Big Data: *How to avoid a* Big Mess

(Injecting some caution into the discussion)

Rob Tibshirani, Stanford

`Email:tibs@stanford.edu`

`http://www-stat.stanford.edu/~tibs`

Big data is a good thing

- These are exciting times for biostatisticians
- Statistics is a relatively young field ($<$ 100 years old), born in the small data age
- Today we need new statistical tools for the analysis of big, complex datasets. And computation plays a key role.
- Along with my students and collaborators, I spend most of my time developing, testing and applying new tools

What mess?

When

BIG DATA meets Small Signal

the signal has to be **teased out** of the data

- There are many ways to do this: a **small change** in the analysis details can cause a **large change** in the results.
- It is too easy to distort your findings, either by **fooling yourself** or **on purpose**.

The result

Journals are publishing many flawed, misleading or wrong papers (Ionannidis 2005)

8-11
NO, I'M NOT ASKING YOU TO LIE.
I'M ASKING YOU TO EMBELLISH THE TRUTH FOR THE SAKE OF MARKETING...
NOAH'S PUBLISHER
ucomics.com/nonsequitur
© '06 Wiley Ink, Inc.
Dist. by Universal Uclick
wileyink@earthlink.net

A big mess that will get worse

- In the age of big data, more than ever scientists need the ability to **examine, reproduce and judge** the strength of findings in a published study.
- But many (most?) published works cannot be reproduced (Ionannidis 2005)
- The **incentives** are all in the wrong direction: journals want to publish the latest, hottest research. Scientists want grants and tenure.
- Journals make it difficult (even **discourage**) critical feedback on published papers

Damage to the authors?

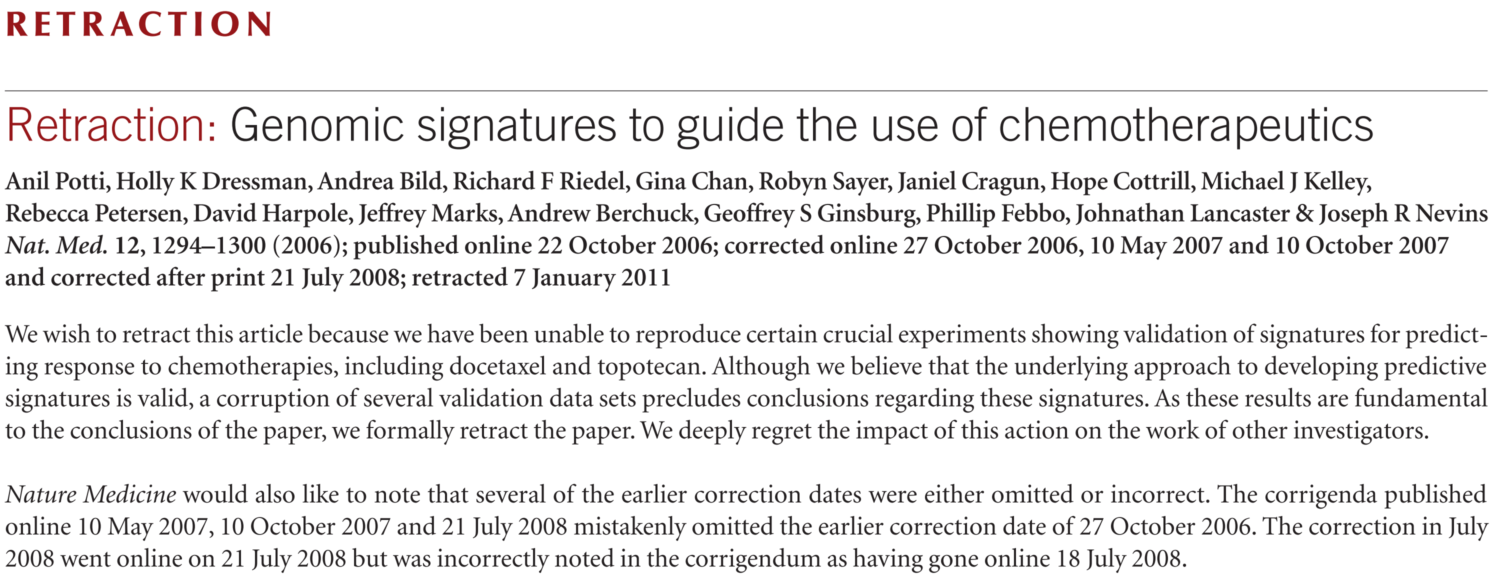
- There are **few negative consequences** when a flawed paper is published. It can take years for the flaws to be determined.
- By then the paper's findings are "**enshrined**" as truth (since it was **published!**), and people often continue to cite discredited papers years after the fact (Ionannidis 2005).
- And the authors have moved on to new work.

A domino effect

- Exaggerated claims by one group set the bar unrealistically high, and force exaggeration by the next group.
- Otherwise they will have difficulty publishing their work on the same topic

Two Examples

Predicting response to chemotherapies



RETRACTION

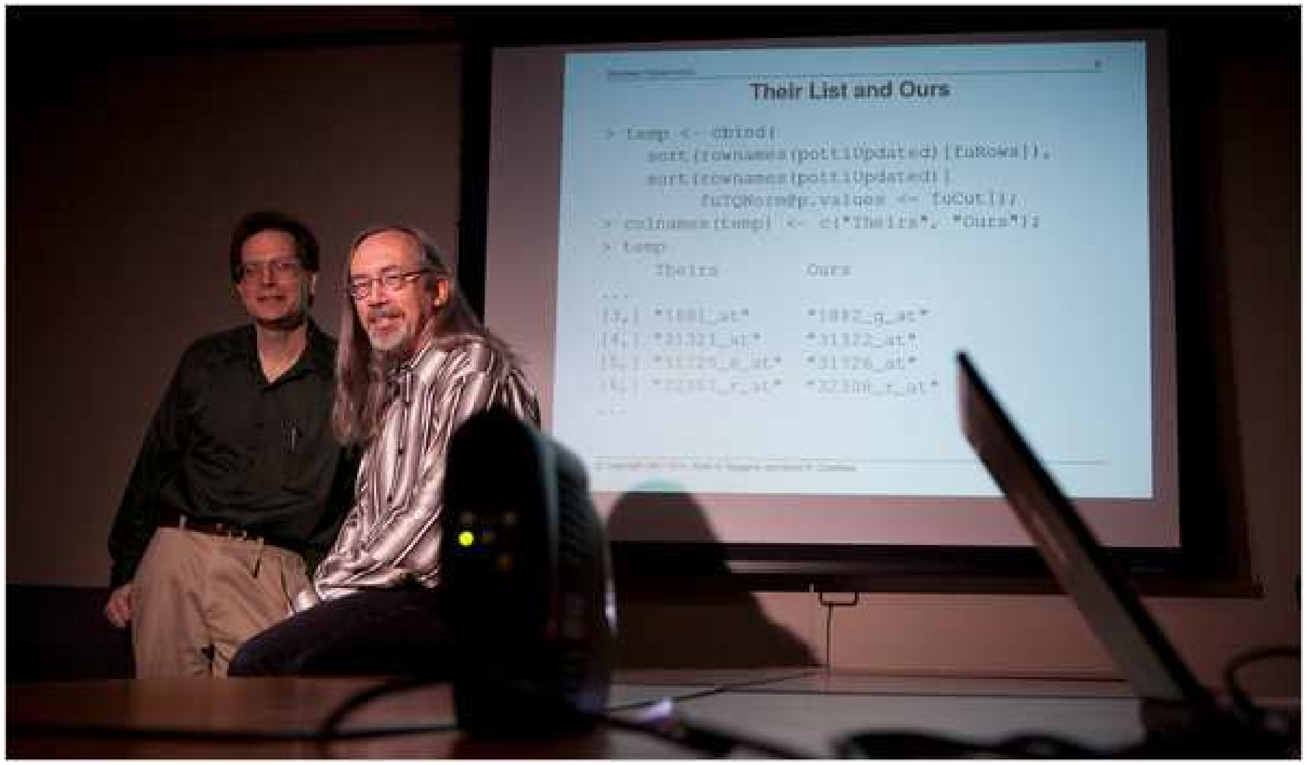
Retraction: Genomic signatures to guide the use of chemotherapeutics

Anil Potti, Holly K Dressman, Andrea Bild, Richard F Riedel, Gina Chan, Robyn Sayer, Janiel Cragun, Hope Cottrill, Michael J Kelley, Rebecca Petersen, David Harpole, Jeffrey Marks, Andrew Berchuck, Geoffrey S Ginsburg, Phillip Febbo, Johnathan Lancaster & Joseph R Nevins
***Nat. Med.* 12, 1294–1300 (2006); published online 22 October 2006; corrected online 27 October 2006, 10 May 2007 and 10 October 2007 and corrected after print 21 July 2008; retracted 7 January 2011**

We wish to retract this article because we have been unable to reproduce certain crucial experiments showing validation of signatures for predicting response to chemotherapies, including docetaxel and topotecan. Although we believe that the underlying approach to developing predictive signatures is valid, a corruption of several validation data sets precludes conclusions regarding these signatures. As these results are fundamental to the conclusions of the paper, we formally retract the paper. We deeply regret the impact of this action on the work of other investigators.

Nature Medicine would also like to note that several of the earlier correction dates were either omitted or incorrect. The corrigenda published online 10 May 2007, 10 October 2007 and 21 July 2008 mistakenly omitted the earlier correction date of 27 October 2006. The correction in July 2008 went online on 21 July 2008 but was incorrectly noted in the corrigendum as having gone online 18 July 2008.

Baggerley and Coombes

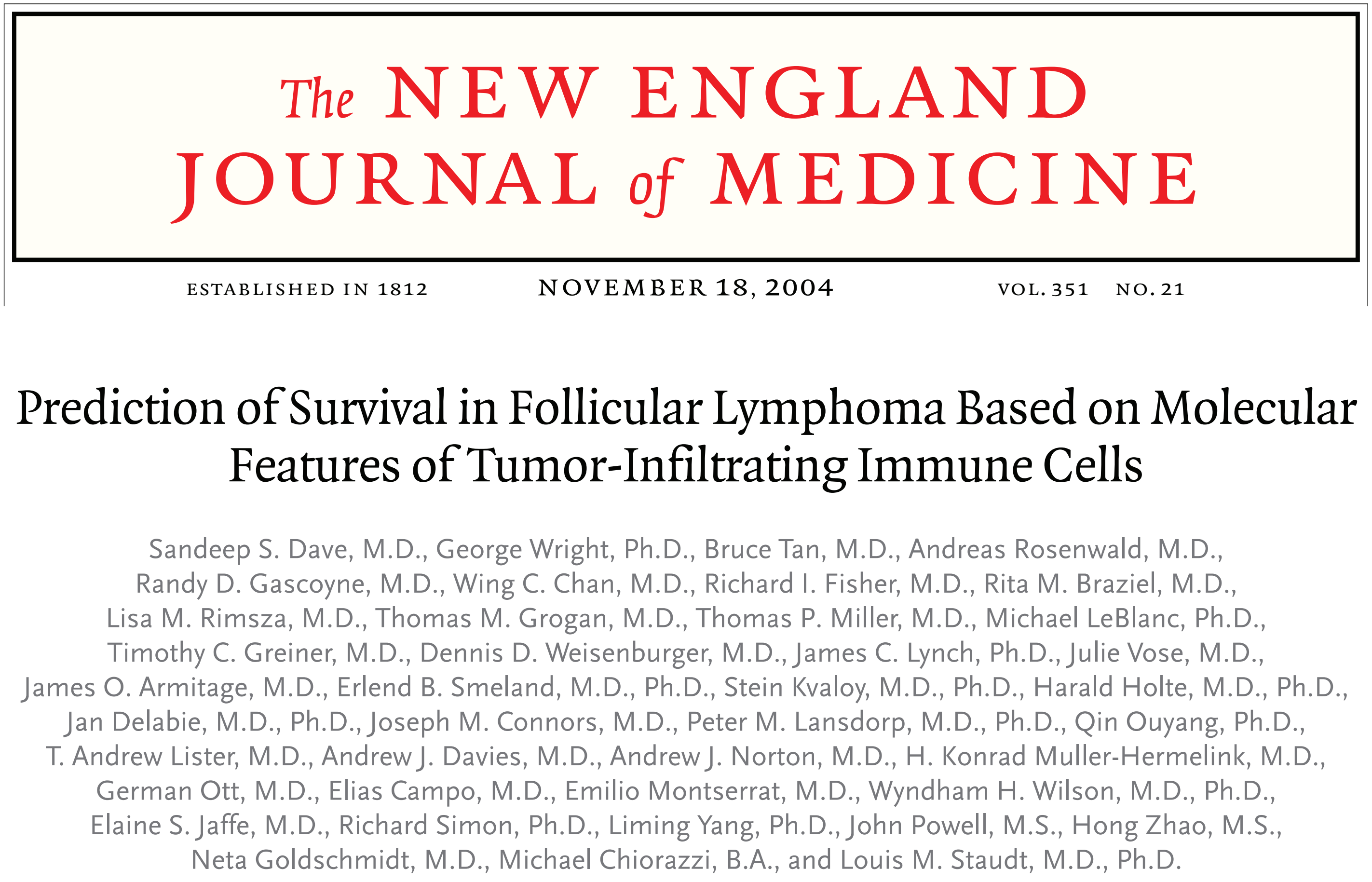


Their List and Ours

```
> temp <- cbind(
    sort(rownames(pottiUpdated)[fuRows]),
    sort(rownames(pottiUpdated)[
        fuTQNorm@p.values <= fuCut]));
> colnames(temp) <- c("Theirs", "Ours");
> temp
     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"
...
```

+ > 1500 hours!

One of my own experiences:

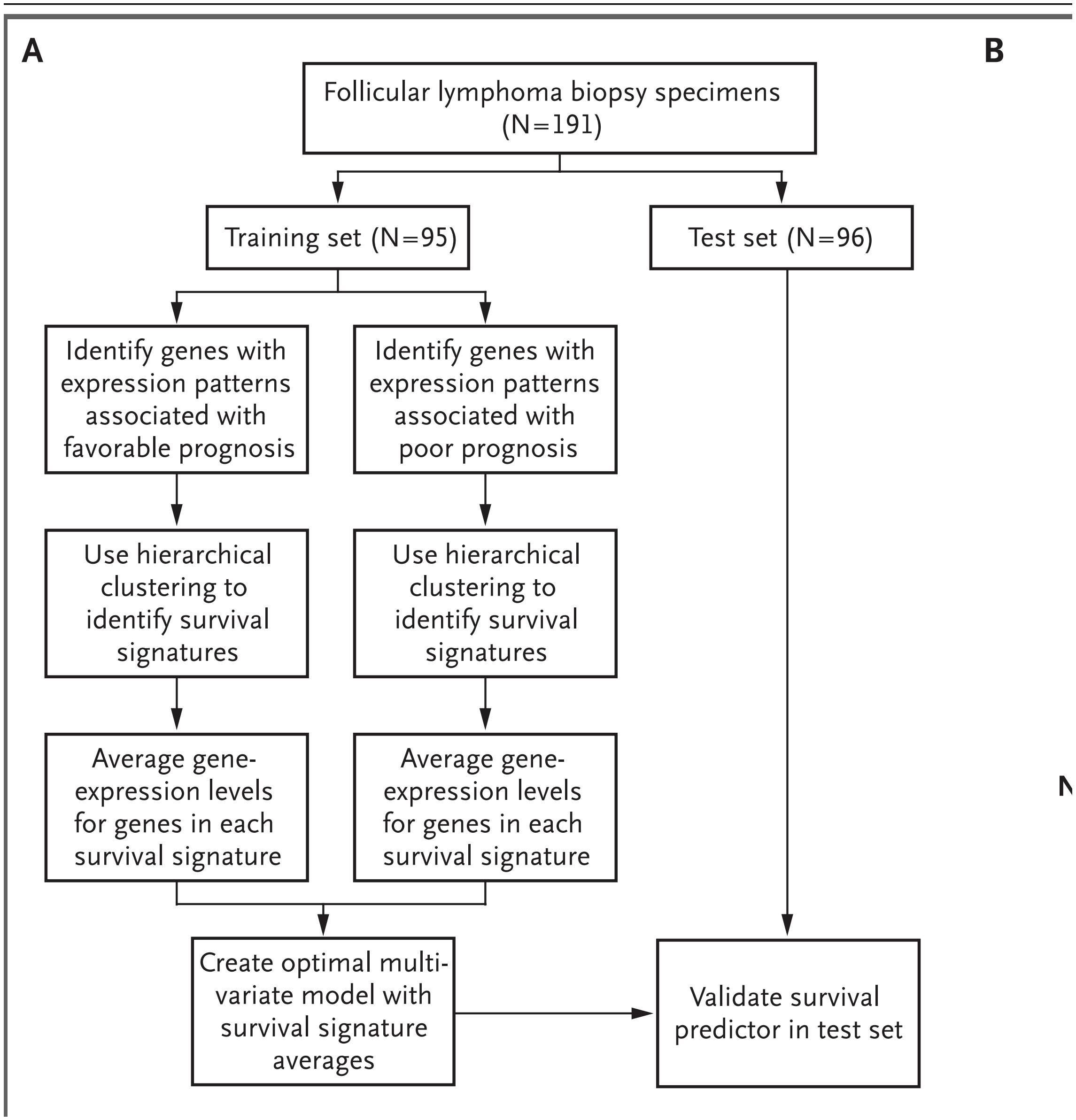


The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 NOVEMBER 18, 2004 VOL. 351 NO. 21

Prediction of Survival in Follicular Lymphoma Based on Molecular Features of Tumor-Infiltrating Immune Cells

Sandeep S. Dave, M.D., George Wright, Ph.D., Bruce Tan, M.D., Andreas Rosenwald, M.D.,
Randy D. Gascoyne, M.D., Wing C. Chan, M.D., Richard I. Fisher, M.D., Rita M. Braziel, M.D.,
Lisa M. Rimsza, M.D., Thomas M. Grogan, M.D., Thomas P. Miller, M.D., Michael LeBlanc, Ph.D.,
Timothy C. Greiner, M.D., Dennis D. Weisenburger, M.D., James C. Lynch, Ph.D., Julie Vose, M.D.,
James O. Armitage, M.D., Erlend B. Smeland, M.D., Ph.D., Stein Kvaloy, M.D., Ph.D., Harald Holte, M.D., Ph.D.,
Jan Delabie, M.D., Ph.D., Joseph M. Connors, M.D., Peter M. Lansdorp, M.D., Ph.D., Qin Ouyang, Ph.D.,
T. Andrew Lister, M.D., Andrew J. Davies, M.D., Andrew J. Norton, M.D., H. Konrad Muller-Hermelink, M.D.,
German Ott, M.D., Elias Campo, M.D., Emilio Montserrat, M.D., Wyndham H. Wilson, M.D., Ph.D.,
Elaine S. Jaffe, M.D., Richard Simon, Ph.D., Liming Yang, Ph.D., John Powell, M.S., Hong Zhao, M.S.,
Neta Goldschmidt, M.D., Michael Chiorazzi, B.A., and Louis M. Staudt, M.D., Ph.D.


A
B
Follicular lymphoma biopsy specimens (N=191)
Training set (N=95)
Test set (N=96)
Identify genes with expression patterns associated with favorable prognosis
Identify genes with expression patterns associated with poor prognosis
Use hierarchical clustering to identify survival signatures
Use hierarchical clustering to identify survival signatures
Average gene-expression levels for genes in each survival signature
Average gene-expression levels for genes in each survival signature
Create optimal multivariate model with survival signature averages
Validate survival predictor in test set

My findings

- After careful reconstruction of their analysis pipeline, **the signal disappeared** when training and test samples were swapped
- Same thing, when any one of a handful of tuning parameters were changed slightly
- My critique was published after much effort, squeezed down to 400 words!
- Authors never conceded anything was amiss

Journals are starting to listen

- **Nature's new policy**: Encourages authors to be transparent and **abolishes space restrictions** in the methods section.
- Other journals where reviewers sometimes ask to run code: **PLOS Comp Bio**, **Nature Biotechnology**
- **JNCI**: "Methods descriptions should be succinct but sufficiently detailed to allow replication of the study by a qualified investigator." **NEJM**- microarray data must be made available: nothing specific about analysis scripts.

Statistics journals

They are not leading the charge!

- **J. Stat. Software** requires code; **Biostatistics** rewards papers that provide code with special “kite markings”.
- No other major statistics journal (to my knowledge) has code requirements.

A written description is not good enough

- There is a **big difference** between a written description of what was done, and an actual software script to reproduce the Figures and Tables

- in my experience as a “forensic statistician”, it can take many weeks or months **and the cooperation of the authors** to go from one to the other

What we need to do

- the scientific community as a whole has to **embrace the importance of reproducibility**— researchers, granting agencies and journals
- authors should make **serious efforts** to provide data and software scripts (analogous to a wet lab protocol) so that readers can **easily** reproduce their results and judge the strength of their findings

What we need to do: continued

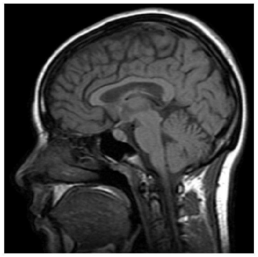
- Biostatisticians and bioinformaticians should **insist** on this at their own institutions: **don't put your name on a paper** if there is no clear and complete and clear analysis script.
- This might be painful!

Software may be the key

- We need software environments to make reproducibility easier for researchers. (we can't expect all researchers to have the sophistication for R)
- It is NOT OK to use proprietary software that reviewers or readers cannot access

One initiative

"**Verifiable Computational Research**" project: M. Gavish and D. Donoho (Stanford): `www.verifiable-research.org`.



Published figures have a barcode; when scanned, it links to a website that reproduces the figures with time-stamped data and code

We all need to step up

- I hope that this joint Stanford-Oxford collaboration will set a good example
- The "**big cheeses**" in the field (some of you are here), need to step up and make their own analyses reproducible.

SHORT TERM PAIN FOR LONG TERM GAIN