More Data, Please!

Keith Baggerly1*

In this issue of Clinical Chemistry, Witwer (1) surveys the state of data reporting in microRNA (miRNA)2 studies and finds it wanting. I echo his concerns and agree the problems can and should be addressed. Data-reporting problems are affecting a number of areas beyond miRNA studies. Here I try to cast some of the issues (and solutions) in broader context.

Incomplete Data Reporting Is a Broad Problem

Witwer (1) observes in his review of miRNA studies that “data submission was reported at publication for <40% of all articles.” Sadly, the problem of poor data reporting is neither new nor restricted to miRNA data. Poor access to data motivated the 2001 MIAME (minimum information about a microarray experiment) standards in the first place (2). In 2008, Ochsner et al. (3) reviewed microarray studies from 2007 across 20 journals and found “the rate of deposition of datasets was <50%.” Related concerns motivated the 2011 editorial by Baggerly and Coombes (4).

Incomplete Reporting Precludes Reproduction and Replication

In considering the implications, I distinguish between reproduction (derivation of the reported results from the same raw data) and replication (obtaining qualitatively similar results from new experiments). Ideally, reproduction (which should be faster and cheaper) should precede replication as a sanity check. Poor data access hinders both. Even when data are supplied, reproducibility should not be presumed; in their survey of 18 microarray studies, Ioannidis et al. (5) were able to access data for 10 studies but could reproduce quantitative results for just 2.

Given this poor rate of reproduction, poor replication rates, such as the rate of 6 of 53 reported by Begley and Ellis (6), for even “landmark” studies are not a huge surprise.

It is not clear to what extent the lack of data and documentation is driven by a “reluctance to share data,” as opposed to a lack of appreciation for how difficult reproduction and replication can be. In practical terms, the point is moot, but below I touch on motivation for those in the latter camp.

Reproduction Is Needed Since Intuition May Be Lacking

If we do not have good intuition about whether the results “make sense,” we need to be able to check the process by which the results were derived. Experience has shown that it is necessary, particularly when results involve fairly new biological entities (e.g., miRNA) or high-dimensional data (e.g., microarrays) and thus intuition is empirically lacking, even when we think intuition is present. As illustrations, Baggerly and Coombes (7) gave examples of microarray studies in which gene lists were supplied and “explained” in light of the hypothesized biology, despite the fact that the lists referred to the wrong genes owing to a simple indexing error. Furthermore, as data complexities increase, so do the numbers of processing steps that can appreciably affect the results. Consequently, “a detailed account of preprocessing and normalization” (1) means we need both the data and the code for reproduction to succeed.

The Implications Can Be Severe

Poor reproducibility in the “omics” context has already allowed inaccurate genomic signatures to progress to guiding care in clinical trials (7). Furthermore, lack of clarity with respect to the required level of reproducibility let the trials resume even after questions were raised. The clinical trials were terminated only after sustained effort to clarify the science and the introduction of extraneous, albeit catalyzing, information (CV fabrication).

This instance prompted an Institute of Medicine (IOM) review of the level of evidence that should be required before omics-based signatures are used in a clinical context. The IOM report, published in March 2012 (8), also calls for the depositing of data, code, and annotation.

1 University of Texas MD Anderson Cancer Center, Houston, TX.
2 Address correspondence to the author at: University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Box 447, Houston, TX 77030-4009. E-mail kabagg@mdanderson.org.
3 Received December 27, 2012; accepted January 3, 2013.
4 Previously published online at DOI: 10.1373/clinchem.2012.200501
5 Nonstandard abbreviations: miRNA, microRNA; MIAME, minimum information about a microarray experiment; IOM, Institute of Medicine; REMARK, reporting recommendations for tumor marker prognostic studies; FDA, US Food and Drug Administration; NCI, National Cancer Institute.

The latest version is at Papers in Press. Published January 22, 2013 as doi:10.1373/clinchem.2012.200501
Copyright (C) 2013 by The American Association for Clinical Chemistry
The Problem Is Fixable

The problem is not insoluble! Indeed, we already have good guidelines for many aspects of reporting, including both MIAME (2) and REMARK (reporting recommendations for tumor marker prognostic studies) (9). The latter in particular discusses types of information an investigator should be prepared to report, and why.

The IOM report also looks at “how to do it right” and provides detailed discussion, not just of data reporting but also of issues to be considered at all stages of the process of moving a biomarker to the clinic: from discovery (when anything goes, as long as a “locked down” rule emerges at the end) to validation (when the locked-down rule is subjected to analytical, clinical, and biological validation) to evaluation (when the rule is used to assess performance in new patient samples).

The Tools Are There

Another reason for optimism in addressing the problem is that increasingly better tools for documentation, largely stolen from the open-source community, are being introduced, “lowering the bar” as it were. In addition to the tools mentioned by Baggerly and Coombes (4), I draw particular attention to Markdown (10), which allows the blending of code with an e-mail–like syntax. I am increasingly using these tools in writing my own reports.

The Time Is Now

As Witwer notes in his introduction, issues of (ir)reproducibility are drawing increasing attention. In addition to the principles discussed by Begley and Ellis (6) and the IOM report (8), the rate of retractions in the scientific literature has been increasing, as ably documented by the blog Retraction Watch (11). These have recently included some egregious cases of fraud (e.g., Diederik Stapel, Anil Potti) in which access to data contributed to the problems, leading to soul-searching and negative attention at a time of tight funding. The US Food and Drug Administration (FDA) is watching closely. In October, the National Cancer Institute (NCI) posted guidelines (12) to better ensure both reproducibility and replicability, which researchers should be prepared to address before getting NCI funds to run a clinical trial. Reproducibility is becoming an issue of self-interest as well as altruism.

Some Further Cautions (Outsourcing, Common Errors)

In closing, I touch on 2 somewhat tangentially related points. Witwer (Discussion, points 6 and 7) alludes to potential issues with outsourcing analyses and inadequate statistical input. Just as I hope to see data and code from a report’s authors, investigators should expect data and code with any outsourced analysis. I echo his call for clear integration of an analysis, particularly as it relates to (a) experimental design and (b) validation.

With respect to experimental design, complete confounding (e.g., running all “group 1” samples in March and all “group 2” samples in June) is distressingly common and can easily lead to experimental noise being mistaken for biological signal. Baggerly et al. (13) discussed in detail an instance involving proteomics. In this case, a “home brew” test was being marketed before the FDA intervened. The reason this issue is of such concern is that systemic “batch effects” are ubiquitous. Leek et al. (14) surveyed a wide range of high-throughput assays and found examples of major distortions affecting each one. This problem will not go away—if an assay is sensitive enough to detect subtle changes in analyte concentrations, it is sensitive enough to detect changes in reagent lots—but proper design can allow for it.

With respect to validation, I add a caution to Witwer’s discussion pertaining to improper approaches: “validating” the performance of a model with the data used to construct the model, or fitting a model to all the data and then using those coefficients in an approximate “cross-validation.” Dupuy and Simon (15) surveyed a large cohort of microarray reports with respect to the statistical methods used and found such validation errors to be frequent. With high-dimensional data, it is surprisingly easy to find patterns even in random noise.

In both of these cases (and the others discussed above), being able to reproduce the analysis (and to see what, if anything, might have gone wrong) both increases our confidence and lets us progress more quickly.

Author Contributions: All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.

Authors’ Disclosures or Potential Conflicts of Interest: Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: NIH grant P50 CA083639-09.
Expert Testimony: None declared.
Patents: None declared.
References