cell phone use and brain cancer. The biasing effect increased with recollection time and was higher in European countries. The authors argued that if Dr. Whitaker's article had taken these factors into account it would have shown no significant association between cell phone use and brain cancer. The commentary was funded by the cell phone industry. The authors of the commentary contacted Dr. Whitaker and asked her to provide them with the original data from her study, so they could reanalyze it.

- How should Dr. Whitaker respond to these critiques?

- Should Dr. Whitaker provide any of the authors with original data?

- Should Dr. Whitaker have anticipated these possible critiques in developing the study design and in analyzing and interpreting the data?

- What role [if any] should the journal's original anonymous peer-reviewers play in responding to the critique? They were the ones who had approved the paper in the first place. Doesn't the journal's editorial board have a role in dealing with disputes arising from their published articles?

- If the question about the validity of the survey had been raised by scientists NOT funded by the cell-phone industry, what difference would that make to your answers to these questions?

Introduction to Enhancing Reproducibility, Cases 2014

Considerable attention has been focused on the inability of scientists and corporations to reproduce results of pre-clinical studies, especially those using animal models. Several papers[1], including a Nature Commentary by Drs. Collins and Tabak[2], expressed concern about this irreproducibility. NIH has been exploring issues affecting reproducibility and ways to improve scientific fidelity. In their Commentary, they announced that NIH will be taking the lead in developing a training module on enhancing reproducibility and transparency of research results emphasizing experimental design. This module should be ready within the year for testing. The following case studies for 2014 preview these ideas.

These case studies cover examples of specific areas of concern, which include

1. Deficiencies in reporting and bias
2. The importance of blinding and randomization
3. Defining exclusion criteria and how to handle 'outliers'
4. Determining correct sample size to reduce chance observations

[1] http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0007824
http://www.bwfund.org/newsroom/newsletter-articles/special-report-biomedical-research-are-all-results-correct
Special Report on Reproducibility
http://www.nature.com/nature/journal/v490/n7419/full/nature11556.html

[2] http://www.nature.com/news/policy-nih-plans-to-enhance-reproducibility-1.14586

Case #5 – Reproducibility I: Sample Composition and Reproducibility

Dr. Sanchez is a respected researcher who studies a fatal human neurodegenerative disease for which there is no effective treatment. He suddenly realizes that data from a cell culture model predict that a commonly prescribed cancer drug could prevent neuronal cell death. Dr. Sanchez asks his postdoc Dr. Fisher to test this hypothesis in a transgenic mouse model of the disease. Based on prior experience, they inject 6 animals with the drug and 6 animals with a vehicle control. Preliminary results are promising for these young mice: there is a 'trend' ($p=0.06$) showing 10% better neurological function in the treatment group. Based on this, Dr. Fisher injects another three animals with the drug and is happy to see that the results are now statistically significant ($p<0.05$). Although the normal phenotype was not fully rescued, they both agree that even a 10% improvement in patient function could be clinically significant, and the findings should be submitted for publication in a top-tier journal.

1. How should the animal group size and composition be determined? Is it legitimate to add more animals after the first group is analyzed? When should a study be repeated?
2. Would your opinion of the work change if you knew that the testing of neurological function was performed by the same person who injected each animal? What if the animals were assigned randomly to the control or treatment groups?
3. What if there had been 90% improvement in neurological function?

The journal sends the manuscript to two referees. Dr. Williams rates the paper highly and recommends accelerated publication because of the clinical significance. Dr. Johnson is unenthusiastic and concerned that the manuscript does not mention sample size estimates, randomization, blinding, or repeating. He also notes that the experiment used young animals, whereas the human disease occurs later in life. His review states: "To extrapolate these results to the clinical setting, the authors should replicate them, such as in a more reasonable model of the disease (i.e., in older animals.)"

Fisher and Sanchez immediately test aged mice, but to their surprise see no statistically significant effect ($p=0.35$). Nonetheless, because the initial results were so important, and because they included mechanistic in vitro experiments supporting the original hypothesis in the revised manuscript, they add a statement to the discussion: "Even though further preclinical development of the drug seems warranted, some caution may be needed because the effects in older animals are more modest (data not shown)." The journal editor is convinced by telephone of the importance of the study and accepts the article despite Dr. Johnson's remaining concerns.

1. At what point does failure to replicate an experiment become a concern?
2. How well can peer reviewing of articles address problems of reproducibility?

After publication, the article garners a great deal of attention. Several independent groups, including the company that markets the drug, try to replicate the initial finding, but all fail to show that the compound prevents neurodegeneration in animal models. A consortium of researchers asks the journal to publish a second paper refuting the study, but it declines to because the consortium "cannot present a new conceptual advance beyond negative data."

1. What mechanisms exist (or should exist) for publishing data that raise serious doubts about the validity of a published study?
2. Would having the raw data available in the original manuscript have altered the outcome?
3. Is there anything special about $p<0.05$?

<u>Useful references:</u>
Scott et al., Amyotroph Lateral Scler 2008; 9: 4-15
Simmons et al., Psychological Science 2011; 22: 1359-1366
Sullivan and Feinn, J Grad Med Educ 2012; 4: 279-282

Case #6 – Reproducibility II: Prostate Cancer Serum Biomarker Study

Dr. Simmons is an oncologist at the NIH Clinical Center whose translational research is focused on prostate cancer. In addition to seeing patients enrolled in his two active clinical trials, he has pursued studies of prostate cancer early detection biomarkers in collaboration with the mass-spectrometry (MS) laboratory of Dr. Wallace in NCI-Frederick in order to develop a screening test that performs better than the commonly used prostate-specific antigen (PSA) test. His most recent study includes a clinical series of 120 prostate cancer cases from which he has collected and stored pre-protocol fasting serum over the past 3 years. These samples were collected from the NIH Clinical Center, Howard University Medical Center and the University of Maryland Medical Center in Baltimore. Dr. Simmons brings the cancer case samples (previously aliquotted on four 30-well plates) on dry-ice with him to the monthly Wallace lab meeting, during which one of the post-doctoral fellows describes quality control (QC) and other findings from their new, highly sensitive ultra-high performance LC-MS / GC-MS metabolomic platform capable of identifying over 800 serum metabolites. After the meeting, they discuss the new study, and Wallace mentions that she has serum samples from 120 control subjects as well as additional serum QC duplicate samples ready for assay, and the set of four cancer case plates and five control subjects/QC plates are run the following week with excellent reproducibility within the duplicate serum QC samples. Multivariable analysis shows a significantly different metabolic pattern in serum from the prostate cancer cases compared to the control subjects ($P<0.0001$), with substantially higher pyruvate and acetoacetate concentrations in the cancer cases.

1. What are the implications of having only cancer case serum on some plates, with control and QC samples on other plates?

2. Was the manner in which control subjects selected appropriate? What if women were included?

3. Are there potential biases from having cancer cases enrolled in three different clinics?

Drs. Simmons and Wallace are excited by the findings, and calculate that the sensitivity and specificity for both compounds are 95-99%, which is far better than that reported for serum PSA by most investigators. They may have discovered a new prostate cancer screening test. They hurriedly draft the manuscript, obtain NCI clearance, and send the report to the New England Journal of Medicine (NEJM). Despite generally positive reviews the manuscript is returned to them for revision. The most critical comment is from Referee #1 who questions the biological plausibility of prostate tumors causing an elevation in serum pyruvate and acetoacetate, and asks that a more detailed description of sample collection, handling, and storage be added to the methods. Dr. Simmons asks his clinical fellow to track down the information from the Wallace lab and check the literature for any information relevant to their results. The fellow reports back that the serum from the control subjects were collected appropriately and frozen at -80oC in 7.5 ml aliquots immediately after their collection 5 years ago, but that they were thawed and re-aliquotted into 1 ml vials 2-3 years later. Dr. Wallace cannot find documentation in her lab regarding who re-aliquotted the samples and how it was done, or any information about the control subjects (e.g., age, gender, or fasting status). At the same time, the clinical fellow finds a recent article describing degradation of several blood metabolites following multiple thaw-refreeze cycles.

1. After discussing their data, Wallace and Simmons are unsure as to whether the sample handling and storage are responsible for the metabolite difference in their data. Is there anything they can do to address this?
2. Who should have been responsible for lab documentation of the control subjects' serum processing?
3. Having reviewed their data, Wallace and Simmons decide to withdraw their manuscript from the NEJM. What can they do to make their data and report acceptable to another journal?

Comments and Guidelines for NIH Ethics Cases 2014

- The honest and accurate presentation of scientific findings is the most important thing a scientist can do. Illustrations must provide an accurate representation of the data obtained.

- Many recent cases of scientific misconduct in both the intramural and extramural communities involve inappropriate data manipulation using programs (such as Photoshop) or inappropriate statistical analysis. As a result, journals now analyze images to detect inappropriate manipulations and often obtain separate statistical reviews of submissions.

- Changes in brightness, contrast, etc. should be applied simultaneously to all panels in a figure, including positive and negative controls. Parts of images or graphical data should not be arbitrarily modified. For digital images, the original data file must always be kept, with its original name (as recorded in a notebook); subsequent modified versions, and versions finalized for publication must be maintained as separate files.

- For safety, two copies/versions of data should be kept (e.g., original + figure version, two hard copies, hard copy + scan, computer file + backup, etc.).

- When a new technique is introduced into a laboratory, it should be validated by rigorous positive and negative controls.

- When experiments do not "work" every time, more science can often be learned by thorough trouble-shooting than by just repeating the experiment. Controls should always be part of the repeat experiments, because they will tell you something about outliers, loading differences, etc.

- Using an appropriate number of experimental animals is important both for statistical considerations and for the ethical implications of animal usage. Power calculations are required in human clinical studies and can be helpful in guiding animal experimental designs, including by having the investigator think about statistical approach ahead of time. Adding new animals or samples merely to reach an arbitrary level of statistical significance can be risky and lead to false-positive findings.

- Carefully consider and report choice of sample size, and when appropriate, the use of randomization, blinding, numbers of repeat experiments, any exclusions, and failures of replication. Many journals now require such reporting explicitly and it is good practice to gather this information before submission so it can be critically evaluated by the submitting laboratory before being even more critically reviewed by an external referee.

- Mechanisms are needed to publish data that raise serious doubts about previously published studies. It is always appropriate to have differences of opinion on data interpretation, and being wrong is not an ethical issue! However, having robust data and access to all datasets including those where data doesn't replicate is important to be able to have those discussions.

- The scientific integrity and credibility of clinical trial data depend on sound trial designs, with clearly identified primary and secondary endpoints and a description of statistical methods to be employed. This is a requirement for clinical studies under the jurisdiction of the FDA.

- Appropriate case and control subjects should be carefully selected in order to avoid potential bias, and their biospecimens (e.g., blood) should be collected, processed, stored, and assayed similarly.

- Lab notebooks should be thoroughly documented, and methods sections in manuscripts should provide detail sufficient to permit replication of the study.

- Institute/Center clearance is required for all NIH intramural manuscripts.

==========