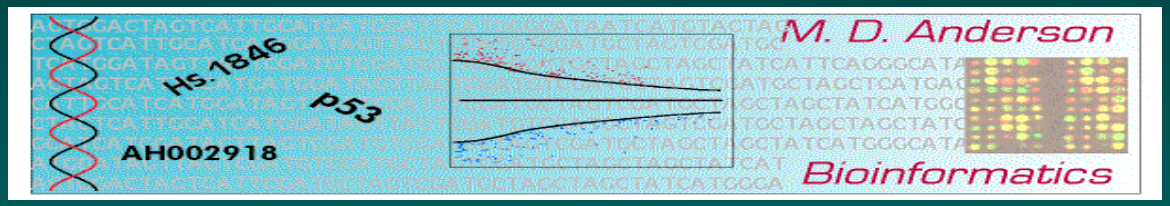
Rigor and Reproducibility: Introduction, Analysis, and Statistics

Keith A. Baggerly
Bioinformatics and Computational Biology
kabagg@gmail.com

GCC Rigor and Reproducibility, May 14, 2026



Science: Trust, Because we can Verify

Memories of Chem Lab

Color change, precipitation, charge differences

Classic experiments have been done over and over again

Biology and medicine can have additional jitters

Science benefits from a large scale presumption of regularity

Challenges at the Frontier

Tools and Instrumentation have gotten better

Our questions have gotten more subtle

Full descriptions of what we did are more involved

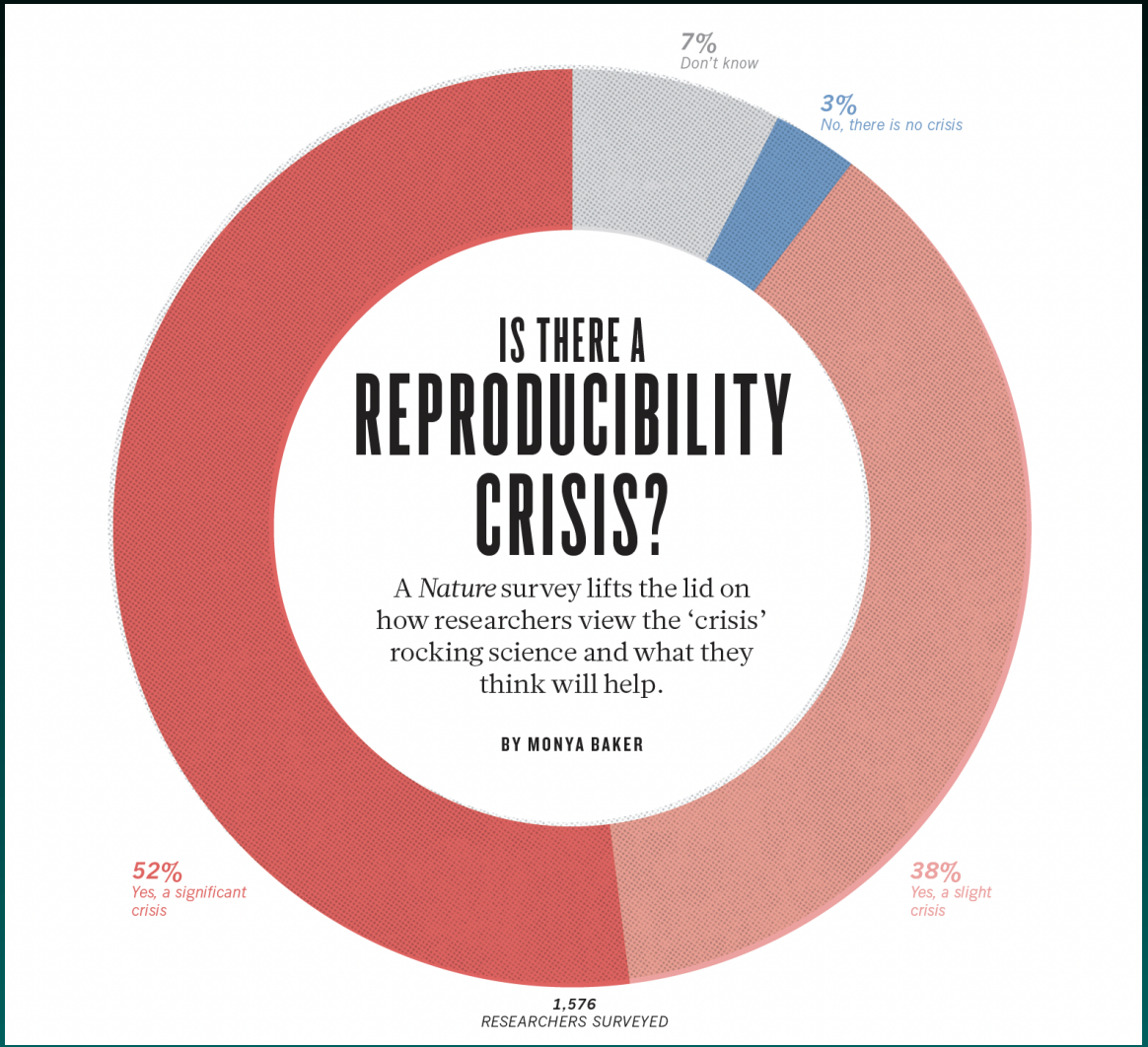
Analysis and interpretation of observations

Distinguishing “reproduction” from “replication”

Occasionally, the process fails

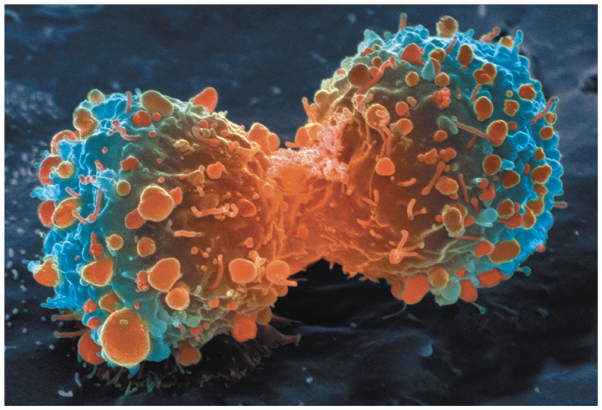
What if it fails more than “occasionally”?

Houston, We Have a Crisis



Monya Baker 2016, Nature, 533:452-4.

This Didn't Come Out of Nowhere



Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

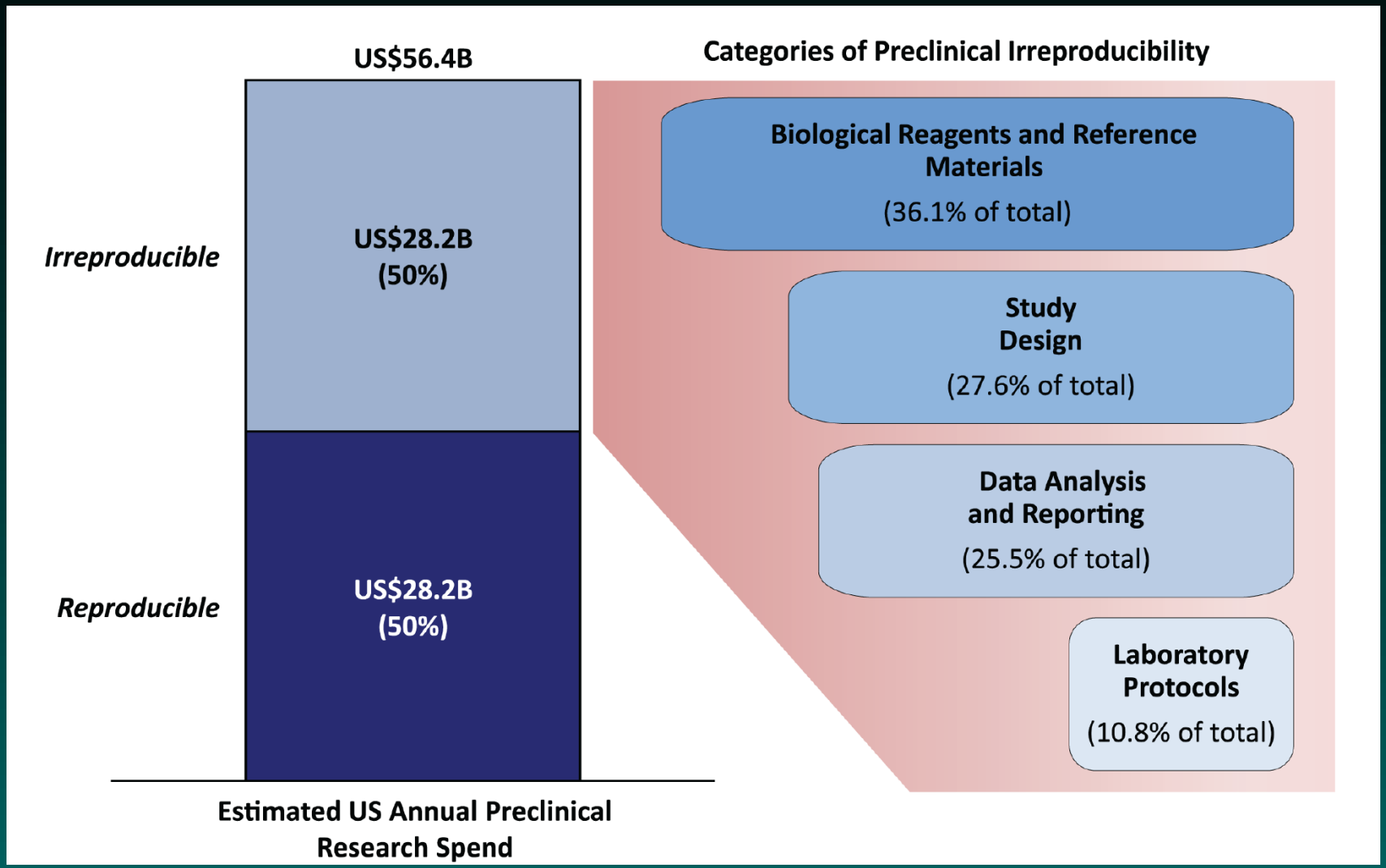
Begley & Ellis 2012, Nature, 481:531-3. Amgen 6/53

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

Prinz et al 2011, Nat Rev Drug Disc, 10:328-9. Bayer 14/67

Some Cost Breakdowns (Take 1)



Freedman et al (2015), PLoS Biology, 13(6):e1002165

Why is this such a Problem?

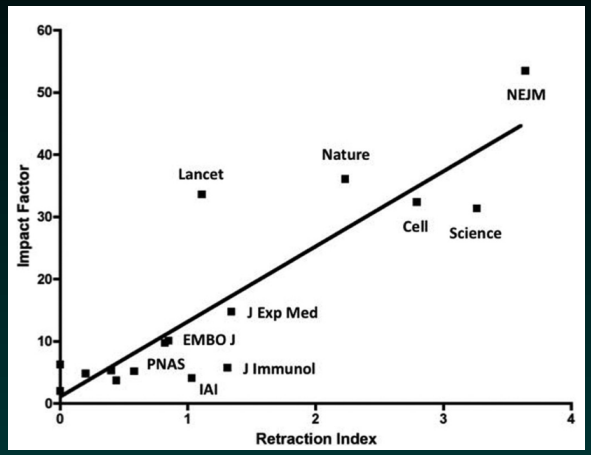
On the Folly of Rewarding A, While Hoping for B

STEVEN KERR
Ohio State University

Kerr 1975, Acad Mgmt J, 18(4):769-83

What gets rewarded in academia?
Publish or Perish!
High Impact Publications: “CNS”!

Perverse Evidence of Incentives



Fang & Casadevall 2011, Infec & Immun, 79(10):3855-9.

RESEARCH ARTICLE

Do individual and institutional predictors of misconduct vary by country? Results of a matched-control analysis of problematic image duplications

Daniele Fanelli[1]*, Matteo Schleicher[1], Ferric C. Fang[2], Arturo Casadevall[3], Elisabeth M. Bik[4]

Fanelli et al 2022, PLoS One, 17(3):e0255334.

See also: Retraction Watch and Elisabeth Bik!

Funders: Bringing B Closer to A

2013: OSTP states data sharing reqs are coming

2014: NIH and Journals introduce common checklists

2016: NIH reviewers must score study design

2023: NIH reqs for Data Mgmt and Sharing

2022: OSTP: federally funded research will be open access

Education:
Describe extent and nature of the problems
Make fixes more apparent

Why is Reproducibility Important in H-T B?

Our intuition about what “makes sense” is very poor in high-d.

To use “omics-based signatures” as biomarkers, we need to know they’ve been assembled correctly.

Without documentation, we may need to employ (lengthy!) *forensic bioinformatics* to infer what was done.

Let’s look at examples in the context of a specific problem: *can we predict which patients will respond to which chemotherapeutics?*

Using Cell Lines to Predict Sensitivity

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, 12:1294-300.

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

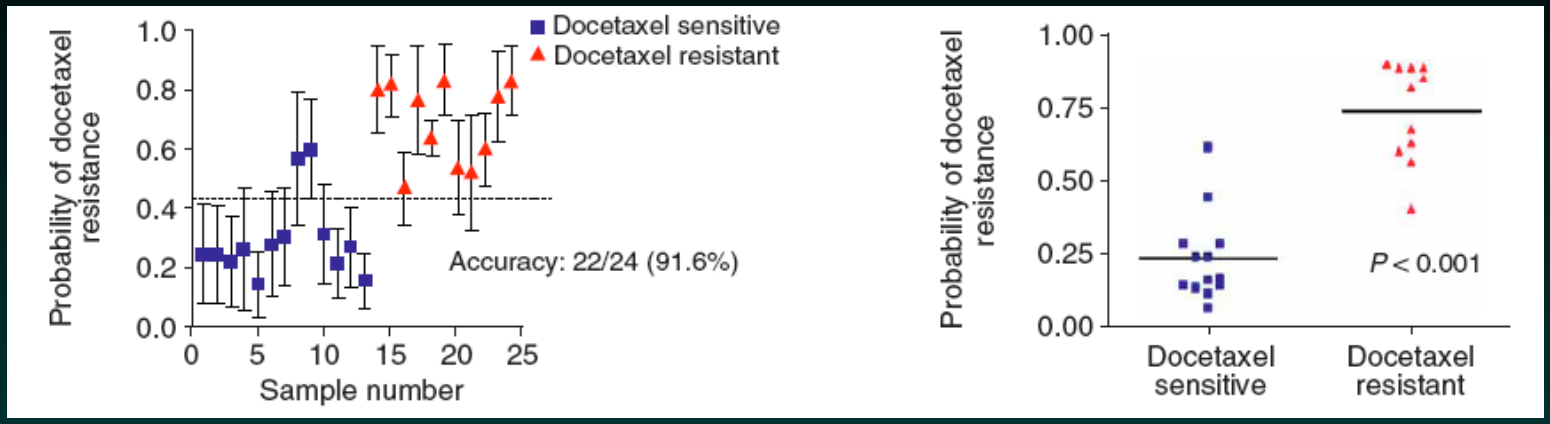
They provide examples using 7 commonly used agents.

This got people at MDA very excited.

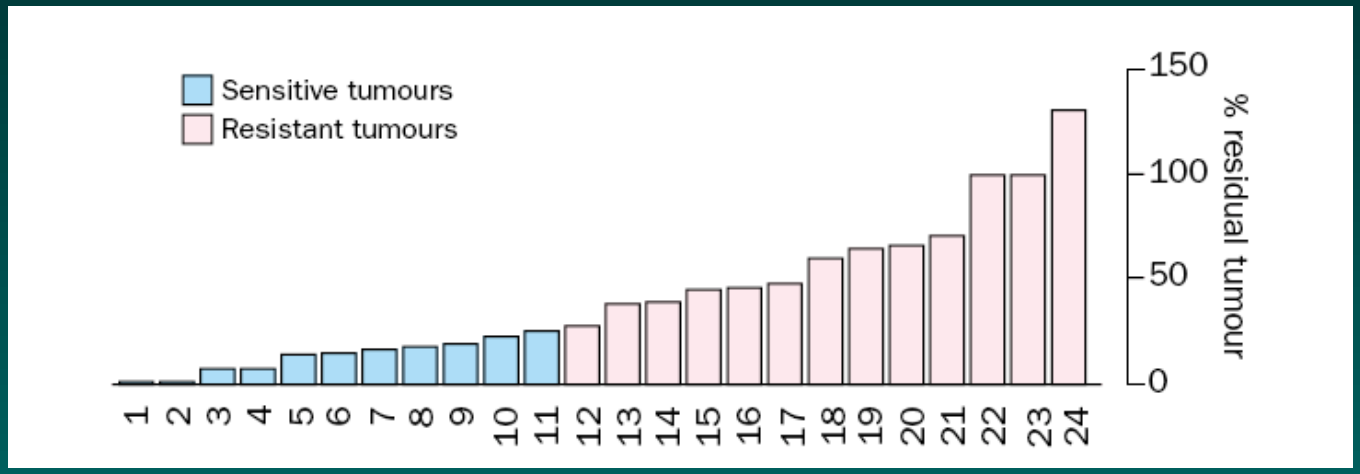
Their Gene List and Ours

```
> 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"
...
```

Predicting Response: Docetaxel

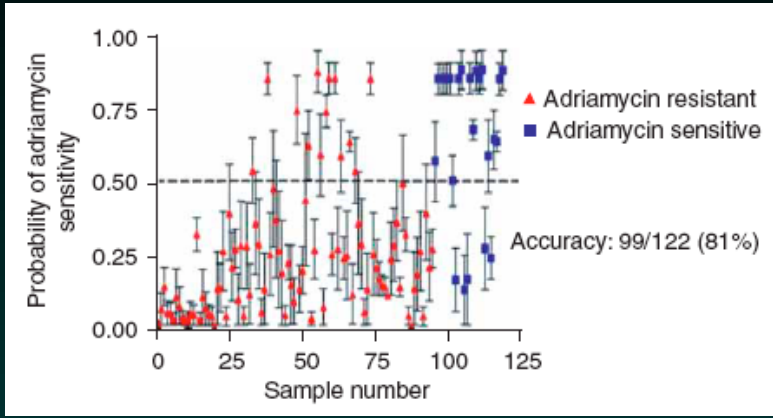


Potti et al (2006), Nature Medicine, 12:1294-300, Fig 1d

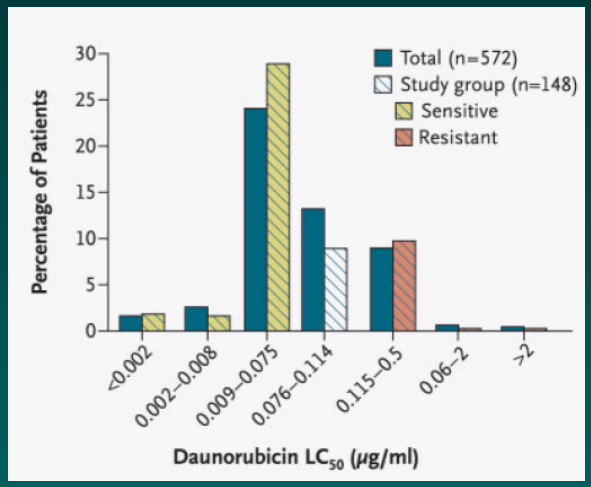


Chang et al, Lancet 2003, 362:362-9, Fig 2 top

Predicting Response: Adriamycin



Potti et al (2006), Nature Medicine, 12:1294-300, Fig 2c



Holleman et al, NEJM 2004, 351:533-42, Fig 1

Partial Timeline

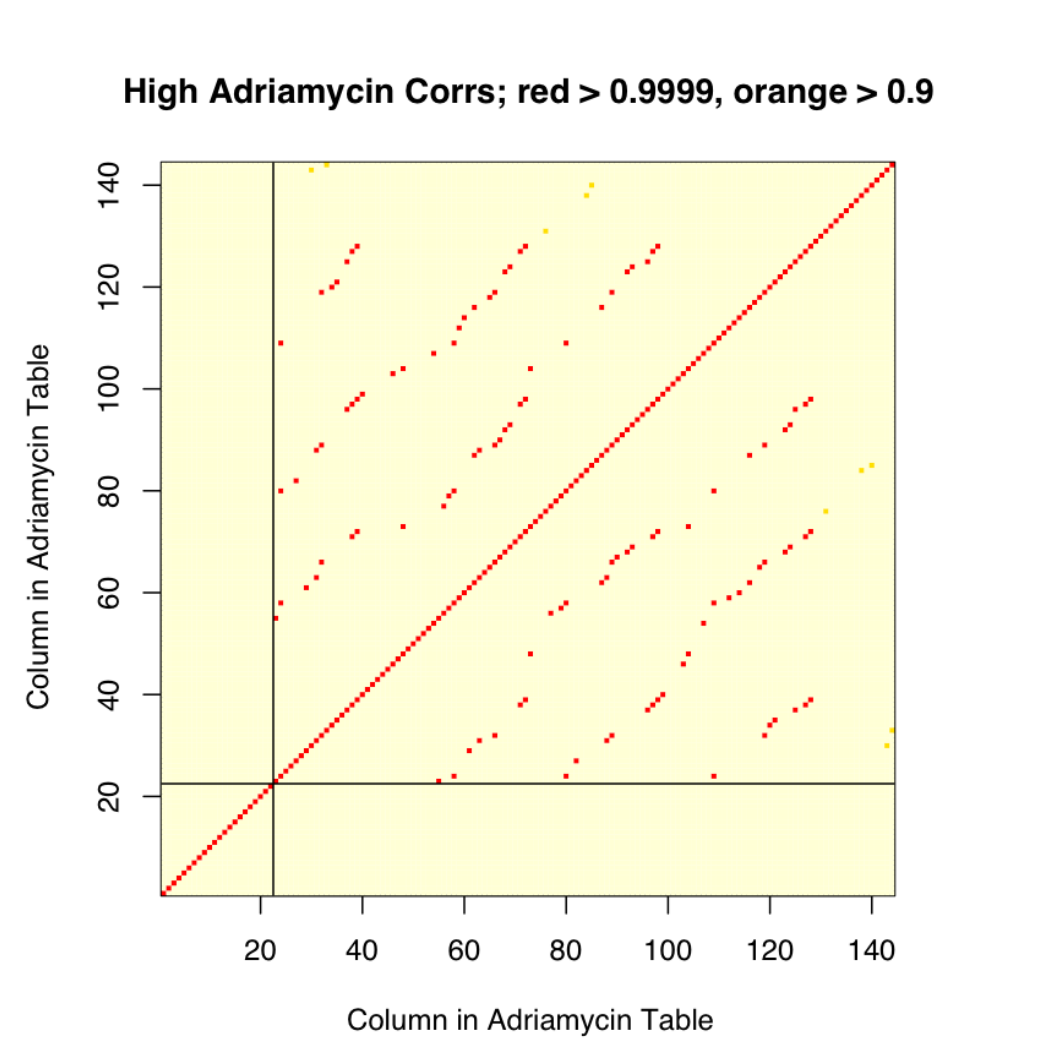
2006:

* Nov 8: Our first questions to Potti and Nevins.
* Nov 21: Our first report describing errors.
* Nov-Dec: More reports/questions: Nov 27, Dec 4, 13, 27.

2007:

* Jan 24: We meet with Nevins at M.D. Anderson. We urge him to review the data.
* Feb-Apr: New data and code are posted. Some numbers change. We tell them we don't think it works.
* Apr 25: We send Potti and Nevins a draft for comment.
* May: We find problems with outliers. Potti and Nevins continue to insist it works, and want to "bring this to a close".

Adriamycin 0.9999+ Correlations



Redone Aug 08, “using ... 95 unique samples”.

Validation 1: Hsu et al

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. ERCC1, ERCC4 and DNA repair genes are identified as “important”.

With some work, we matched the heatmaps. (Gene lists?)

The 4 We Can't Match

203719_at, ERCC1,
210158_at, ERCC4,
228131_at, ERCC1, and
231971_at, FANCM (DNA Repair).

Another problem –

The 4 We Can't Match

203719_at, ERCC1,
210158_at, ERCC4,
228131_at, ERCC1, and
231971_at, FANCM (DNA Repair).

Another problem –

The last two probesets aren't on the U133A arrays that were used. They're on the U133B.

Validation 2: Bonnefoi et al

Validation of gene signatures that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the EORTC 10994/BIG 00-01 clinical trial

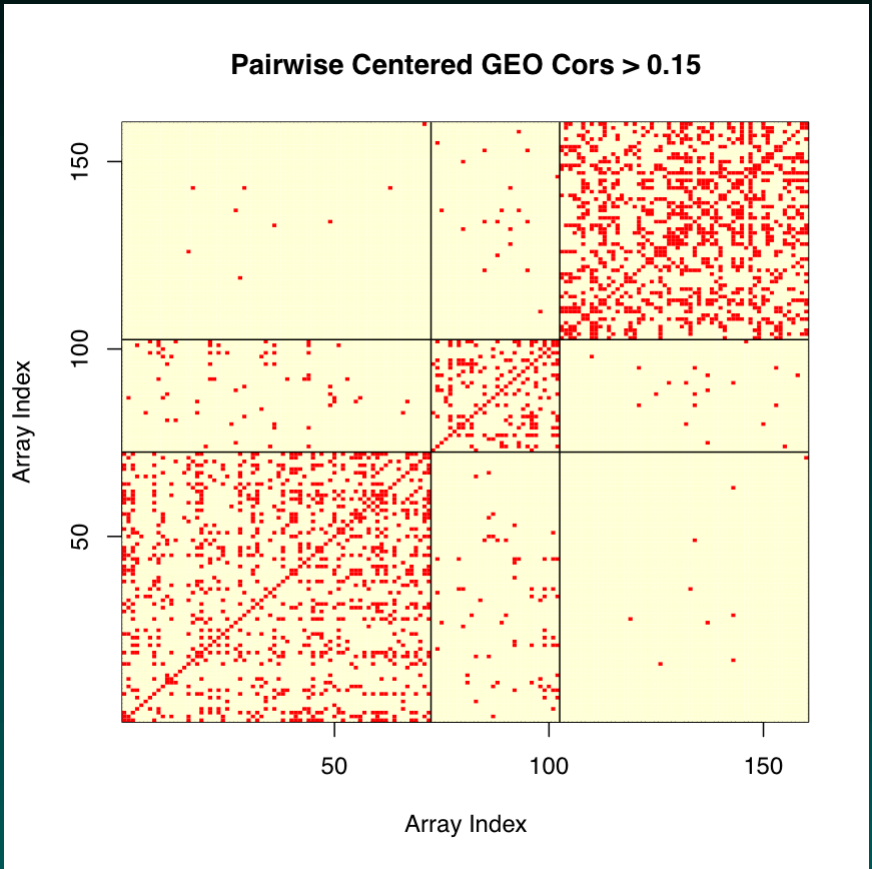
Hervé Bonnefoi, Anil Potti, Mauro Delorenzi, Louis Mauriac, Mario Campone, Michèle Tubiana-Hulin, Thierry Petit, Philippe Rouanet, Jacek Jassem, Emmanuel Blot, Véronique Becette, Pierre Farmer, Sylvie André, Chaitanya R Acharya, Sayan Mukherjee, David Cameron, Jonas Bergh, Joseph R Nevins, Richard D Iggo

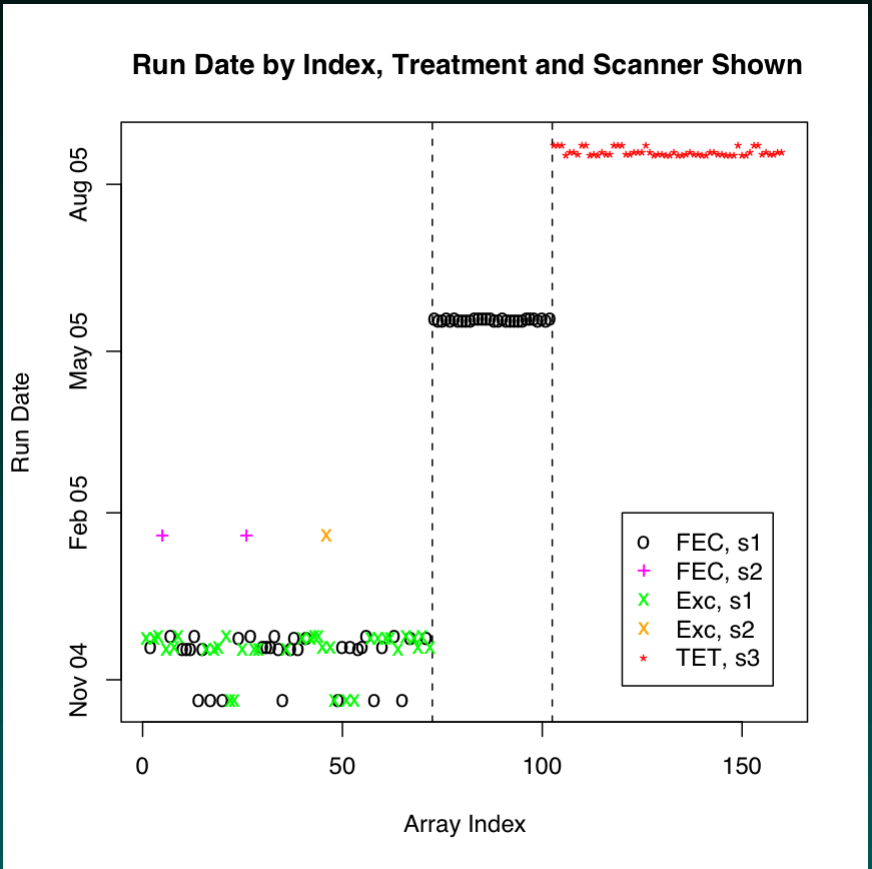
Lancet Oncology, Dec 2007, 8:1071-8. (early access Nov 14)

Similar approach, using signatures for Fluorouracil, Epirubcin (used Adriamycin), Cyclophosphamide, and Taxotere (Docetaxel) to predict response to one of two combination therapies: FEC and TET.

Potentially improves ER- response from 44% to 70%!

We Might Expect Some Differences...





High Sample Correlations

Array Run Dates

See Leek et al, Nat Rev Genet, 2010 for more examples.

How Are Results Combined?

Potti et al predict response to TFAC, Bonnefoi et al to TET and FEC. Let P() indicate prob sensitive. The rules used?

How Are Results Combined?

Potti et al predict response to TFAC, Bonnefoi et al to TET and FEC. Let P() indicate prob sensitive. The rules used?

$$P(TFAC) = P(T)+P(F)+P(A)+P(C)-P(T)P(F)P(A)P(C).$$

How Are Results Combined?

Potti et al predict response to TFAC, Bonnefoi et al to TET and FEC. Let P() indicate prob sensitive. The rules used?

$$P(TFAC) = P(T)+P(F)+P(A)+P(C)-P(T)P(F)P(A)P(C).$$

$$P(ET) = \max[P(E), P(T)].$$

How Are Results Combined?

Potti et al predict response to TFAC, Bonnefoi et al to TET and FEC. Let P() indicate prob sensitive. The rules used?

$$P(TFAC) = P(T)+P(F)+P(A)+P(C)-P(T)P(F)P(A)P(C).$$

$$P(ET) = \max[P(E), P(T)].$$

$$P(FEC) = \frac{5}{8}[P(F) + P(E) + P(C)] - \frac{1}{4}.$$

How Are Results Combined?

Potti et al predict response to TFAC, Bonnefoi et al to TET and FEC. Let P() indicate prob sensitive. The rules used?

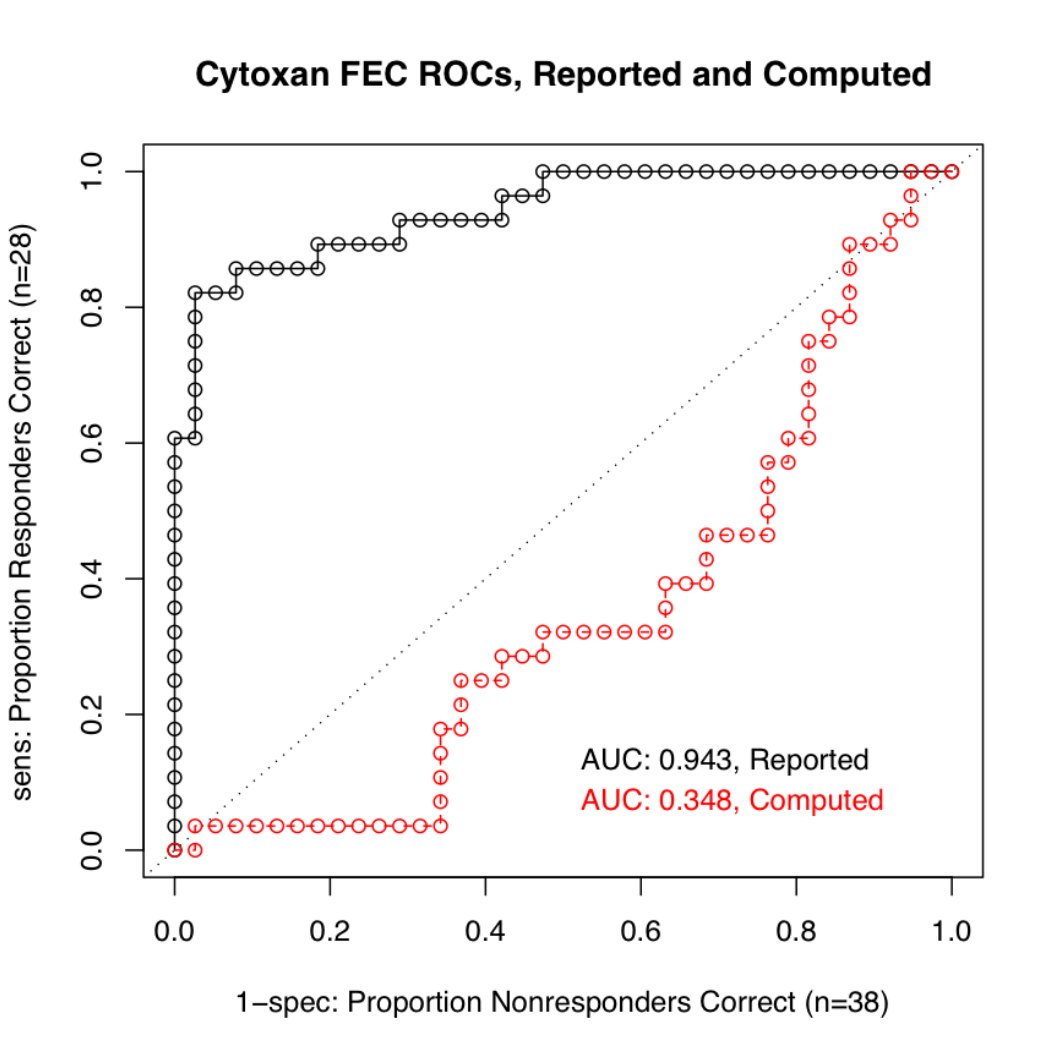
$$P(TFAC) = P(T)+P(F)+P(A)+P(C)-P(T)P(F)P(A)P(C).$$

$$P(ET) = \max[P(E), P(T)].$$

$$P(FEC) = \frac{5}{8}[P(F) + P(E) + P(C)] - \frac{1}{4}.$$

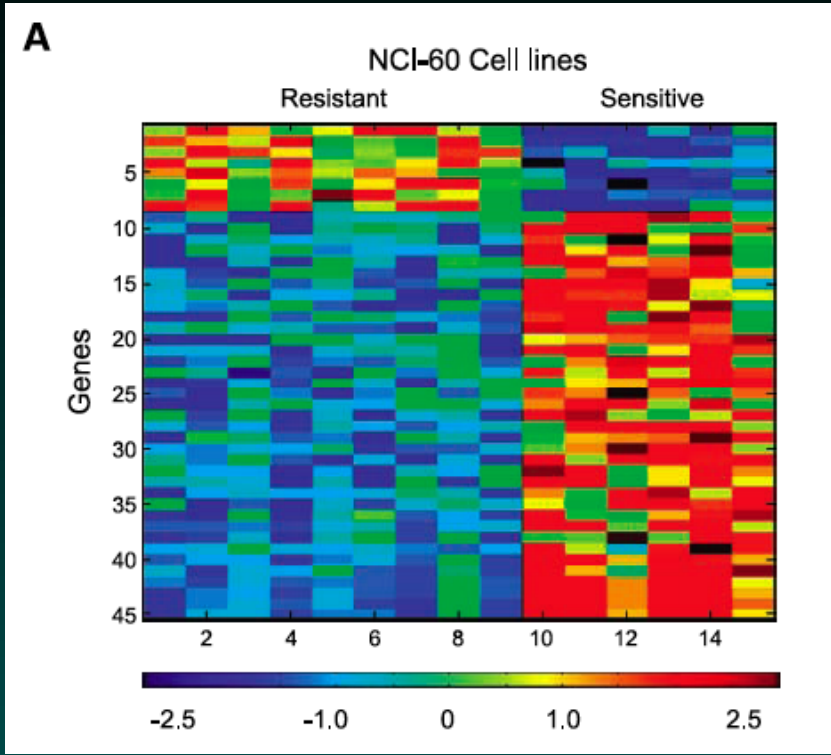
Each rule is different.

Predictions for Individual Drugs?



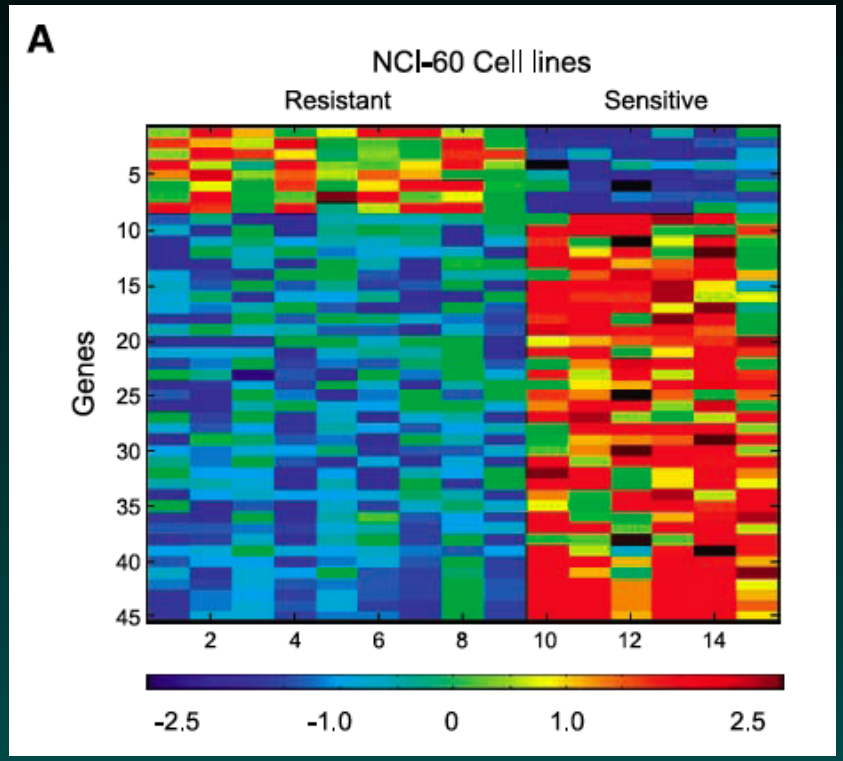
Does cytoxan make sense?

Temozolomide Heatmaps

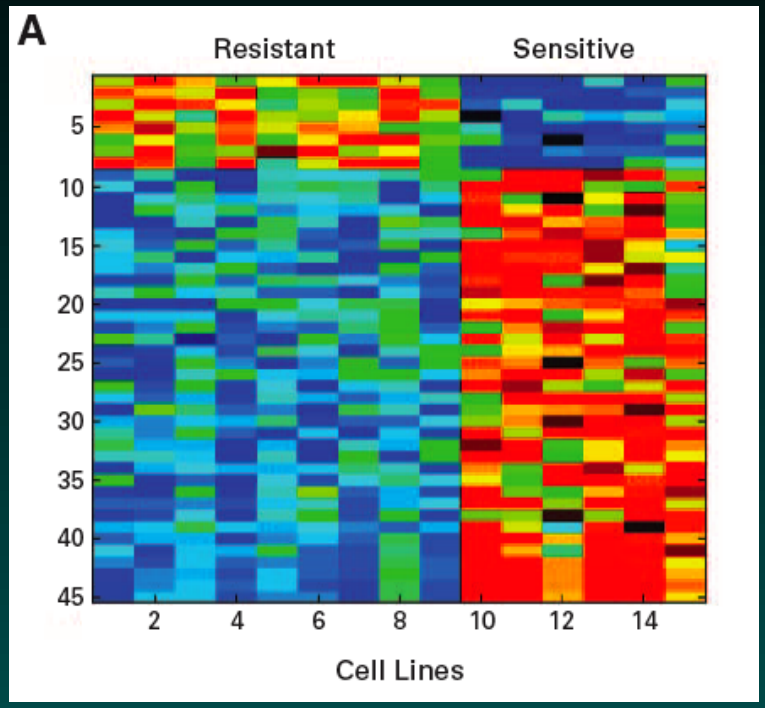


Augustine et al., 2009, *Clin Can Res*, **15**:502-10, Fig 4A.
Temozolomide, NCI-60.

Temozolomide Heatmaps



Augustine et al., 2009, *Clin Can Res*, **15**:502-10, Fig 4A. Temozolomide, NCI-60.



Hsu et al., 2007, *J Clin Oncol*, **25**:4350-7, Fig 1A. Cisplatin, Gyorffy cell lines.

Why We Care

Jun 2009: we learn clinical trials had begun.
2007: pemetrexed vs cisplatin, pem vs vinorelbine.
2008: docetaxel vs doxorubicin, topotecan vs dox (Moffitt).

Why We Care

Jun 2009: we learn clinical trials had begun.
2007: pemetrexed vs cisplatin, pem vs vinorelbine.
2008: docetaxel vs doxorubicin, topotecan vs dox (Moffitt).

Sep 1, 2009: We submit a paper describing case studies to the *Annals of Applied Statistics*.

Sep 14, 2009: Paper accepted and available online at the *Annals of Applied Statistics*.

Sep-Oct 2009:

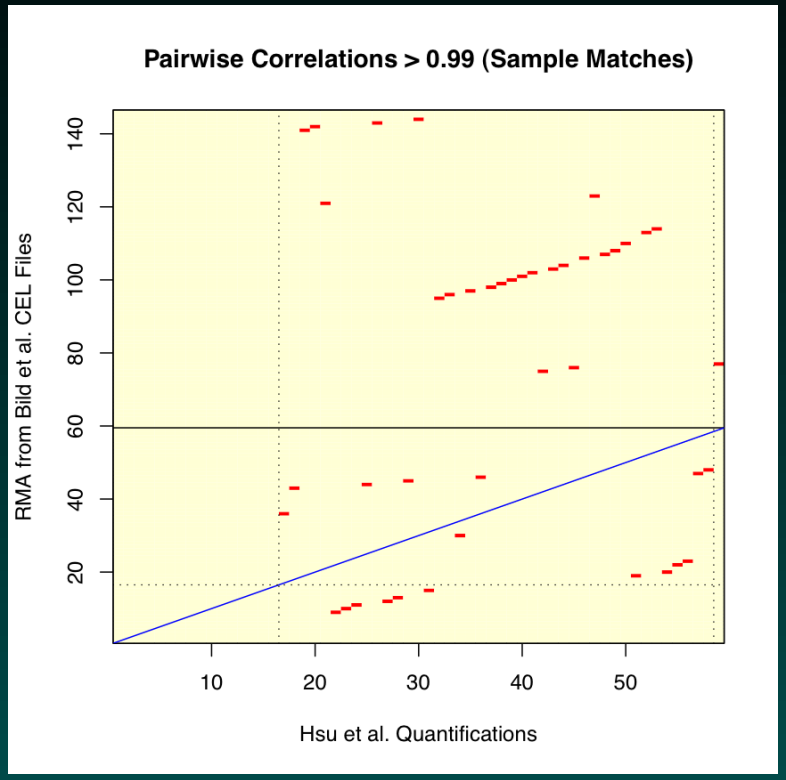
Story covered by *The Cancer Letter*; Oct 2, Oct 23.
NCI raises concerns with Duke's IRB behind the scenes.
Duke starts internal investigation, suspends trials.

New Data

Early-Nov ’09 (mid-investigation), the Duke team posted new data for cisplatin and pemetrexed (in lung trials since ’07).

These included quantifications for the 59 ovarian cancer test samples (from GSE3149, which has 153 samples) they used to validate their predictor.

We Tried Matching The Samples



43 samples are mislabeled.

16 samples don't match because the genes are mislabeled.

All of the validation data are wrong.

We reported this to Duke and to the NCI in mid-November.

Jan 29, 2010

THE CANCER LETTER

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

Duke In Process To Restart Three Trials Using Microarray Analysis Of Tumors

By Paul Goldberg

Duke University said it is in the process of restarting three clinical trials using microarray analysis of patient tumors to predict their response to chemotherapy.

Their investigation's results *"strengthen ... confidence in this evolving approach to personalized cancer treatment."*

We Asked for the Data

"While the reviewers approved of our sharing the report with the NCI, *we consider it a confidential document*" (Duke). A *future paper* will explain the methods.

This did give us one more option...

We Asked for the Data

"While the reviewers approved of our sharing the report with the NCI, *we consider it a confidential document*" (Duke). A *future paper* will explain the methods.

This did give us one more option...

In May 2010, we obtained a copy of the reviewers' report from the NCI under FOIA (Cancer Letter, May 14).

We (and others) didn't think it justified restarting trials.

There was no mention of our Nov 2009 report.

A Catalyzing Event: July 16, 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

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

Oct 22/9: First call for paper retraction.

Nov 9: Duke terminates trials.

Nov 19: call for Nat Med retraction, Potti resigns

Other Developments

117 patients were enrolled in the trials.
Sep, 2011: Patient lawsuits filed (11+ settlements).

Misconduct investigation (Jul 2010-Nov 2015).
10/6+ 10 full/partial retractions, FDA Review

Jul 8, 2011: Front Page, NY Times.

Feb 12, 2012: 60 Minutes

Mar 23, 2012: IOM Report Released

April/May, 2015: Last lawsuits settled

Nov 9, 2015: Official ORI finding of fraud

Mar 21, 2018: NIH imposes new requirements on Duke

Some Cautions/Observations

This case is pathological.

But we've seen similar problems before.

The most common mistakes are simple.

Confounding in the Experimental Design
Mixing up the sample labels
Mixing up the gene labels
Mixing up the group labels
(Most mixups involve simple switches or offsets)

This simplicity is often hidden.

Incomplete documentation

This is not an Isolated Problem

Ioannidis et al. (2009), *Nat. Gen.*, **41**:149-55. Tested reproducibility of microarray papers. Could reproduce 2/18.

Begley and Ellis (2012), *Nature*, **483**:531-3. Amgen attempted replication of clinical "breakthroughs" prior to further study. Validated 6/53.

NCI focus meeting Sep 2012.

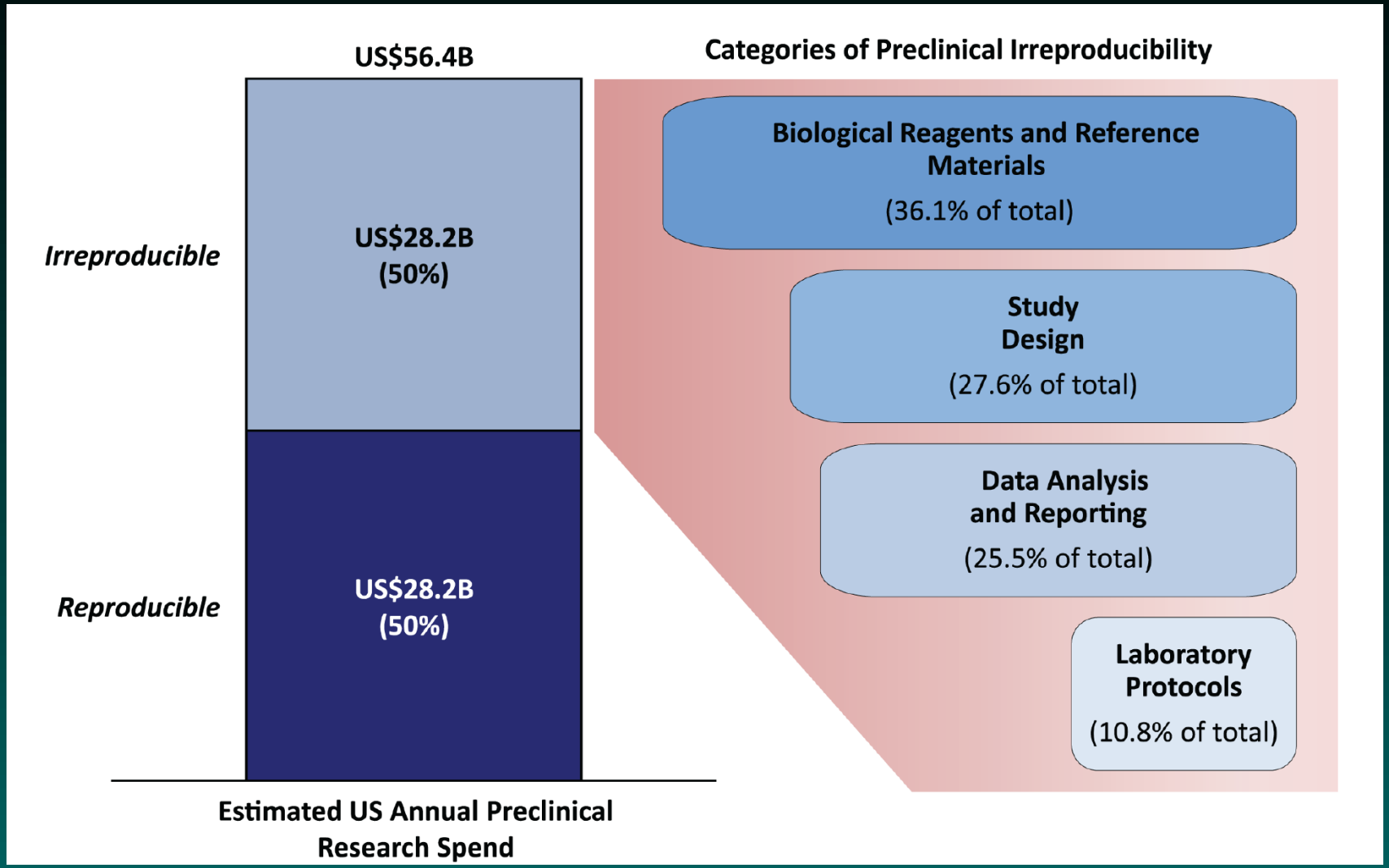
Collins and Tabak (2014), *Nature*, **505**:612-3.

Rigor and Reproducibility, NIH, 2016

SISBID RR Short Course July, 2015, 2016, 2017, 2018

GCC Short Course (YouTube), Parts 1, 2, 3

Some Cost Breakdowns (Take 2)



Freedman et al (2015), PLoS Biology, 13(6):e1002165

What Have We and Others Suggested?

Exploiting a Teachable Moment...

Baggerly et al *Nature* (2010)

Give us your data, your code, your huddled masses

Records of data provenance

Checking existence as a task for journals and reviewers
(are there links? are they live?)

NCI Guidelines in *Nature* Oct 2013

NIH Guidance Jan 2023

Some Things NIH Asks For Now

- Cell line identity verification (Labels correct?)
- Discussion of experimental design (Confounding avoided?)
- Data sharing plan (Told us what you did?)

What might I look for in an R01?

This isn't impossible.

We've done it.

It's easier to do today than when we started.

(GCC Reproducibility Short Course, 2018)

This Time is Different...

F.D.A. Blocked Publication of Research Finding Covid and Shingles Vaccines Were Safe

The agency's scientists and data contractors reviewed millions of patient records for studies that were pulled back before release.

NY Times, May 5, 2026

Jake Scott paper links and commentary

Einstein: The right to search for truth implies also a duty;

one must not conceal any part
of what one has recognized to be true.

Other Recent RR Challenges

NEJM, Data Parasites, and Immunotherapy

COVID-19, Real-Time Research, and Surgisphere

Protein Folding, AI, and AlphaFold

CASP14 - Blinded Validation, Prespecified Success Metric

Inherent Randomness and Black Box Evaluation
Statistical Agreement and Clinical Trials

Will AI Solve Everything?
Well-Annotated Gold Standard Datasets

Will people act according to what the data suggest?

Sholto David and the whistleblower option!

Reasons for Hope

1. Our Own (Evolving!) Experience
2. Better tools (knitr, markdown, GitHub, the tidyverse)
3. Journals, Code and Data
4. The IOM, the FDA, and IDEs*
5. The NCI and Trials it Funds
6. Project TIER (see protocols, course materials)
7. Center for Open Science
8. NIH Rigor and Reproducibility Initiative

Acknowledgments

Kevin Coombes

Yang Zhao, Ying Wang, Shelley Herbrich
Shannon Neeley, Jing Wang
David Ransohoff, Gordon Mills
Jane Fridlyand, Lajos Pusztai, Zoltan Szallasi

M.D. Anderson Ovarian, Lung and Breast SPOREs

Baggerly and Coombes (2009), *Annals of Applied Statistics*, **3(4)**:1309-34.
http://bioinformatics.mdanderson.org/Supplements/ReproRsch-All/Modified/StarterSet

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

Thanks!