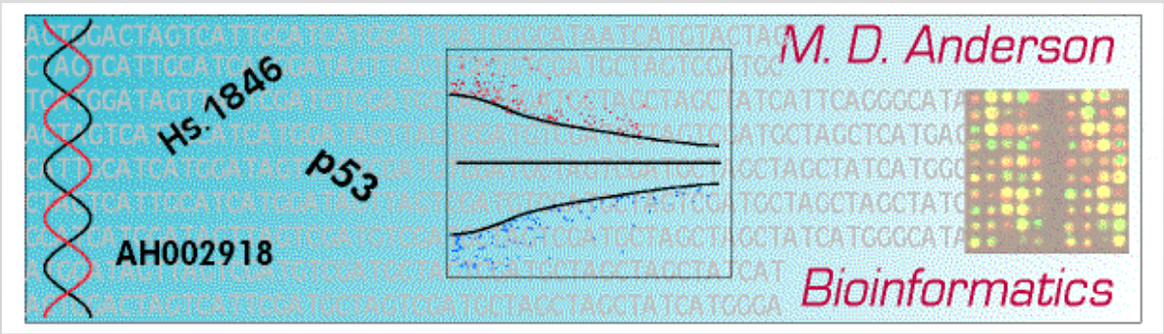
The need for publicly verifiable and reproducible data and analyses

Kevin R. Coombes
Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kcoombes@mdanderson.org

Research Integrity, Mohonk, 8 August 2012



Reproducibility: A Case Study

ture.com/naturemedicine

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

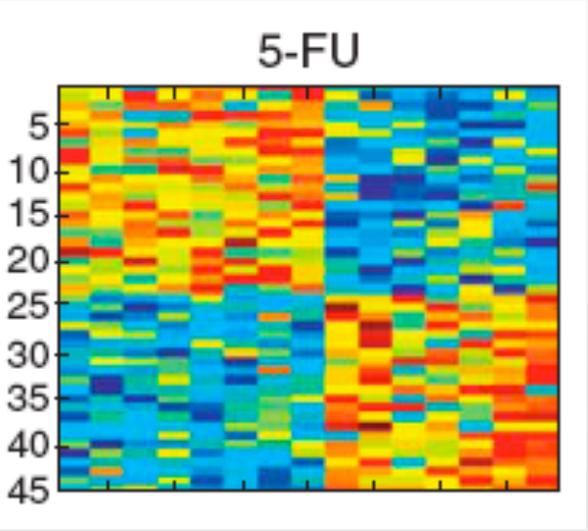
Potti et al (2006), Nature Medicine, 12:1294-1300.

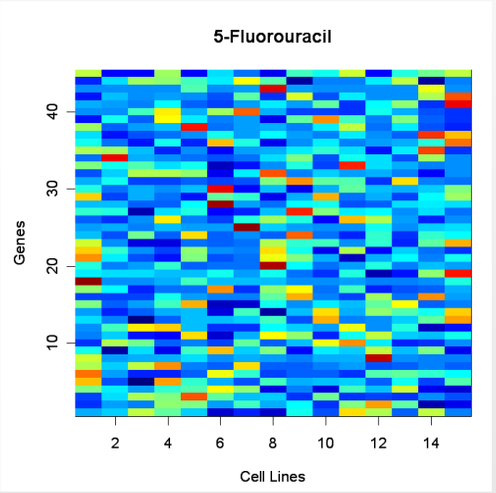
Their main conclusion: we can use microarray data from cell lines (the NCI60) to define drug response "signatures", which can then predict whether patients will respond.

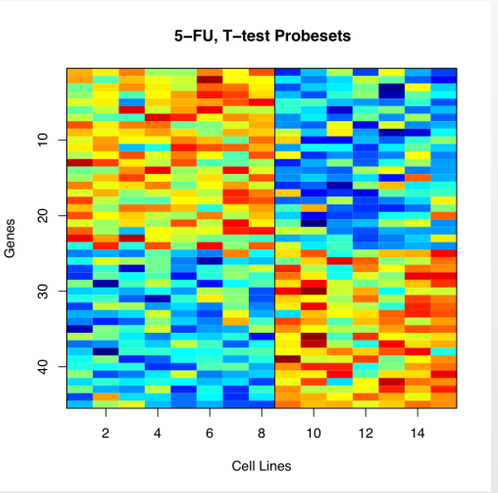
They provided examples using 7 commonly used drugs.

This got people at M.D. Anderson very excited. And, since all the data was publicly available from the people who originally generated it, we could try it ourselves....

Gene Lists Were Off-by-One







Nat Med Paper Their Genes Our t-tests

```
     Theirs          Ours
...
[3,] "1881_at"       "1882_g_at"
[4,] "31321_at"      "31322_at"
[5,] "31725_s_at"    "31726_at"
[6,] "32307_r_at"    "32308_r_at"
```

And There Were Other Genes...

The 50-gene list for docetaxel contains 19 outliers: genes that do not have small t-test p-values after correcting for the off-by-one error.

The initial paper on the test data (Chang et al) gave a list of 92 genes that separated responders from nonresponders.

Entries 7-20 in Chang et al's list comprise 14 out of the 19 outliers.

But there were five other genes: ERCC1, ERCC4, ERBB2, BCL2L11, TUBA3. These are the genes named explicitly in the discussion of the paper to explain the biology.

Mysterious Genes Appeared Again

Pharmacogenomic Strategies Provide a Rational Approach to the Treatment of Cisplatin-Resistant Patients With Advanced Cancer

David S. Hsu, Bala S. Balakumaran, Chaitanya R. Acharya, Vanja Vlahovic, Kelli S. Walters, Katherine Garman, Carey Anders, Richard F. Riedel, Johnathan Lancaster, David Harpole, Holly K. Dressman, Joseph R. Nevins, Phillip G. Febbo, and Anil Potti

J Clin Oncol, Oct 1, 2007, 25:4350-7. Same approach, using Cisplatin and Pemetrexed. For cisplatin, U133A arrays were used for training. In the discussion, ERCC1, ERCC4, and DNA repair genes are identified as "important".

With some work, we matched the heatmaps. We were unable to match four genes: 203719_at (ERCC1), 210158_at (ERCC4), 228131_at (ERCC1), and 231971_at (FANCM (DNA Repair)). *The last two probesets aren't on the U133A arrays that were used. They're on the U133B.*

Sensitive/Resistant Labels Were Reversed

Figure 1E, Potti et al. (2006) Nature Medicine, 12:1294–1300.

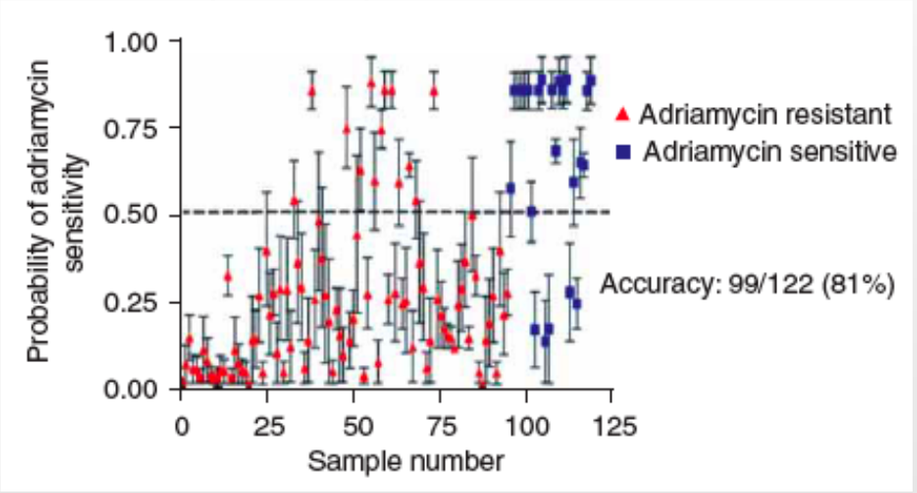
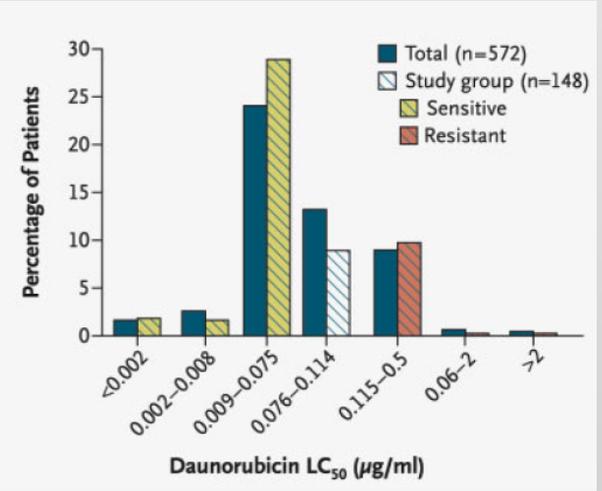
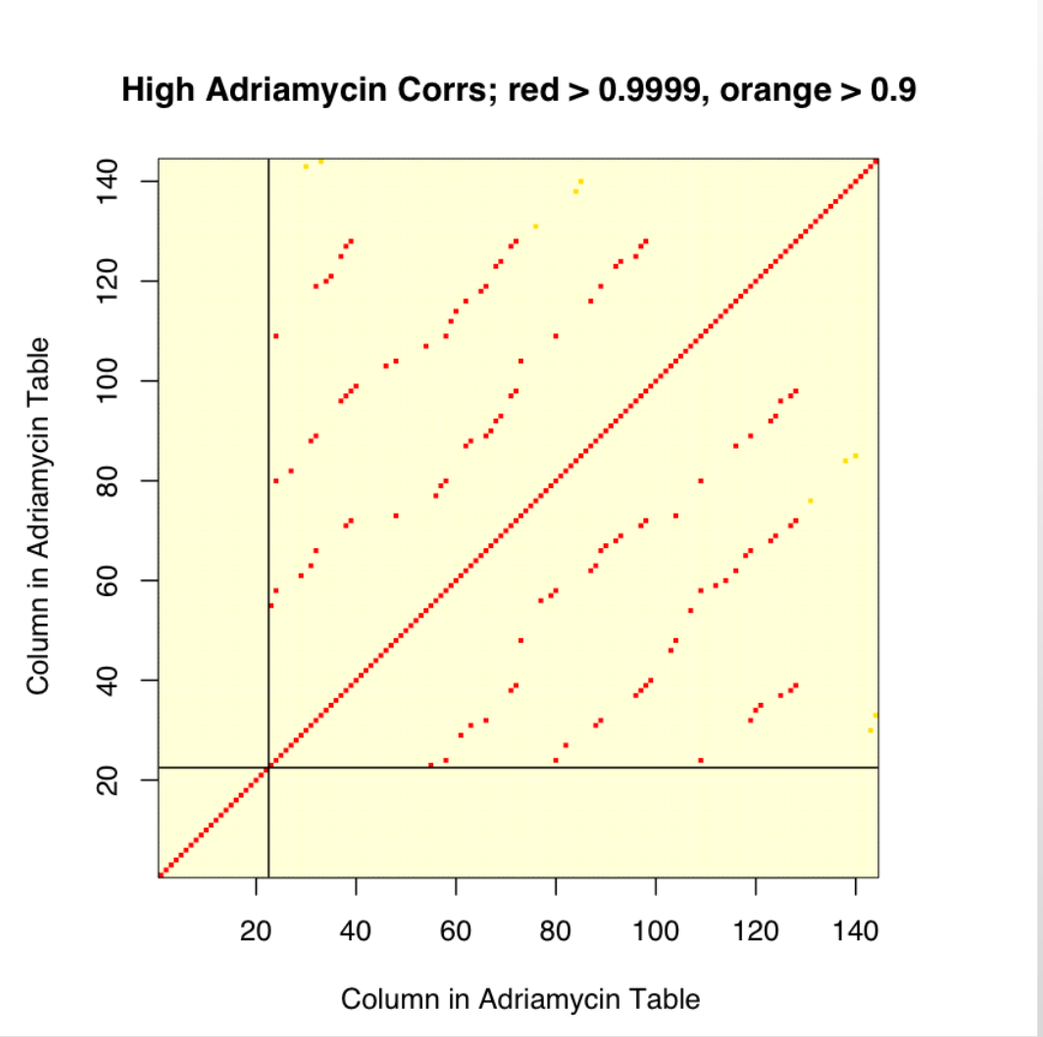


Figure 1, Holleman (2004), NEJM, 351:533–542.

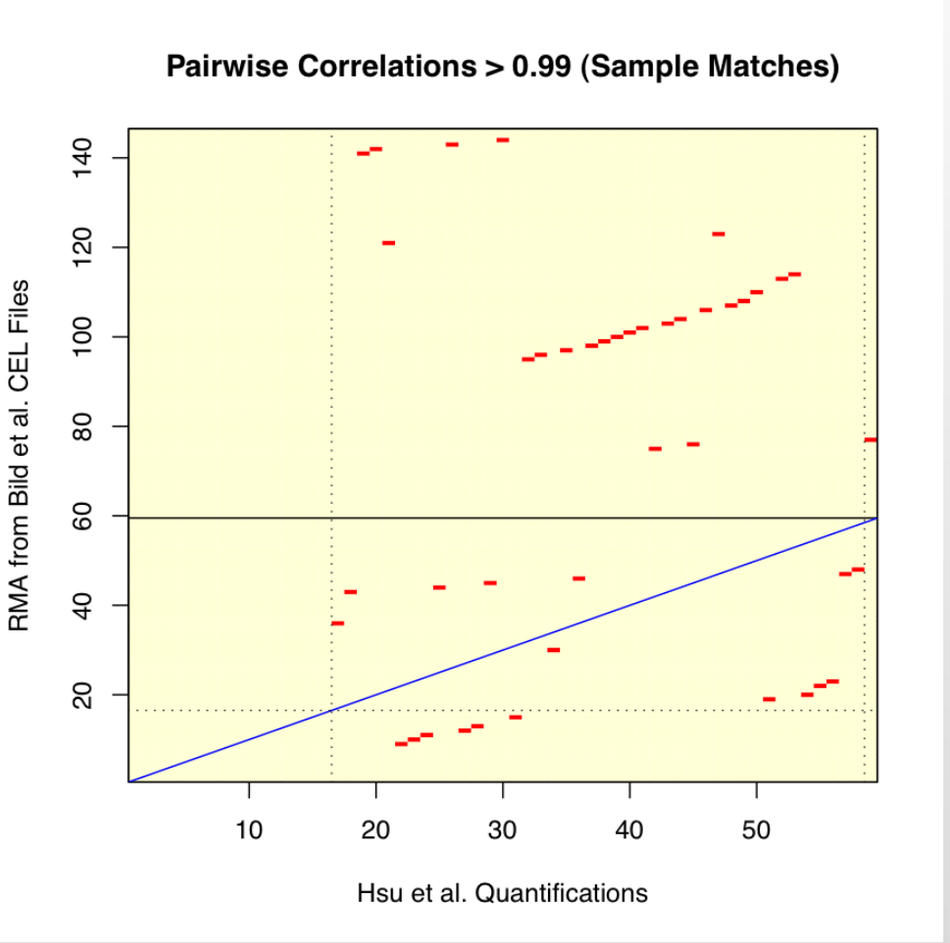


The Same Samples Were Included Multiple Times



Red dots mark sample matches. Expect a diagonal line.

Samples Did Not Match the Labels



43 samples are mislabeled; 16 don't match at all. Expect all red dots to fall on the blue line.

A Partial Timeline

6 November 2007: Our letter to Nature Medicine.
14 September 2009: Our paper in *Annals of Applied Statistics*.
Sep-Oct 2009: Duke starts internal investigation, suspends trials.
29 January 2009: Duke restarts trials.
16 July 2010:


THE CANCER LETTER

PO Box 9905 Washington DC 20016 Telephone 202-362-1809

Prominent Duke Scientist Claimed Prizes He Didn't Win, Including Rhodes Scholarship

By Paul Goldberg

19/20 July 2010: Letter to Varmus; Duke resuspends trials.

October/November 2010

THE CANCER LETTER

Vol. 36 No. 39
Oct. 29, 2010

PO Box 9905 Washington DC 20016 Telephone 202-362-1809

Nevins Retracts Key Paper By Duke Group, Raising Question Of Harm To Patients

By Paul Goldberg

Duke genomic researcher Joseph Nevins has notified his co-authors that he is retracting a paper that provides the scientific justification for two controversial clinical trials conducted at the university.

In an Oct. 22 email, Nevins, a senior author on the paper published in the Oct. 1, 2007, issue of the Journal of Clinical Oncology, acknowledged that patients at Duke were being assigned to cancer therapy based on a biomarker test that he now realizes is inaccurate.

Personalized Medicine:
Experts Say Harm
To Patients Plausible
In Duke Studies
. . . Page 2

The Chronicle

HOME | NEWS | SPORTS | OPINION | RECESS | TO

Article | Comments (2)

NEWS » HEALTH & SCIENCE » RESEARCH AT DUKE

Suspended cancer trials terminated

By JULIA LOVE
November 9, 2010

The University voluntarily canceled three clinical trials that drew from the research of Dr. Anil Potti, a cancer researcher whose research is currently under investigation.

The trials had previously been suspended. Duke researchers stopped admitting new patients to the two studies on lung cancer and one study on breast cancer July 18. The principal investigators made the decision to permanently end the trials

Print Article
Email Article
Download PDF
ShareThis

Potti Retraction Scorecard (June 2012)

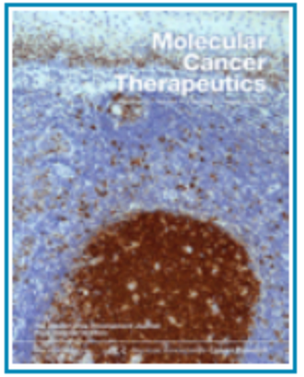
Retraction Watch

Tracking retractions as a

A "retraction in part" for Anil Potti and colleagues, in Molecular Cancer Therapeutics

with one comment

A partial retraction has joined the ten retractions and five corrections of Anil Potti's papers, this one of a 2008 paper in *Molecular Cancer Therapeutics*. The move comes 14 months after the retraction of the *Nature Medicine* paper upon which much of the *Molecular Cancer Therapeutics* paper was based.



Here's the notice:

> Retraction in Part: A Genomic Approach to Identify Molecular Pathways Associated with Chemotherapy Resistance
>
> We wish to retract Table 1 and Supplemental Table 1 from our article entitled "A genomic approach to identify molecular pathways associated with chemotherapy resistance," which was published in the October 2008 issue of *Molecular Cancer Therapeutics* (1).

Did the System Work?

It's not clear.

A scientifically irrelevant issue (the Rhodes scholarship) brought increased scrutiny. This scrutiny was leveraged to focus attention back to the science, and eventually led to retractions and to the termination of clinical trials.

An earlier investigation at Duke did not find or did not understand the seriousness of the scientific errors that we complained about, and so the clinical trials were resumed.

Why was it so hard for the investigators to see the problems clearly? We spent 1500+ hours figuring out what happened.

Scientific Integrity, Take 1

Federal research rules focus on misconduct, defined as:

- Falsification,
- Fabrication,
- Plagiarism.

The National Science Foundation and the Office of Research Integrity confirm about 20 cases of misconduct per year.

(Source: AAAS Forum on Science and Technology Policy, 2006)

This is **not** the primary subject of my talk.

Scientific Integrity, Take 2

In a survey of several thousand NIH-funded scientists (Martinson et al., *Nature*, 2005; **435**:737-738):

- 0.3% admitted to falsifying data (in the previous three years)
- 1.4% admitted to plagiarism
- 6.0% failed to present data that contradicted their previous research
- 10.8% withheld details of methodology or results
- 13.5% used inadequate or inappropriate research designs
- 15.5% changed the design, methodology, or results in response to pressure from a funding source
- 27.5% had inadequate record keeping related to their research

Prediction, classification, etc.

Diagnosis = prediction of presence or absence of disease

Prognosis = prediction of good or poor outcome

Selecting therapy = prediction of who will or will not respond

Prediction is hard, especially about the future.

– Yogi Berra ?
– Niels Bohr ?
– Mark Twain ?

See The Economist, July 2007

Prediction, classification, etc.

Diagnosis = prediction of presence or absence of disease

Prognosis = prediction of good or poor outcome

Selecting therapy = prediction of who will or will not respond

Prediction is hard, especially about the future.

– Yogi Berra ?

– Niels Bohr ?

– Mark Twain ?

See The Economist, July 2007

[I]f we found a character in a novel represented as habitually uttering true predictions of the future, we should cry out at once against the improbability.

– Charles Astor Bristed (aka Carl Benson), *Casual Cogitations*, **The Galaxy**, 1873; 16:196–201.

Our PubMed Search

(“Prognosis”[Mesh] OR “Diagnosis”[Mesh] OR “Statistics as Topic/diagnostic use”[Mesh] OR “Statistics as Topic/classification”[Mesh])
AND (“Proteomics”[Mesh] OR “Microarray Analysis”[Mesh])
NOT (“Review Literature as Topic”[Mesh] OR “Review ”[Publication Type])
AND cancer*
AND *various*[DP]*†

Our Review of Abstracts

- Randomly selected 40 articles published in 2007
- Only $14/40$ (35%) appeared (based on the abstract) to use high-throughput technologies for discovery or validation of predictive models.
- Only $11/14$ were available electronically for complete review.
- Only $5/11$ (based on full article) included predictive models.

Suggests that there are at least ten articles per month trying to build high-throughput models to make predictions about cancer.

Our Review of Articles

- Clinical data clearly available for $4/11$ (36%).
- Assay data publicly available for $4/11$ (36%); raw assay data only available for $2/11$ (18%).
- Quantification software named in $7/11$ (64%).
- Preprocessing steps fully described in $4/11$ (36%); at least partially described in $7/11$ (64%).
- Prediction software named in $4/5$ (80%)
- Prediction algorithm described in $4/5$ (80%)
- Prediction model fully reported in $2/5$ (20%)

Overall, only $3/11$ (27%) were **potentially** verifiable.

The Problem is Ubiquitous

Ochsner et al., Nat Meth (2008):

Deposition rates at GEO and ArrayExpress for 20 journals that "require" MIAME compliance are actually below 50%.

Ioannidis et al., Nat Gen (2009):

Tried to reproduce analyses from 18 quantitative microarray papers published over a two-year span (during MIAME compliance). Succeeded for only 2. Reproducibility was impossible even in principle for 10 of them.

Advertising or Science?

"An article about computational science in a scientific publication is **not** the scholarship itself, it is merely **advertising** of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

– Jonathan Buckheit and David Donoho
Wavelab and Reproducible Research
In: Wavelets and Statistics,
A. Antoniadis, ed., Springer, 1995

The statistical analysis of high throughput ("omics") data is a computational science.

Letter to Nature

Baggerly K. Disclose all data in publications. Nature. 2010 Sep 23; 467:401.

Signed on behalf of seven co-authors. Full author list:

Keith Baggerly, Kevin Coombes, Peter Diggle, Steve Goodman, Rafael Irizarry, John Quackenbush, Rob Tibshirani.

The authors were the writing committee for

Scientists for Reproducible Research

http://groups.google.com/group/reproducible-research

Recommendations in Letter to Nature

1. Make Primary Data publicly available.

2. Establish Provenance for re-analyses by supplying primary source references. Document any transformations.

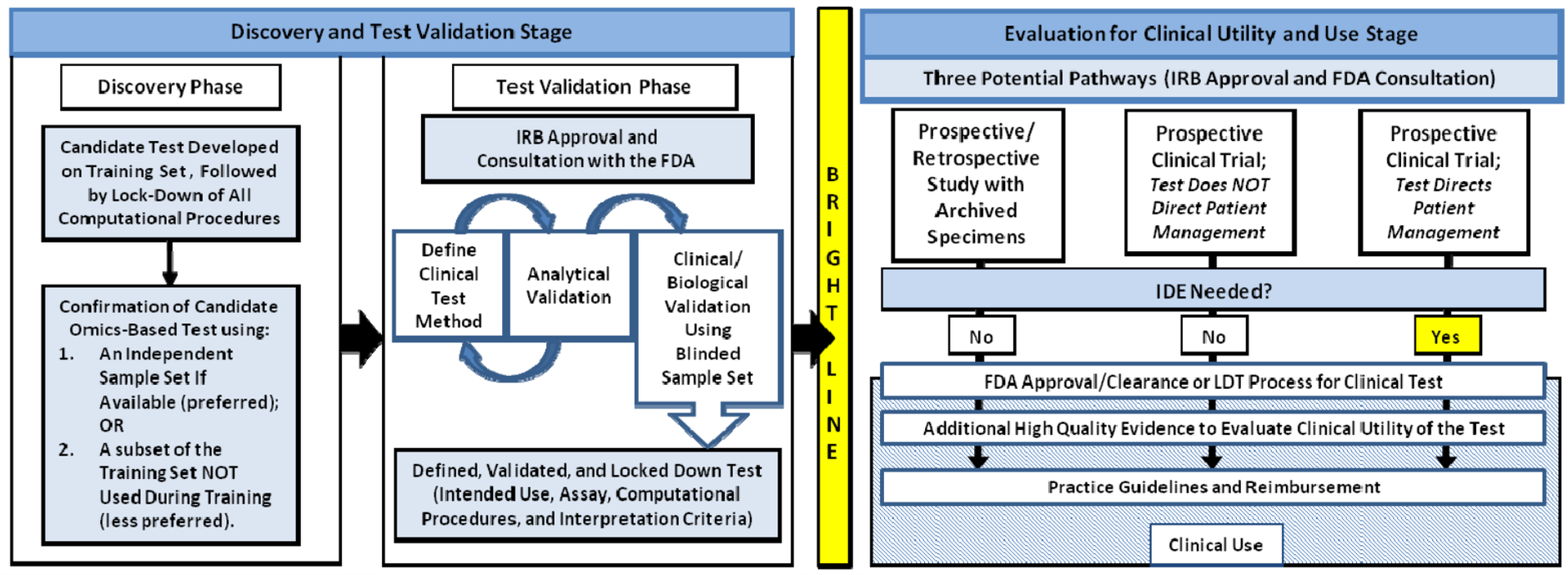
3. Make complete Software Code available.

4. Provide detailed, step-by-step Analytical Protocols with parameters, software source, version, and operating system for any steps that cannot be supplied as code.

5. If a pre-specified Research Protocol exists, supply it.

Justify any omissions, and describe alternate steps to assure reproducibility.

Institute of Medicine Report



- Independent, blind validation.
- Make data and code available
- Lock-down the predictive model.
- Define intended use.
- Move to CLIA-certified lab.
- Must get IDE from FDA if test affects patient management.

Reproducibility? or Correctness?

Answer comes from key concept of Open Source software:

Given enough eyeballs, all bugs are shallow.
– Eric Raymond's formulation of Linus's Law in "The Cathedral and the Bazaar"

Somebody finds the problem, and somebody else understands it. And I'll go on record as saying that finding is the bigger challenge.
– Linus Torvalds

For statistical analyses of high-throughput biological data, reproducibility is a necessary step before anyone can begin to tell if the analysis was correct.

Correctness?

ARTICLE

Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting

Alain Dupuy, Richard M. Simon

Background Both the validity and the reproducibility of microarray-based clinical research have been challenged. There is a need for critical review of the statistical analysis and reporting in published microarray studies that focus on cancer-related clinical outcomes.

J Natl Cancer Inst. 2007; **99**:147–57.

Main finding: In a careful review of articles published in 2004, Dupuy and Simon found that $21/42$ (50%) microarray studies contained at least one major statistical flaw.

Included studies for class comparison, class discovery, and class prediction where the methods were adequately described.

The Scope of the Problem

- About two-thirds of articles provide insufficient detail for independent verification. (They are "not even wrong", in the words of Wolfgang Pauli.)
- Of articles with enough detail, about half make a common statistical error that can be detected by reading the methods.
- At least some of the remaining articles either don't apply the methods they claim to apply, or apply them incorrectly.
- Some of the remaining articles may not generalize to new data because of inappropriate experimental designs or a narrow spectrum of cases or controls.
- Thus, it is possible that as few as 5% of the articles using high throughput technologies for prediction are correct.

The Institutional Challenge

Insisting that code and data be made publicly available would not have prevented the problems at Duke. We might have found the problems earlier; others might have been able to confirm them more easily.

We asserted: Their analysis was wrong. The data and code did not support the published claims. Thus, Duke was running clinical trials based on flawed science, and could potentially harm patients.

Scientific misconduct is irrelevant to these assertions. Simple errors could lead to the same situation.

However, insitutions do not know how to deal with assertions that the science is wrong. Existing mechanisms for investigating misconduct take too long to correct ongoing clinical trials.

Acknowledgements

Keith Baggerly

Shannon Neeley, Jing Wang

David Ransohoff, Gordon Mills

Jane Fridlyand, Lajos Pusztai, Zoltan Szallasi

M.D. Anderson Prostate SPORE, Lung SPORE, Breast SPORE, and Ovarian SPORE

http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified

http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified/StarterSet