



Improving Validation Practices in "Omics" Research

John P. A. Ioannidis, *et al.*
Science **334**, 1230 (2011);
DOI: 10.1126/science.1211811

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by [clicking here](#).

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines [here](#).

The following resources related to this article are available online at www.sciencemag.org (this information is current as of December 3, 2011):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:

<http://www.sciencemag.org/content/334/6060/1230.full.html>

A list of selected additional articles on the Science Web sites **related to this article** can be found at:

<http://www.sciencemag.org/content/334/6060/1230.full.html#related>

This article **cites 20 articles**, 7 of which can be accessed free:

<http://www.sciencemag.org/content/334/6060/1230.full.html#ref-list-1>

This article appears in the following **subject collections**:

Medicine, Diseases

<http://www.sciencemag.org/cgi/collection/medicine>

be several reasons for this. Many observations in nature require unusual expertise and experience, whereas laboratory experiments should be designed so they can be repeated with the same results by a naïve observer. Such an observer would lack the experience and expertise of a hardened field researcher like Jane Goodall and, if inserted into the wilds of Gombe with unhabituated chimpanzees, would have no chance of observing chimps using tools. Subsequent studies confirmed Goodall's original observations on tool use in chimps when similar abilities were observed in other chimp populations, other primates, and even New Caledonian crows (15). These were important, but in the end they were not needed to ensure the veracity of her original observations.

When field observations lead to field and laboratory experiments, however, rigor, controls, and replication similar to those used in the more traditional laboratory sciences are expected (14). Field biology recently has benefitted from an influx of

technology in which audio and video recordings, remote sensing, and satellite tracking are important aids in data collection. The videos that eventually appeared of chimps using tools and bats eating frogs, for example, provided added value to the original observations. These tools enhance the reliability of field observations, allow observations at a scale not previously possible, and are now a welcome addition to the field biologist's tool kit.

Field observations are good at telling us what happens in nature. Experimentation is better at demonstrating cause and effect. Experiments in the field encompass the variables under which animals function, whereas those in the laboratory allow for some control over these variables. Each has its virtues. All findings in the field and the laboratory make predictions which, if supported, add further support to what we think we know or, if not supported, lead us to doubt our interpretations. All of this is science and, if done well, is good science.

References

1. C. Darwin, *The Descent of Man and Selection in Relation to Sex* (Murray, London, 1871).
2. M. J. Ryan, *The Túngara Frog, A Study in Sexual Selection and Communication* (Univ. Chicago Press, Chicago, 1985).
3. M. J. Ryan, *Science* **209**, 523 (1980).
4. A. S. Rand, M. J. Ryan, *Z. Tierpsychol.* **57**, 209 (1981).
5. M. D. Tuttle, M. J. Ryan, *Science* **214**, 677 (1981).
6. M. J. Ryan, M. D. Tuttle, L. K. Taft, *Behav. Ecol. Sociobiol.* **8**, 273 (1981).
7. M. J. Ryan, M. D. Tuttle, A. S. Rand, *Am. Nat.* **119**, 136 (1982).
8. S. H. Hurlbert, *Ecol. Monogr.* **54**, 187 (1984).
9. R. A. Page, M. J. Ryan, *Curr. Biol.* **16**, 1201 (2006).
10. K. L. Akre, H. E. Farris, A. M. Lea, R. A. Page, M. J. Ryan, *Science* **333**, 751 (2011).
11. V. Bruns, H. Burda, M. J. Ryan, *J. Morphol.* **199**, 103 (1989).
12. J. Goodall, *Nature* **201**, 124 (1964).
13. N. B. Davies, *Cuckoos, Cowbirds and Other Cheats* (Princeton University, Princeton, NJ, 2000).
14. A. R. Palmer, *Annu. Rev. Ecol. Syst.* **31**, 441 (2000).
15. G. R. Hunt, *Nature* **379**, 249 (1996).

10.1126/science.1214532

PERSPECTIVE

Improving Validation Practices in “Omics” Research

John P. A. Ioannidis¹ and Muin J. Khoury^{2*}

“Omics” research poses acute challenges regarding how to enhance validation practices and eventually the utility of this rich information. Several strategies may be useful, including routine replication, public data and protocol availability, funding incentives, reproducibility rewards or penalties, and targeted repeatability checks.

The exponential growth of the “omics” fields (genomics, transcriptomics, proteomics, metabolomics, and others) fuels expectations for a new era of personalized medicine. However, clinically meaningful discoveries are hidden within millions of analyses (1). Given this immense biological complexity, separating true signals from red herrings is challenging, and validation of proposed discoveries is essential.

Some fields already employ stringent replication criteria. For example, in genomics, genome-wide association studies demand high statistical significance (P values $< 5 \times 10^{-8}$) and perform large-scale replication efforts within international consortia (2). Conversely, other fields continue

to perform “mile-long, inch-thick” research (3), in which many factors are tested once (“discovered”) but are rarely further validated. Studies in gene expression profiling and transcriptomics sometimes try to validate the results using different assays within single populations as well as statistical techniques such as cross-validation, which do not require the evaluation of additional, independent samples. However, such methods do not guarantee good performance across different populations. Moreover, very often cross-validation overestimates classifier performance, probably because biases are introduced in the process (4, 5). Independent external validation usually yields more conservative results, but may also be inflated because of optimism, selective reporting, and other biases (5, 6). Independent external validation by completely different teams remains rare.

Even strong replication of omics results does not automatically imply the potential for successful adoption in clinical or public health practice. Demonstrating clinical validity requires evaluation of the predictive value in real-practice populations, whereas clinical utility requires evaluation of the balance of benefits and harms associated with the adoption of these technologies

for different intended uses (7). Ideally, randomized clinical trials are needed to assess whether omics information improves patient outcomes. Long-term, large-scale trials, such as those under way for Oncotype^{DX} (a diagnostic test that analyzes a panel of 21 genes within a breast tumor to assess the likelihood of disease recurrence and/or patient benefit from chemotherapy) and MammaPrint (a breast cancer signature of 70 genes) also require careful consideration of design issues (8, 9), because information on available classifiers constantly changes and new classifiers are proposed. There is at least one recent unfortunate example, where gene signatures were moved into clinical trial experimentation with insufficient previous validation. Three trials of gene signatures to predict outcomes of chemotherapy in treating non-small-cell lung cancer and breast cancer were suspended in 2011 after the realization that their supporting published evidence was nonreproducible (10).

Many scientists now demand reproducible omics research (11). This requires access to the full data, protocols, and analysis codes for published studies so that other scientists can repeat analyses and verify results. Fortunately, several public data repositories exist, such as the Gene Expression Omnibus, ArrayExpress, and the Stanford Microarray Database. There have also been many calls for diverse comprehensive study registries, such as for tumor biomarkers, a field riddled with uncertainty because of suboptimal study design and data quality, and a poor replication record (12, 13). Many leading journals are now working to adopt policies to make public deposition of data and protocols a prerequisite for publication (14). However, the practice of making this information accessible is applied inconsistently; furthermore, it is challenging to verify that complete data and protocols are indeed

¹Stanford Prevention Research Center, Department of Medicine and Department of Health Research and Policy, Stanford University School of Medicine and Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA 94305, USA. ²Office of Public Health Genomics, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA; and Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, MD 20852, USA.

*To whom correspondence should be addressed. E-mail: muk1@cdc.gov

deposited, files are usable, and results repeatable. An empirical assessment of 18 published papers of microarray studies showed that independent analysts could perfectly reproduce the results of only two of the studies, and it sometimes took over a month to reproduce a single figure (15). Moreover, in some fields that started with good prospects for public data deposition, such as genome-wide association studies, many investigators have become more resistant to sharing data because of perceived confidentiality issues (16).

Data, protocol, and code deposition for access by qualified investigators should be more widely adopted. Possible steps in this direction include making data-protocol-algorithm deposition a prerequisite for publication in all journals, not just select ones, and mandating this procedure for all future funded proposals and for all data submitted to regulatory agencies. Some resources are needed to support the development and expansion of such repositories, but hopefully once the infrastructure is set up and streamlined, maintenance should not be expensive. Moreover, federal funding agencies could offer financial bonuses to investigators whose publicly available data have demonstrably been used by multiple other investigators, with citations made to the publicly available data sets. Actual repeatability exercises require considerable effort, and it is

currently not feasible to have a “police force” of analysts repeating all published and deposited data and analyses. However, targeted repeatability checks are possible by specialists in bioinformatics/biostatistics (or by other field experts), with a priority for postulated discoveries that have potentially high clinical impact. For example, a repeatability check of all relevant data can become a prerequisite to clinical trial experimentation. Given that few discoveries (currently much less than 1%) progress that far in the translation process, such a repeatability check is logistically manageable without an inordinate cost. Such checks can be competitively outsourced to bioinformatics experts, for example, who have not been involved in any of the original work. Funding agencies may also outsource spot checks of repeatability in a small number of randomly selected funded omics projects and may offer bonuses to investigators with repeatable results and penalize the future funding of investigators with nonrepeatable results. In this fashion, the penalties accrued from instances of nonreproducible results may help to cover the costs of the validation exercises.

Another possibility is to make collections of massive amounts of omics data available to all investigators who are interested in analyzing them. Rewards (funding) could go to those who

succeed in producing the best models (for example, models that achieve the best predictive discrimination for cancer), which can also be validated by other, independent data sets. This is a mechanism similar to what Netflix has used. The movie rental company launched an open public competition to generate models that would best predict movie ratings based on available collected data on movie preferences (17). Such an approach makes data available on an egalitarian basis to all potentially interested investigators, maximizes data transparency, and rewards investigators that use the best methods and get the most-reproducible results.

Some fields have recognized the need to reevaluate even earlier stages of measurement validity. For example, analytic validity remains a challenge in proteomics. One study (18) tested the ability of 27 laboratories to evaluate standardized samples containing 20 highly purified recombinant human proteins with mass spectrometry, a simple challenge compared to the thousands of proteins involved in clinical samples. Only seven laboratories reported all 20 proteins correctly, and only one lab captured all tryptic peptides of 1250 daltons. This exemplifies how factors such as missed identifications, contaminants, and poor database matching can create havoc. In such fields, efforts to improve interlaboratory reproducibility of data should precede efforts to promote independent validation in large samples from different data sets (19, 20). Several efforts such as the recently launched Human Proteome Project (21) and other Human Proteome Organisation (www.hupo.org) initiatives are currently under way to create networks of laboratories; interactions that will hopefully enhance reproducibility.

The validation of large-scale omics discoveries involves multiple steps (Fig. 1). Some steps are easier to accomplish in specific fields than others—for example, genotyping of common gene variants has reached almost-perfect analytical validity, with measurement error <0.01%, and the genome community continues to pursue the development of reference materials for use in validating whole-genome sequencing and variant calling; on the other hand, the field of proteomics is struggling with measurement error basics. Given that new measurement platforms continuously emerge, and information is increasingly combined from many of them, one has to decide which approaches are more reproducible and potentially informative. Empirical replication with large-scale studies is therefore increasingly necessary. One may argue that this is not easy because of technical and cost considerations. However, similar arguments were made for fields such as human genome epidemiology, which then saw the cost of DNA sequencing decrease over a billionfold over the past 20 years and the amount of information increase proportionally. Costs could decrease for other technologies as

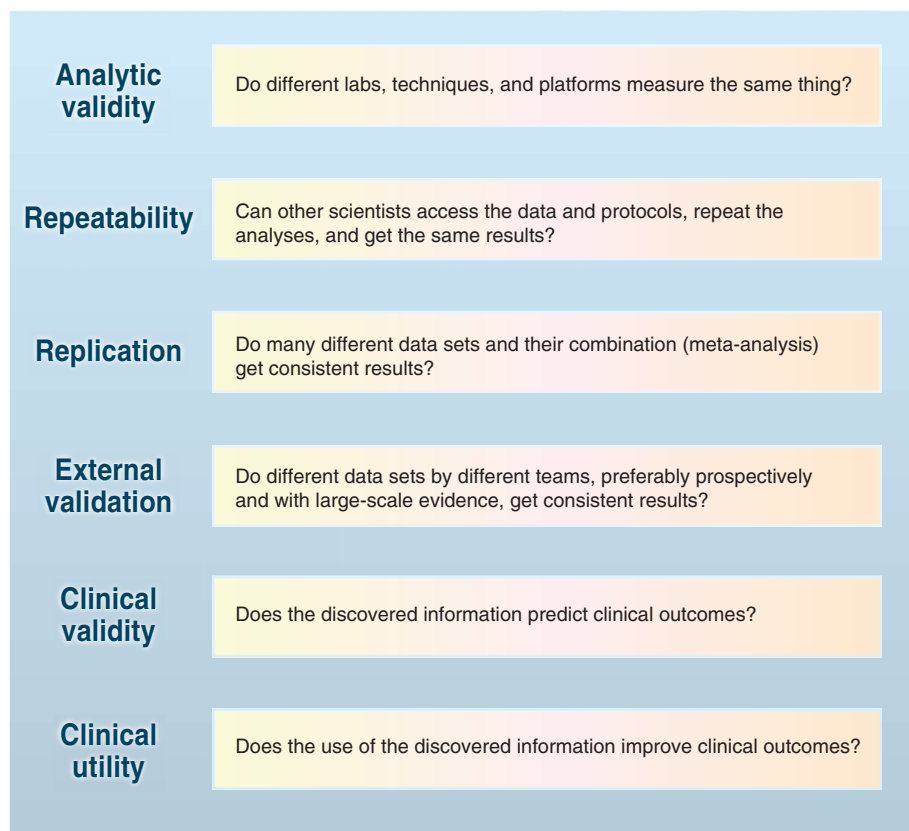


Fig. 1. The validation of omics research for use in medicine and public health requires fulfilling multiple steps. [Adapted from (7)]

technologies attract the attention of many investigators, especially in large consortia, thereby driving data reproducibility in a field. Funding incentives, reproducibility rewards and/or non-reproducibility penalties, and targeted requirements for repeatability checks may enhance the public availability of useful data and valid analyses.

References and Notes

1. J. P. A. Ioannidis, *PLoS Med.* **2**, e124 (2005).
2. S. J. Chanock *et al.*, *Nature* **447**, 655 (2007).
3. R. D. Riley *et al.*, *Health Technol. Assess.* **7**, 1 (2003).
4. A. Dupuy, R. M. Simon, *J. Natl. Cancer Inst.* **17**, 99 (2007).

5. P. J. Castaldi, I. J. Dahabreh, J. P. Ioannidis, *Brief. Bioinform.* **12**, 189 (2011).
6. M. Jelizarow, V. Guillelot, A. Tenenhaus, K. Strimmer, A. L. Boulesteix, *Bioinformatics* **26**, 1990 (2010).
7. S. M. Teutsch *et al.*, EGAPP Working Group, *Genet. Med.* **11**, 3 (2009).
8. J. A. Sparano, S. Paik, *J. Clin. Oncol.* **26**, 721 (2008).
9. F. Cardoso *et al.*, *J. Clin. Oncol.* **26**, 729 (2008).
10. K. Baggerly, K. R. Coombes, *Ann. Appl. Stat.* **3**, 1309 (2009).
11. K. Baggerly, *Nature* **467**, 401 (2010).
12. F. Andre *et al.*, *Nat. Rev. Clin. Oncol.* **8**, 171 (2011).
13. J. P. A. Ioannidis, O. A. Panagiotou, *JAMA* **305**, 2200 (2011).
14. A. A. Al-Sheikh-Ali, W. Qureshi, M. H. Al-Mallah, J. P. Ioannidis, *PLoS ONE* **6**, e24357 (2011).

15. J. P. A. Ioannidis *et al.*, *Nat. Genet.* **41**, 149 (2009).
16. A. D. Johnson, R. Leslie, C. J. O'Donnell, *PLoS Genet.* **7**, e1002269 (2011).
17. D. F. Ransohoff, *Clin. Chem.* **56**, 172 (2010).
18. A. W. Bell *et al.*, HUPO Test Sample Working Group, *Nat. Methods* **6**, 423 (2009).
19. H. Mischak *et al.*, *Sci. Transl. Med.* **2**, 46ps42 (2010).
20. M. Mann, *Nat. Methods* **6**, 717 (2009).
21. P. Legrain *et al.*, *Mol. Cell. Proteom.* **10**, M111.009993 (2011).

Acknowledgments: The opinions expressed in this paper are those of the authors and do not reflect an official position of the Department of Health and Human Services.

10.1126/science.1211811

PERSPECTIVE

The Reproducibility of Observational Estimates of Surface and Atmospheric Temperature Change

B. D. Santer,^{1*} T. M. L. Wigley,² K. E. Taylor¹

Although concerns have been expressed about the reliability of surface temperature data sets, findings of pronounced surface warming over the past 60 years have been independently reproduced by multiple groups. In contrast, an initial finding that the lower troposphere cooled since 1979 could not be reproduced. Attempts to confirm this apparent cooling trend led to the discovery of errors in the initial analyses of satellite-based tropospheric temperature measurements.

“We have produced, using objective techniques, a long-term series of average Northern Hemisphere temperatures” (1, 2). This innocuous sentence was published in 1982 by Phil Jones and colleagues at the University of East Anglia’s Climatic Research Unit (CRU). The sentence was a prologue to the modern era of scientific attempts to estimate large-scale changes in the Earth’s average surface temperature. Building on earlier work by American, British, Russian, and Japanese teams [reviewed in (1)], CRU researchers took on the difficult challenge of transforming surface temperature measurements from many hundreds of meteorological stations into credible scientific estimates of temperature changes over land areas of the planet. Temperature changes over ocean areas (3, 4) and over the whole globe (land plus ocean) were soon to follow (5). Few could have imagined the far-reaching scientific, societal, and political repercussions of this seemingly routine research.

The CRU team soon joined forces with scientists at the UK Meteorological Office Hadley

Centre (MOHC), who were refining estimates of observed changes in sea-surface temperature (SST). This scientific partnership led to the development of the Hadley Centre/CRU observational record of combined changes in SST and land-surface temperature (HadCRUT). The HadCRUT data set has provided hard scientific evidence for the warming of Earth’s surface over the past 150 years (6).

The Hadley Centre and CRU efforts to construct successive versions of the HadCRUT data set were open and transparent, and are documented in many peer-reviewed papers. From the very beginning of this research, CRU and MOHC scientists recognized the difficulties involved in estimating the true (but unknown) temperature change in the physical climate system. In order to do this, it was necessary to account for the effects of nonclimatic factors, such as temporal changes in the type of thermometer used to make temperature measurements, the thermometer location and its immediate physical surroundings, and the time of observation. Jones and colleagues found that even if they made different (but reasonable) choices in data set construction, their bottom-line conclusion—that the surface of our planet experienced rapid warming over the second half of the 20th century—was rock solid.

“An extraordinary claim requires extraordinary proof” (7). The claim that the planet had warmed markedly during the 20th century had extraordinary societal implications, and was therefore subjected to extraordinary scrutiny. Groups at the NASA/Goddard Institute for Space Studies in New York (GISS) and at the National Oceanic and Atmospheric Administration’s National Climatic Data Center (NCDC) in North Carolina independently attempted to reproduce the HadCRUT results. Although all three teams used raw temperature measurements from similar (but nonidentical) sets of observing stations, they made different choices in the treatment of these raw measurements and the calculation of area averages (8). In spite of these differences, the GISS and NCDC analyses confirmed the “warming Earth” findings of the CRU and MOHC scientists (9, 10).

Other lines of evidence substantiated the CRU/MOHC, GISS, and NCDC estimates of planetary temperature increase. The surface warming was consistent with the independently monitored retreat of snow and Northern Hemisphere sea-ice cover, the widespread melting and retreat of glaciers, the rise in global-mean sea level, and the increase in the amount of water vapor in the atmosphere. These and many other independent observations provided the scientific underpinning for the finding that “warming of the climate system is unequivocal” (11).

Yet doubts about the reality of 20th-century surface warming remained, especially in the aftermath of the 2009 event colloquially referred to as “Climategate.”

Reproducibility—the independent verification of prior findings—is at the core of “the spirit of science” (12). An additional verification effort has helped to underscore the robustness of the finding of “unequivocal” surface warming. Richard Muller, representing a fourth research group, testified to a Congressional committee in March of 2011 that his team had independently reproduced the land-surface warming found previously by the other three groups (13). The data set developed during this verification study has now been publicly released, and scientific papers describing the findings of this fourth group are currently undergoing peer review.

¹Program for Climate Model Diagnosis and Intercomparison, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA. ²National Center for Atmospheric Research, Boulder, CO 80307–3000, USA.

*To whom correspondence should be addressed. E-mail: santer1@llnl.gov