Clinical Chemistry 61:2
347–348 (2015)

Opinion

Learning from Our Mistakes: The Future of Validating Complex Diagnostics

Stephen R. Master[1*] and Viktor Mayer-Schönberger[2*]

In 2009, Google unveiled "Flu Trends," a program designed to estimate rates of influenza infection based solely on the use of search terms submitted by users across the US. In their initial published report of the work in *Nature (1)*, the authors demonstrated a striking match between their Google search–based estimates and the official flu statistics for 2007–08. In addition to the novelty of using search data for public health purposes, one of the most interesting aspects of this work was that the predictive algorithm was not developed using preselected, candidate search terms. Rather, using the "big data" available to Google from hundreds of millions of users, developers identified the most predictive search terms out of the 50 million most frequently used terms regardless of whether they "made sense." This unguided approach not only tracked the spread of the flu with high accuracy, but also provided its results 1–2 weeks earlier than similar estimates from the CDC.

However, Google Flu Trends was significantly less successful in predicting CDC winter flu data in 2012 *(2)*. One important reason for the inaccuracy was a change in the behavior of users. Whereas search term statistics up until 2007 were able to adequately predict influenza in 2007–08, changes in term usage led to a subsequent degradation in performance. This raises a critical question: would Google's diagnostic algorithm be more effective if it were allowed to retrain itself and learn new terms over time? Indeed, retrospectively incorporating changes from 2010 and 2011 into the algorithm led to improved predictions for the 2012 flu season *(3)*.

An analogous question has been raised in recent years regarding the use of large numbers of laboratory measurements, such as those obtained using "omics" technologies, in clinical diagnostics. As with Flu Trends, the goal of such assays is to identify analytes that can be combined using an algorithm to reach a diagnosis with high sensitivity and specificity. Unlike Flu Trends, which has been welcomed by public health authorities such as the CDC, acceptance in the laboratory context has been muted. Based on several high-profile controversies over the past decade, regulatory agencies and scientific bodies in particular have issued warnings against the naive use of such approaches.

There is no question that inadequate study design, poorly characterized preanalytical variation, and careless data management have led to inaccurate—even dangerous—errors in assay development. In one particularly well-publicized case, the results of an omics assay using flawed data and bioinformatics were used to guide patients into inappropriate arms of a clinical oncology trial *(4)*. To prevent such disasters, a number of responses have been proposed. For example, some suggest that test development should use analytes with a plausible link to underlying biology *(5)*. Others have advocated increased federal regulation, as exemplified by the recently proposed expansion in active oversight of laboratory-developed tests by the US Food and Drug Administration *(6)*. A third approach exerts strong control over the algorithm itself. This last strategy is most clearly seen in the 2012 Institute of Medicine report on omics assays, which proposed a "bright line" between research/development and clinical validation. In this model, such validation can only proceed once a diagnostic algorithm is fully specified and frozen in its final form *(7)*.

Although these approaches are reasonable and important in our current context, we believe that in the long run they may impose constraints that limit the accuracy and effectiveness of complex diagnostics. The paradigms that drive these constraints are products of a particular world view, outgrowths of a data-deprived environment in which gathering and analyzing data was difficult and time-consuming, and thus data use was limited. Data analyses were not routinely and regularly redone when new data became available, but rather often only when questions about the validity of the original analyses were raised. Because of the dearth of data, these analyses typically used static, relatively simple, parametric statistical models. Even when more complex analyses were undertaken, the use of a static model remained. The improvements in Flu Trends, however, suggest a different approach. We propose that using more modern, Bayesian statistical approaches to continually improve a model based on new data would have a similar effect in

[1] Assistant Professor of Pathology and Laboratory Medicine, Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA; [2] Professor of Internet Governance and Regulation, Oxford Internet Institute, Oxford University, Oxford, UK.
* Address correspondence to this author at: S.R. Master: 613A Stellar-Chance Labs, 422 Curie Blvd., Philadelphia, PA 19041. E-mail srmaster@mail.med.upenn.edu. V.M. Schönberger: Oxford Internet Institute, 1 St Giles, Oxford OX1 3JS, UK. E-mail viktor.ms@oii.ox.ac.uk.
Received September 27, 2014; accepted October 1, 2014.
Previously published online at DOI: 10.1373/clinchem.2014.231407

clinical diagnostics. The rapid evolution of medical practice, whether due to changes in standards of care, new treatment paradigms, or altered patient populations, is not well served by static diagnostic algorithms.

Further, because the lack of data has been combined with the natural human tendency to look for causes, we have privileged causal inquests over correlational ones, even when correlational insights might have provided us with a pragmatic way forward. Of course, biological investigation of causative pathways may provide important statistical advantages when analyzing some data sets. However, history has demonstrated the limitations of relying too heavily on known causes. For example, when the mid-19th century Hungarian physician Ignaz Semmelweis showed that proper hand disinfection correlated with a dramatic decrease in the occurrence of puerperal fever, his suggestion was dismissed because he had no proof of his causal explanation. As a result, tens of thousands of women died an unnecessary death. In point of fact, Semmelweis's proposed causal explanation was wrong, but so was the causal explanation of his peers who refused to wash their hands. In contrast, a more iterative approach that would have permitted small, yet important, steps of trial and error implementing hand disinfection might not have revealed the underlying cause—germ theory was not yet discovered—but could have provided sufficient trust in Semmelweis's correlational insight to implement hygienic practice faster and more comprehensively.

Perhaps the time has come to think in a similar way about pragmatic but bolder steps toward an iterative approach in the laboratory sciences that acknowledges the distinct qualities of large, multidimensional data sets rather than defending the methods and structures that were shaped by a small-data mindset. Restricting classifiers to the use of well-understood analytes, or requiring that algorithms be locked down with no possibility of future updates will, in the long run, slow our ability to learn from our mistakes.

This proposal creates a fairly obvious problem, however, for the validation and regulation of laboratory tests. If laboratories incorporate a dynamic analysis of complex data to improve the performance of diagnostic algorithms, how can consumers of the data have confidence that the laboratory has done its job correctly? Put another way, how can regulatory agencies be assured that tests are still robust and diagnostically reliable? In this context, one might think about establishing oversight on a metalevel, requiring the data collection and statistical analysis processes for updating the classifier to be rigorously specified and reviewed, rather than focusing on the static algorithm based on a necessarily limited pool of data. Given the emerging importance of reproducible data analysis pipelines, this focus for review processes may actually provide a significant step forward compared with current practice. One could even think of a special group of trained big-data experts—"algorithmists," if you want—that are available to regulatory agencies and to the clinical laboratory for review and auditing of their processes and practices, and who may be able to offer suggestions for improvements *(8)*.

We are not suggesting that established methods and protocols in the clinical laboratory should be abandoned tomorrow. Rather we want the community to be cognizant that many of our practices are at least partially shaped by how we have collected and analyzed data in the past. As data-deprivation is replaced by data surfeit, we will need to have a robust discussion on how methods and processes of working with data in clinical laboratories can be rethought as well.

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. Ginsburg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature 2009;457:1012–4.
2. Butler D. When Google got flu wrong. Nature 2013;494:155–6.
3. Copeland P, Romano R, Zhang T, Hecht G, Zigmond D, Stefansen C. Google disease trends: an update. Presented at: ISNTD Bites 2013. Proceedings of the 2nd International Society for Neglected Tropical Diseases ISNTD Bites conference; 2013 Oct 15; London, UK. https://drive.google.com/file/d/0B1UI69AUsTn1WWdJUnJFYnNDbkk/preview?pli=1 (Accessed January 2015).
4. Reich ES. Cancer trial errors revealed. Nature 2011;469:139–40.
5. McDermott JE, Wang J, Mitchell H, Webb-Robertson BJ, Haven R, Ramey J, Rodland KD. Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data. Expert Opin Med Diagn 2013;7:37–51.
6. FDA. Framework for regulatory oversight of laboratory developed tests (LDTs). http://www.fda.gov/downloads/MedicalDevices/ProductsandMedicalProcedures/InVitroDiagnostics/ucm407409.pdf (Accessed October 2014).
7. Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials, Board on Health Care Services, Board on Health Sciences Policy, Institute of Medicine. Micheel CM, Nass SJ, Omenn GS, eds. Evolution of translational omics: lessons learned and the path forward. Washington (DC): National Academies Press; 2012.
8. Mayer-Schönberger V, Cukier K. Big data. Boston: Mariner Books/Houghton Mifflin Harcourt; 2013:178–82.