Experimental Design, Randomization, and Validation

Keith Baggerly[1*]

Featured Article: Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 2004;20:777–85.[2]

Before our article discussed here appeared in 2004, one might have thought a golden age in proteomic diagnostics was at hand. The good news began in 2002 *(1)*, when National Cancer Institute (NCI)[3]/Food and Drug Administration (FDA) researchers claimed to have processed high-throughput measurements of easy-to-get samples with machine learning algorithms to extract previously elusive diagnoses.

The high-throughput measurements were from mass spectrometry, specifically surface-enhanced laser desorption and ionization (SELDI) assays. The samples were minimally processed serum. The machine learning algorithms were, broadly, black boxes taking peak intensities (nominally peptide abundances) and producing categorical calls: "disease," "no disease," or "something else." The diagnosis was whether a woman had ovarian cancer.

The numbers looked impressive. Starting with 216 samples (100 women with ovarian cancer, 100 healthy controls, and 16 women with benign disease), the authors trained on 50 cancer and 50 control spectra and predicted outcomes for the rest. They reported 100% (50 of 50) clinical sensitivity, 92% (42 of 50) clinical specificity, and called all 16 of 16 benign disease cases "something else," suggesting specificity to ovarian cancer. The authors posted data from a first experiment (DS1), a follow-up experiment (DS2) using the same samples but a different SELDI surface (WX2 vs HE4), and another experiment (DS3) using new samples (161 cancer cases, 92 controls; WX2). They reported great results throughout. Some concerns were raised that the results were biologically implausible, but the broader field forged ahead.

Everyone wanted in, including my own institution (MD Anderson). We wanted to treat our patients better. Our group was tasked with exploring the raw data to optimize the algorithm.

We could not get it to work. We kept finding patterns other than those reported, and these new patterns drove the story. The DS1 benign disease samples stood out from other DS1 samples because they looked like DS2 samples. The DS1 "something else" calls were not driven by different biology; they were driven by spectral drift. Sample run order was not randomized, and 1 time batch perfectly intersected 1 outcome group ("complete confounding"). We could perfectly classify DS3 using "electronic noise" spectral regions (complete confounding again). Patterns giving "great results" in individual data sets failed utterly when applied across data sets. New models were fit every time, as opposed to fixing 1 model early and using later samples for prospective validation.

The imminent arrival of a home-brew diagnostic test, OvaCheck, was announced at the annual meeting of the Society of Gynecologic Oncologists in 2004. Basically, the claim was "you send us a serum sample, we will tell you whether this woman needs her ovaries removed." We originally submitted our article discussed here to *The Lancet*; it was rejected as too technical. Our article appeared in *Bioinformatics* in January 2004. One week later, *The New York Times* covered it (not too technical). In February 2004, the FDA asked the companies involved to hold marketing, pending review. Shortly thereafter, FDA clarified that OvaCheck and other in vitro diagnostic multivariate index assays were subject to regulation as medical devices. Better experimental design and prospective validation would be required before marketing. In 2005, the NCI's Scientific Advisory Board rejected a proposed $89 million initiative largely searching for proteomic patterns, refocusing shortly thereafter on a new initiative, the Clinical Proteomic Tumor Analysis Consortium, aimed at better clarifying what the assays could and could not be expected to do. David Ransohoff played a key role here in explaining the issues.

Confounding-driven results have been found with every type of high-throughput assay *(2)*. A 2012 Institute of Medicine report *(3)* highlighted the need for better experimental design and prospective validation. The National Institutes of Health's 2016 Rigor and Reproducibility Initiative *(4)* notes that confounding problems are large contributors today to the irreproducibility of much scientific research.

The importance of basic issues will not be news to clinical chemists. Hopefully, broader absorption will come.

[1] MD Anderson Cancer Center, Houston, TX.
* Address correspondence to the author at: 1617 Calumet St., Houston, TX 77004. E-mail kabagg@gmail.com.

[2] This article has been cited more than 500 times since publication.

Received October 13, 2017; accepted October 18, 2017.
Previously published online at DOI: 10.1373/clinchem.2017.273334

[3] Nonstandard abbreviations: NCI, National Cancer Institute; FDA, Food and Drug Administration; SELDI, surface-enhanced laser desorption and ionization.

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: *No authors declared any potential conflicts of interest.*

References

1. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.
2. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 2010;11:733–9.
3. Omenn G, DeAngelis C, DeMets D, Fleming T, Geller G, Gray J, et al. Evolution of translational omics: lessons learned and the path forward. Washington (DC): National Academies Press; 2012.
4. NIH. Rigor and reproducibility. 2016. https://www.nih.gov/research-training/rigor-reproducibility (Accessed August 2018).