Statistical inference enables bad science; Statistical thinking enables good science.

Christopher Tong, Ph.D.

The 41st Midwest Biopharmaceutical Statistics Workshop

Indianapolis, Indiana

15 May 2018

Disclaimers

- The views expressed do **not** necessarily represent the policies, views, or opinions of my employer or the United States, our hosts, or organizers.
- Work on this presentation was carried out on the author's personal time, **not** as part of my official duties.
- *Acknowledgements*: Bert Gunter and Frank Harrell.

Synopsis

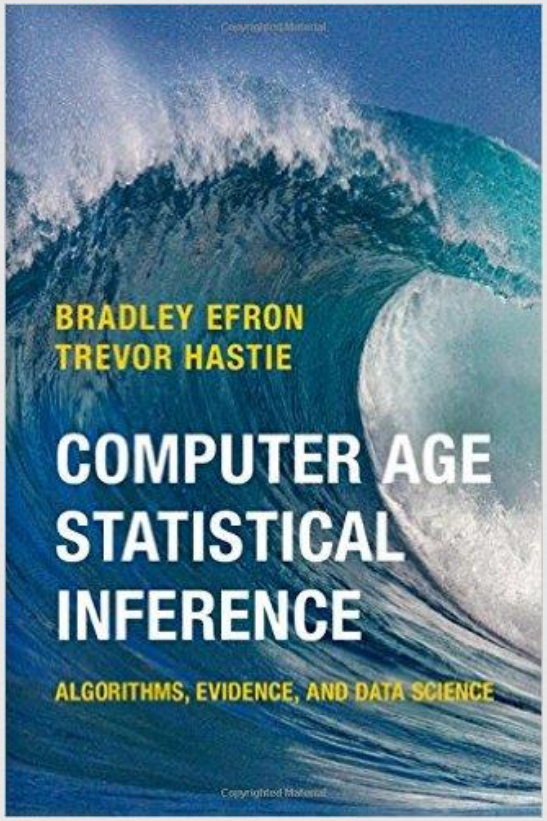
- Statistical inference is typically invalid due to the potential for overfitting when the data are used to make decisions on statistical model building, model criticism, and model selection, all illustrating unquantifiable uncertainty about the model itself. Technical solutions offered by statisticians, and their deficiencies, are discussed. The distinction between exploratory and confirmatory research is made; statistical inference can only be valid in confirmatory studies, when the experimental protocol and statistical analysis plan are completely prespecified and adhered to. This distinction is fully realized in human clinical trials. The remainder of scientific research is largely exploratory, and should not be strait-jacketed by the high bar needed to guarantee valid statistical inference. Most of the great discoveries in the history of science did not need statistical inference. Rather, David Freedman's “shoe leather” approach and Andrew Ehrenberg's “many sets of data” provide better avenues for doing science. Statistics has much to offer science in experimental design and execution, enlightened data display, and disciplined data exploration, and should abandon its focus on the misleading attempt to confer judgments about statistical significance.

Outline

- Model uncertainty and the Optimism Principle.
- Exploratory vs. confirmatory objectives.
- Technical solutions and their deficiencies.
- The Cult of the Isolated Study.
- Statistics without statistical inference?

Model uncertainty and the Optimism Principle

Efron and Hastie (2016)



- On the first page of the first chapter:
- After introducing the sample mean and its standard error:
- **"It is a surprising, and crucial, aspect of statistical theory that the same data that supplies an estimate can also assess its accuracy."**

Do they really believe this?

For prediction **No**

(p. 227: data splitting)

For inference, **Yes, with caveats**

(ch. 20: "Inference after model selection")

The risk of overfitting

- The statistical model is rarely fully pre-specified, but must be constructed after data collection is completed.
- Statistical model building and model criticism is a multi-stage, interactive process.
- Every chance you have to use the data to improve the model fit is also an opportunity to **overfit** the data at hand.
  - "Researcher degrees of freedom" (Simmons et al., *Psych. Sci.*, 2011). "Double dipping" (Kriegeskorte et al., *Nature Neuroscience*, 2009).

The Optimism Principle

- The "optimism principle" was coined by Picard & Cook (*JASA*, 1984):
  - Due to overfitting, a model chosen by a selection process provides a more optimistic explanation of the "training data" than of other data that arise by a similar mechanism.
- It was foreshadowed by T.C. Koopmans (*Econometrica*, 1949) and articulated by Mosteller & Tukey (*Data Analysis and Regression*, 1977).
- Studied by many statisticians and econometricians throughout the 1980s and early 1990s.
  - Via simulation and asymptotic theory; a few real data sets.
  - Freedman's Paradox (*TAS*, 1983).

J. R. Statist. Soc. A (1995)
158, *Part* 3, *pp.* 419–466

Model Uncertainty, Data Mining and Statistical Inference

By CHRIS CHATFIELD†

University of Bath, UK

[*Read before* The Royal Statistical Society *on Wednesday, January 18th, 1995, the President,* Professor D. J. Bartholomew, *in the Chair*]

4.3. *Some General Consequences*

Model selection biases are hard to quantify, but the following general points can be made.

(a) *Least squares theory does not apply when the same data are used to formulate and fit a model.* Yet time series text-books, for example, customarily apply least squares methods to time series models even when the model has been selected as the best fitting model from a wide class of models such as ARIMA models.

(b) After model selection, *estimates of model parameters and of the residual variance are likely to be biased.*

(c) The analyst typically *thinks that the fit is better than it really is* (the optimism principle), and diagnostic checks rarely reject the best fitting model *because it is the best fit*!

(d) *Prediction intervals are generally too narrow.*

J Quant Criminol (2010) 26:217–236
DOI 10.1007/s10940-009-9077-7

ORIGINAL PAPER

Statistical Inference After Model Selection

Richard Berk · Lawrence Brown · Linda Zhao

The Statistical Crisis in Science

Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't hold up.

Andrew Gelman and Eric Loken

American Scientist, 102: 460-465 (2014).

BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 55, Number 1, January 2018, Pages 31–55
http://dx.doi.org/10.1090/bull/1597
Article electronically published on October 4, 2017

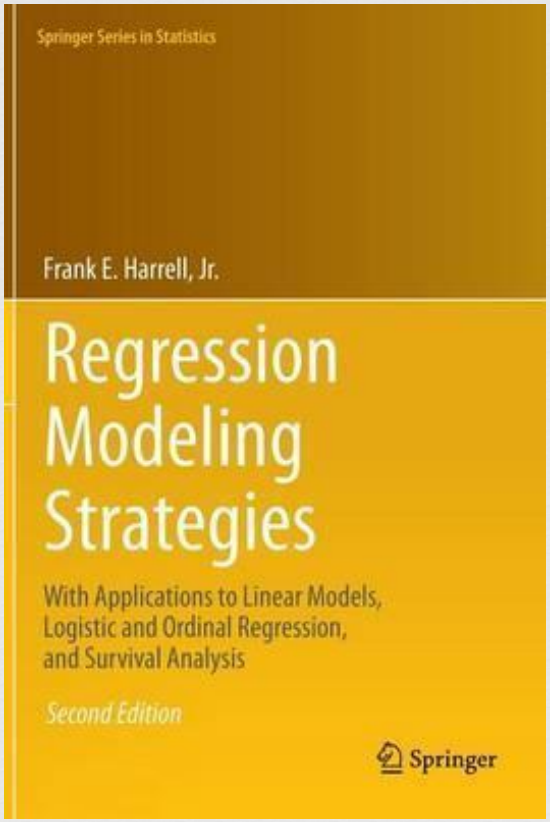
STATISTICAL PROOF?
THE PROBLEM OF IRREPRODUCIBILITY

SUSAN HOLMES

Gelman's comments on the ASA Statement on P-values

- "…knowledge of how many analyses were conducted etc. is not enough. The whole point of the 'garden of forking paths'…is that to compute a valid *p*-value you need to know what analyses *would have been done* had the data been different."
- "Ultimately the problem is not with *p*-values but with null hypothesis significance testing…Whenever this sort of reasoning is being done, the problems discussed above will arise. Confidence intervals, credible intervals, Bayes factors, cross-validation: you name the method, it can and will be twisted, even if inadvertently, to create the appearance of strong evidence where none exists."
- "Statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an 'uncertainty laundering'…"

Frank Harrell (2001/2015)



- "Using the data to guide the data analysis is almost as dangerous as not doing so."

Exploratory vs. confirmatory objectives

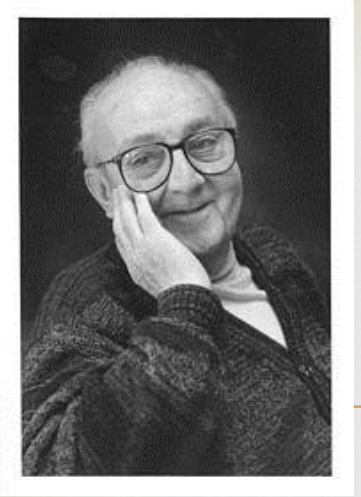
We Need Both Exploratory and Confirmatory

JOHN W. TUKEY*

The American Statistician, Vol. 34, No. 1. (Feb., 1980), pp. 23-25.

Flexibility vs. pre-specification

- Statistical inference is valid only when the study protocol and statistical analysis plan are comprehensively pre-specified prior to data collection, and adhered to during and afterward.
  - This can only be done once a great deal is known about the problem! (Confirmatory studies only.)
- Most research is done long before it is possible to entertain such rigid pre-specification.
  - Exploratory studies: flexible design and flexible analysis must be used if we are to learn what the data has to say, rather than remaining committed to prior expectations.
  - Gelman & Loken (2014) "do not want demands of statistical purity to strait-jacket our science". The price for this flexibility is that valid statistical inference is undermined.



Box's comment on Draper (1995)

J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 45–97

Assessment and Propagation of Model Uncertainty

By DAVID DRAPER†

University of Bath, UK

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, March 16th, 1994,* Professor V. S. Isham *in the Chair*]

George Box (University of Wisconsin, Madison): This is a fine example of how far a first-class statistician can go in the careful reanalysis of dead data sets. But I believe that he should aspire to more.

Statistics has no reason for existence except as a catalyst for scientific enquiry in which only the last stage, when all the creative work has already been done, is concerned with a final fixed model and a rigorous test of conclusions. The main part of such an investigation involves an inductive–deductive iteration with input coming from the subject-matter specialist at every stage. This requires a continuously developing model in which the identity of the measured responses, the factors considered, the structure of the mathematical model, the number and nature of its parameters and even the objective of the study change.

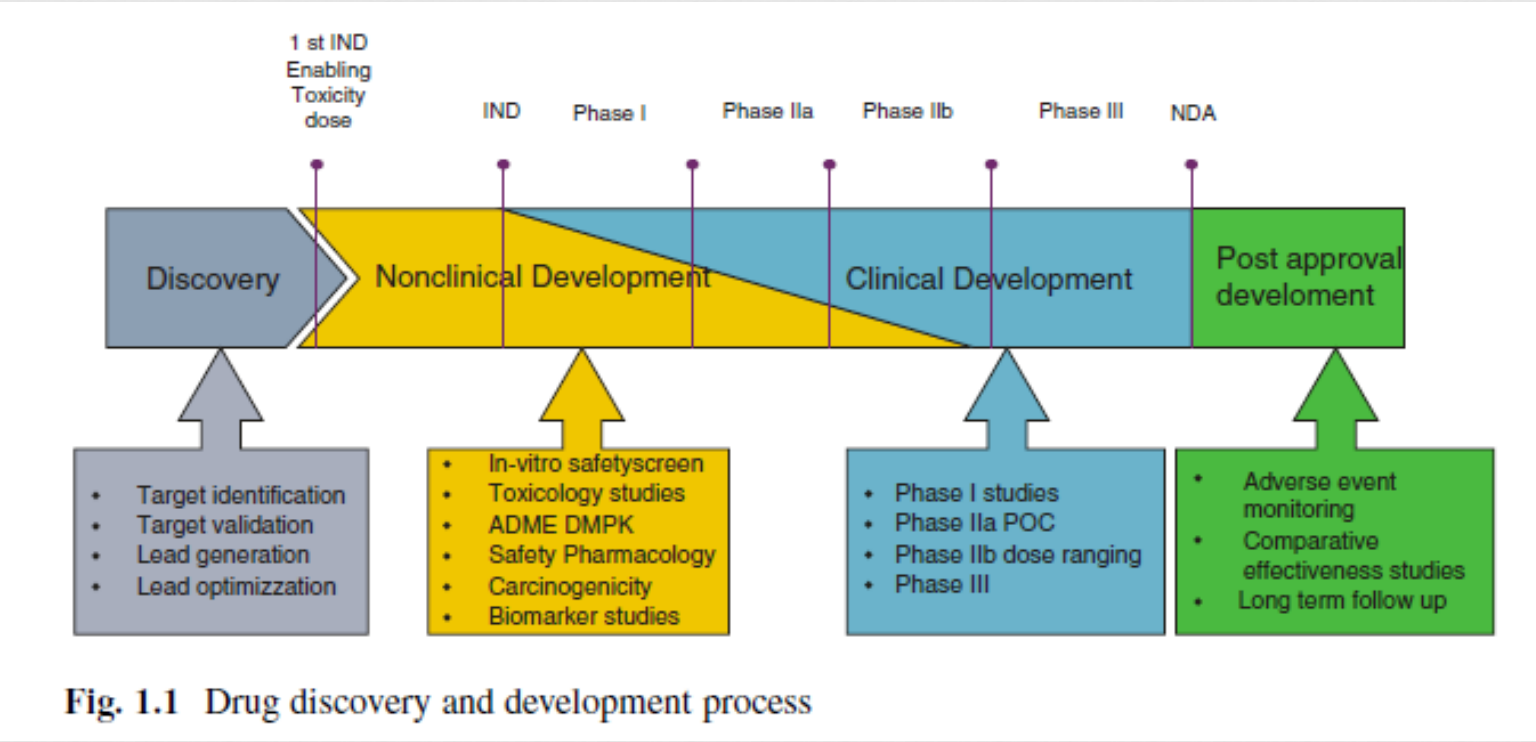


Fig. 1.1 Drug discovery and development process

From Zhang & Su, chapter 1, *Nonclinical Statistics for Pharmaceutical and Biotechnology Industries* (ed. by L. Zhang, Springer, 2016).

Earlier phases can accommodate flexible designs/analyses;
Later phases require pre-specified designs/analyses.
(ICH E8 and E9)

Correlation between Development Phases and Types of Study

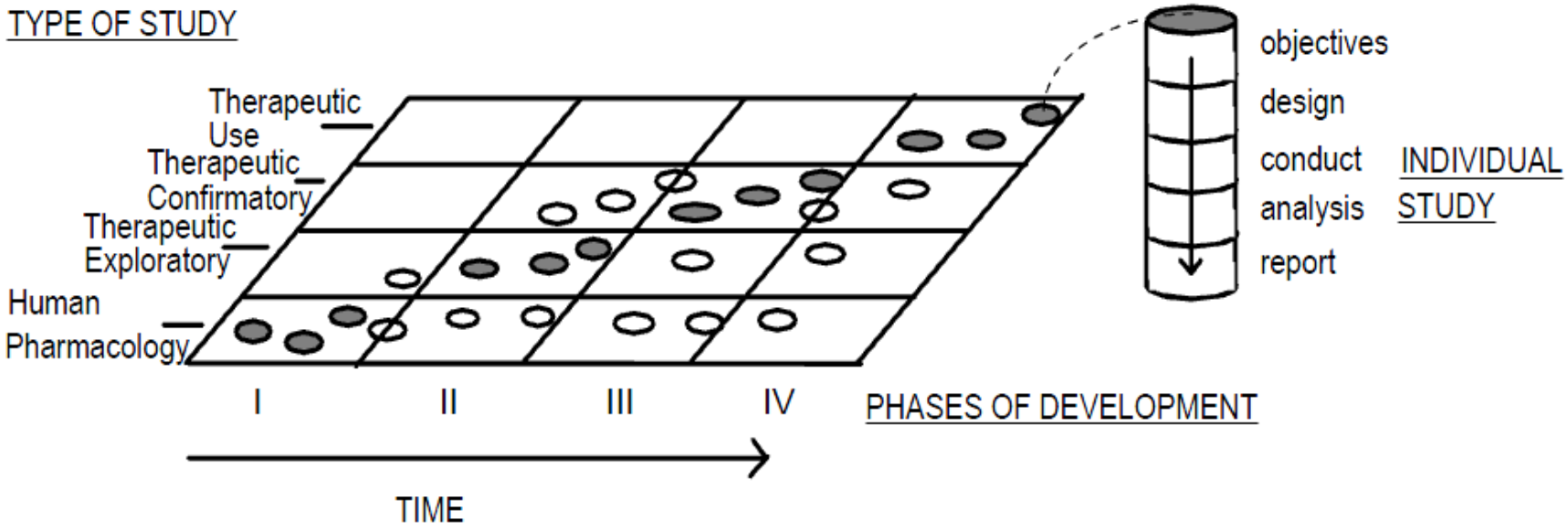


Figure 1 *- This matrix graph illustrates the relationship between the phases of development and types of study by objective that may be conducted during each clinical development of a new medicinal product. The shaded circles show the types of study most usually conducted in a certain phase of development, the open circles show certain types of study that may be conducted in that phase of development but are less usual. Each circle represents an individual study. To illustrate the development of a single study, one circle is joined by a dotted line to an inset column that depicts the elements and sequence of an individual study.*

From ICH E8

Some observations

- While statistical inferences are often reported in earlier phase studies, they are *not* taken as definitive, and licensing decisions are rarely made on their basis.
  - This *shift in attitude* on how to interpret statistical inferences is largely absent in preclinical research, where *p*-values drive publication decisions.
- Subgroup analysis: you should always do them but never believe the results (Sir Richard Peto).

At *Science*, the paradigm is changing. We're talking about asking authors, 'Is this hypothesis testing or exploratory?' An exploratory study explores new questions rather than tests an existing hypothesis. But scientists have felt that they had to disguise an exploratory study as hypothesis testing, and that is totally dishonest. I have no problem with true exploratory science. That is what I did most of my career. But it is important that scientists call it as such and not try to pass it off as something else. If the result is important and exciting, we want to publish exploratory studies, but at the same time make clear that they are generally statistically underpowered, and need to be reproduced.

-Dr. Marcia McNutt, now **President of the National Academy of Sciences.** (Former editor-in-chief, *Science*.)

Source: E. R. Shell, (2016) *Science*, 353: 116-119.

Technical solutions and their deficiencies

Technical solutions (examples)

- Multiplicity adjustments, false discovery rate, false coverage rate.
- Penalization methods and regularization.
- Data splitting (training set/test set).
- Cross-validation (*k*-fold) and resampling methods (bootstrap).
  - Must specify automated model selection procedure.
- Model averaging (Bayesian and frequentist).
- Post-selection inference.

General objection

- Obtaining "more than one set of data, whenever possible, is a potentially more convincing way of overcoming model uncertainty and is needed anyway to determine the range of conditions under which a model is valid." (Chatfield, *JRSS A* 1995).

The Cult of the Isolated Study

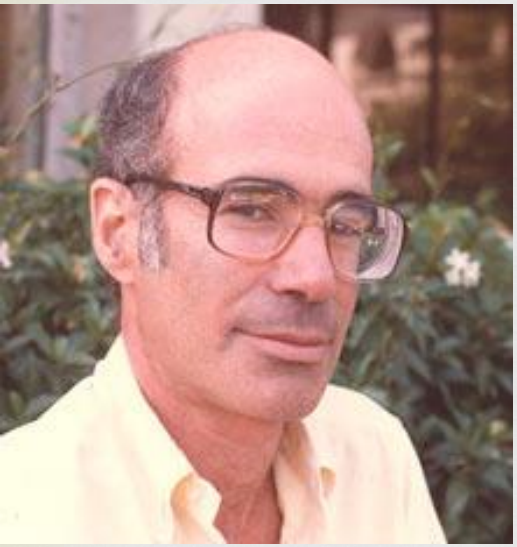
The Cult of the Isolated Study

- John A. Nelder (*JRSS A*, 1986) coined this term.
- Andrew Ehrenberg (*TAS*, 1990) advocated "Many Sets of Data" which "seems the only way in which we can produce results that are generalizable, lawlike, and predictable—which in fact hold for many different sets of data."
- "Medicine is a conservative science and behavior usually does not change on the basis of one study." -- Steven Piantadosi, *Clinical Trials*, 3/e (Wiley, 2017).

Sociological Methodology, Vol. 21, (1991), pp. 291-313

STATISTICAL MODELS AND SHOE LEATHER

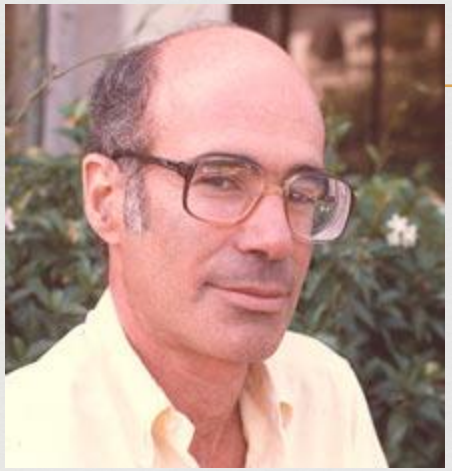
*David A. Freedman**

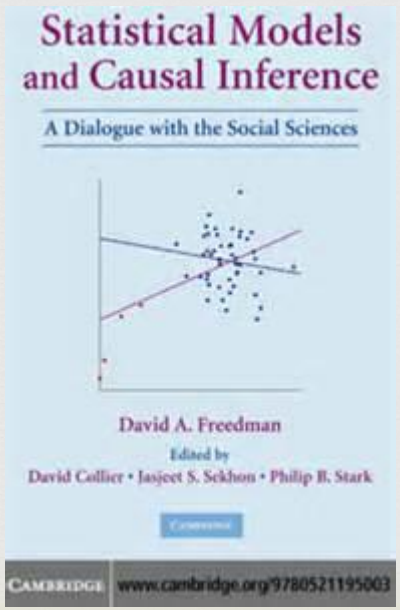


"…statistical technique can seldom be an adequate substitute for good design, relevant data, and testing predictions against reality in a variety of settings."

This requires hard work and an investment of resources ("shoe leather"). Multiple lines of evidence must be sought ("triangulation": Munafo & Davey Smith, *Nature* 2018).

David A. Freedman





"Generally, replication and prediction of new results provide a harsher and more useful validating regime than statistical testing of many models on one data set. Fewer assumptions are needed, there is less chance of artifact, more kinds of variation can be explored, and alternative explanations can be ruled out."

Statistical models and shoe leather (1991)

"I wish we could learn to look at the data more directly, without the fictional models and priors. On the same wish list: We should stop pretending to fix bad designs and inadequate measurements by modeling."

Issues in the foundations of statistics: probability and statistical models
(*Foundations of Science*, 1: 19-39, 1995)

Systematic error

- Can contaminate all the data from a single study.
  - Replication in the same lab: all auxiliary variables remain the same.
  - Replication in a different lab: all auxiliary variables may be changed.
- Youden, "Enduring Values" (*Technometrics*, 1972).
- Seife, "CERN's gamble shows perils, rewards of playing the odds" (*Science*, 2000)
- Bailey, "Why outliers are good for science" (*Significance*, 2018).

Statistics without statistical inference?

History of science, engineering, medicine

- Many great discoveries were made without the use of statistical inference.
  - Periodic table of the elements (Dmitri Mendeleev).
  - Kepler's laws of planetary motion.
  - Germ theory of disease (Ignaz Semmelweis, John Snow, Louis Pasteur, Robert Koch).
  - Molecular structure of DNA (James Watson, Francis Crick, Rosalind Franklin, etc.).
  - Energy quantization (Max Planck).
  - Plate tectonics (Alfred Wegener).
  - Powered flight (Wright brothers).

Statistics without inference

- Statistical principles for study design and execution.
- Enlightened data display and description.
- Disciplined data exploration.
  - Regularized, nonparametric, robust methods.
  - EDA.
- Statistical thinking.
- Example: Mogil & Macleod (*Nature*, 2017) framework.

When Mice Mislead

Tackling a long-standing disconnect between animal and human studies, some charge that animal researchers need stricter safeguards and better statistics to ensure their science is solid

By Jennifer Couzin-Frankel, *Science*, 342: 922-925 (2013)

- In mouse studies of stroke therapeutics, only 36% of studies reported randomization; 29% reported blinding.
- Studies that did **not** report either randomization or blinding "gave substantially and significantly higher estimates of how good these drugs were" – Malcolm Macleod.
- In one case the same drug had 2X effectiveness in a study *without* randomization as one that did!

Conclusions

- Statistical inference, regardless of ideology (frequentist, Bayesian, etc.) is not valid (or needed) except in extremely specialized circumstances.
  - Distracts us from focus on study design & execution; descriptive & exploratory statistics.
  - The Optimism Principle.
- Many Sets of Data and Shoe Leather, not Cult of Isolated Study.
- Statistical inference enables bad science. Statistical thinking enables good science.