Surviving the Reproducibility Crisis
A world Beyond p-values

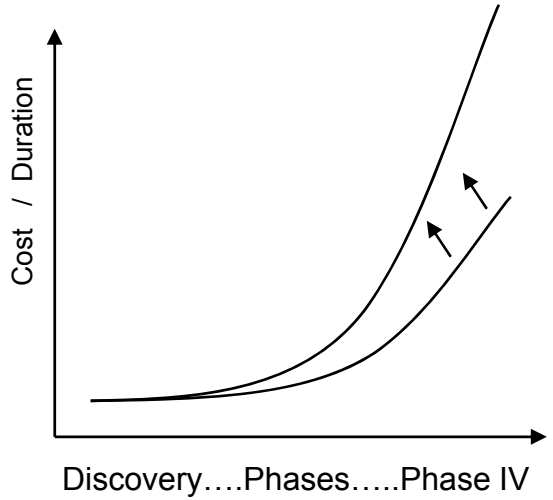
Bruno Boulanger
PharmaLex Statistical Solutions (Arlenda)

PHARMALEX

THE GAME

The background (part of…)

- The cost of studies increases along the R&D chain

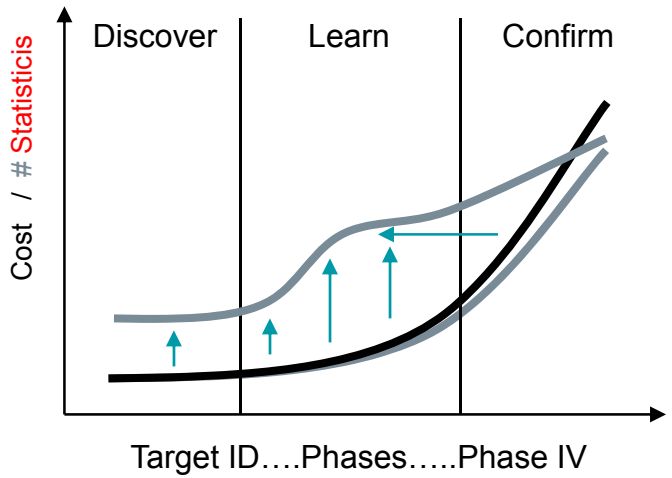


According to Tufts Center

$2.6Billion per approved drug

11.8% of drugs entering Clinical development are approved

Trends

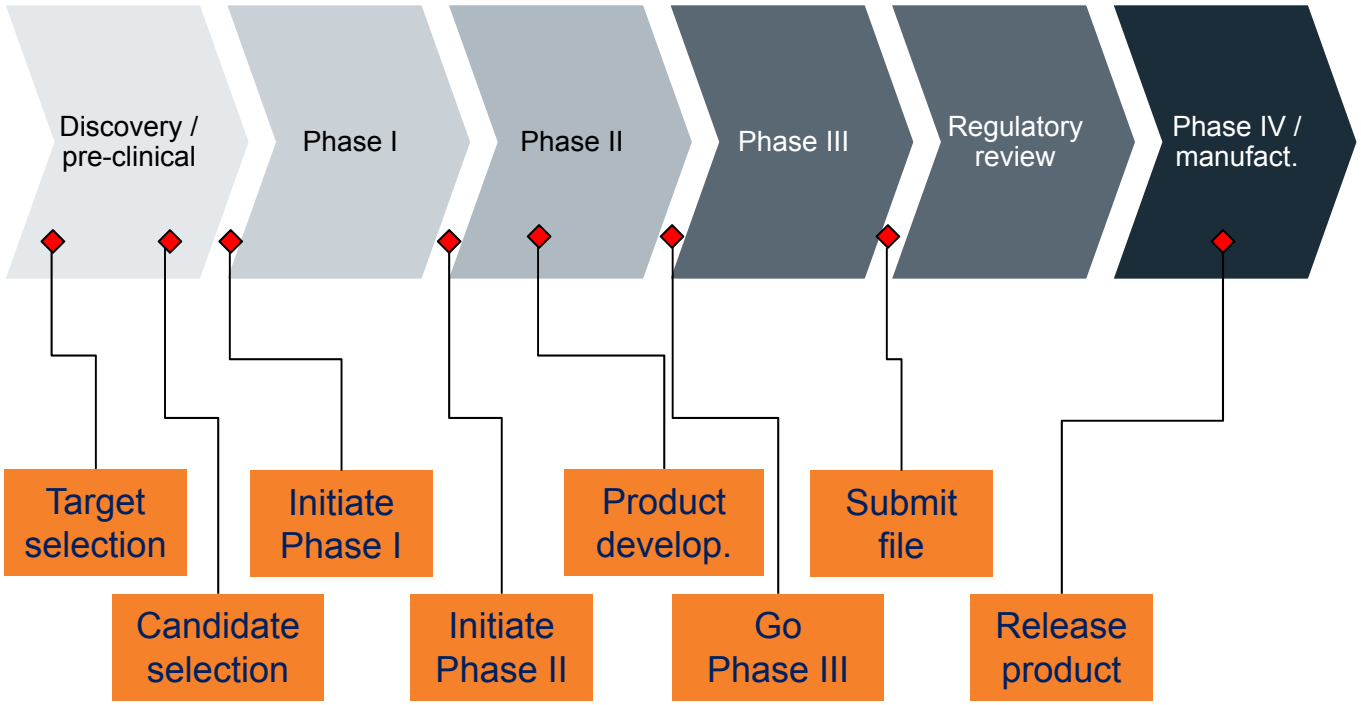


-Predictive models
-Virtual patients
-Biomath models
-Bayesian statistics
-Biomarker
-Translational medicine
-Implementing Technologies

⇒Improve and Predict p(ts)
⇒Increase return on investment

Prediction as development tools
Prediction as support to decisions

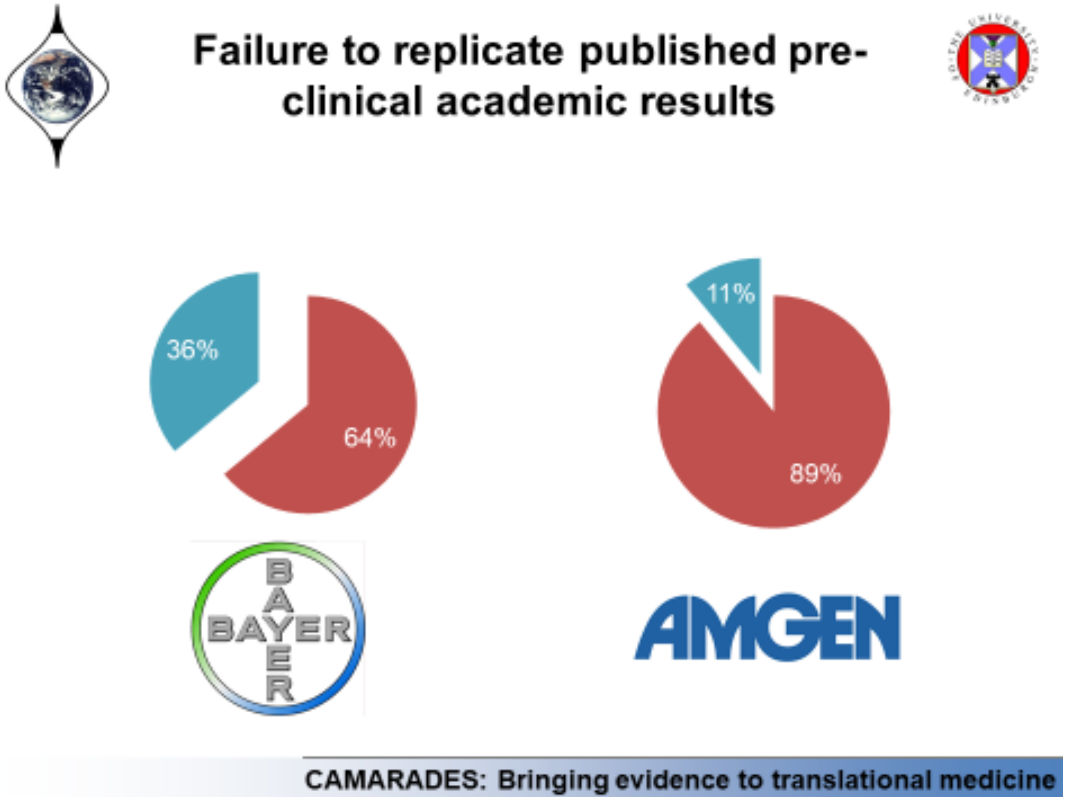
Decisions through drug development and sales



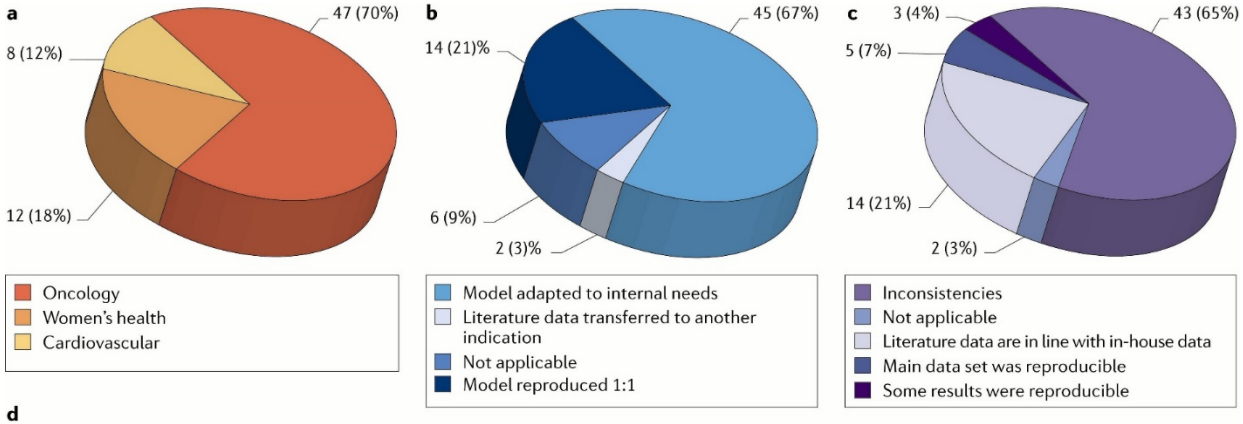
The P value and the

REPRODUCIBILITY CRISIS

The “Bayer” and “Amgen” publications



The “Bayer” publication (2011)



d

	Model reproduced 1:1	Model adapted to internal needs (cell line, assays)	Literature data transferred to another indication	Not applicable
In-house data in line with published results	1 (7%)	12 (86%)	0	1 (7%)
Inconsistencies that led to project termination	11 (26%)	26 (60%)	2 (5%)	4 (9%)

Nature Reviews | Drug Discovery

Current concerns about reproducibility

- "… it has become clear that biomedical science is plagued by findings that cannot be reproduced"
- "Science as a system should place more importance on reproducibility."

Nature's Solution

- From May 2013 Nature introduced editorial methods to improve the consistency and quality of reporting
  - More space given to method sections
  - Key methodological details will be reported
  - **Greater examination of statistics**
  - Encourage transparency, for example by including raw data

- Central to this is a new checklist prompting authors to disclose technical and statistical information

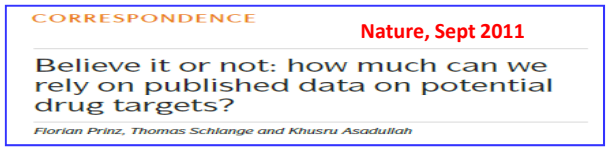
Growing Body of Evidence


OPEN ACCESS Freely available online PLOS ONE

High Impact = High Statistical Standards? Not Necessarily So

Patrizio E. Tressoldi[1]*, David Giofré[1], Francesco Sella[2], Geoff Cumming[3]

February 2013


CORRESPONDENCE

Nature, Sept 2011

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

Florian Prinz, Thomas Schlange and Khusru Asadullah


nature REVIEWS DRUG DISCOVERY

Journal home > Archive > Comment > Full Text

JOURNAL CONTENT

Journal home

Advance online publication

Current issue

Archive

Comment

October 2012

Nature Reviews Drug Discovery 11, 733-734 (Octobe

In search of preclinical robustness

Ian S. Peers[1], Peter R. Ceuppens[1] & Chris Harbron[1]


OPEN ACCESS Freely available online

June 2010

PLoS BIOLOGY

Perspective

Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research

Carol Kilkenny[1]*, William J. Browne[2], Innes C. Cuthill[3], Michael Emerson[4], Douglas G. Altman[5]


nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio &

Archive > Volume 490 > Issue 7419 > Perspectives > Article

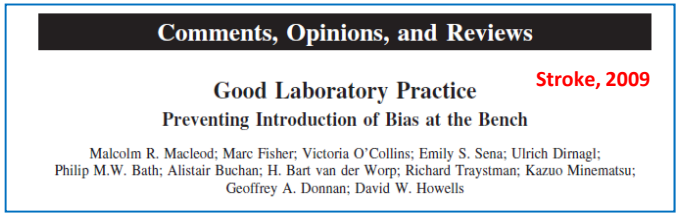
NATURE | PERSPECTIVES OPEN

日本語要約

August 2012

A call for transparent reporting to optimize the predictive value of preclinical research

Story C. Landis, Susan G. Amara, Khusru Asadullah, Chris P. Austin, Robi Blume


Comments, Opinions, and Reviews

Stroke, 2009

Good Laboratory Practice

Preventing Introduction of Bias at the Bench

Malcolm R. Macleod; Marc Fisher; Victoria O'Collins; Emily S. Sena; Ulrich Dirnagl; Philip M.W. Bath; Alistair Buchan; H. Bart van der Worp; Richard Traystman; Kazuo Minematsu; Geoffrey A. Donnan; David W. Howells


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.

Nature, March 2012


ELSEVIER

Pharmacology & Therapeutics

Volume 116, Issue 1, July 2007, Pages 148–175

July 2007

Clinical attrition due to biased preclinical assessments of potential efficacy

Mark D. Lindner

Over a Decade of Discussion

Trends in Pharmacological Sciences

Cell PRESS

Volume 24, Issue 7, July 2003, Pages 341–345

July 2003

Principles: The need for better experimental design

Michael F.W. Festing

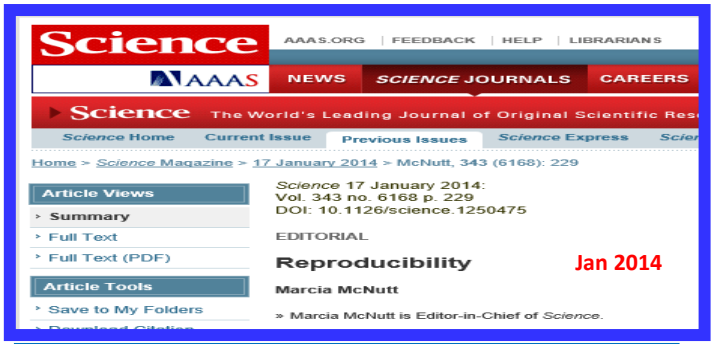
c/o FRAME (Fund for the Replacement of Animals in Medical Experiments), Russell and Burch House, 96–98 North Sherwood Street, Nottingham NG1 4EE, UK

Abstract

Many experiments could be improved with better experimental design and statistical analysis. Badly designed experiments can lead to incorrect conclusions and wasted time and scientific resources. Such experiments are unethical if they involve animals or humans. Good experimental design requires clearly defined objectives and control of the major sources of variation. In this article, a small mouse experiment involving the response of a liver enzyme to the administration of an antioxidant is used to illustrate some important design concepts such as the control and partitioning of sources of variation using factorial and randomized block designs and the estimation of appropriate sample sizes. Scientists clearly need better training in experimental design with better access to consultant statisticians for more complex situations.

“Many scientists ignore the basic principles of experimental design, analyse the resulting data badly, and in some cases reach the wrong conclusions.”

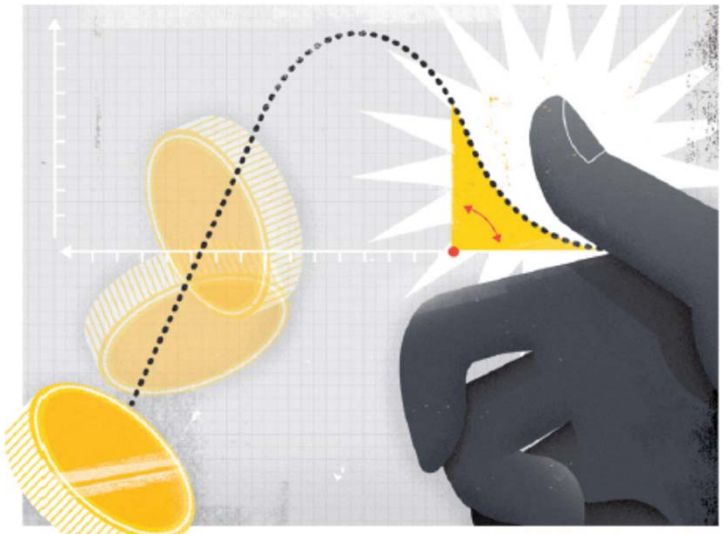
The Articles Keep Coming …



- "Sometimes the fundamentals get pushed aside – the basics of experimental design, the basics of statistics"

Lawrence Tabak, Principal Deputy Director of the NIH

Nature, 2014

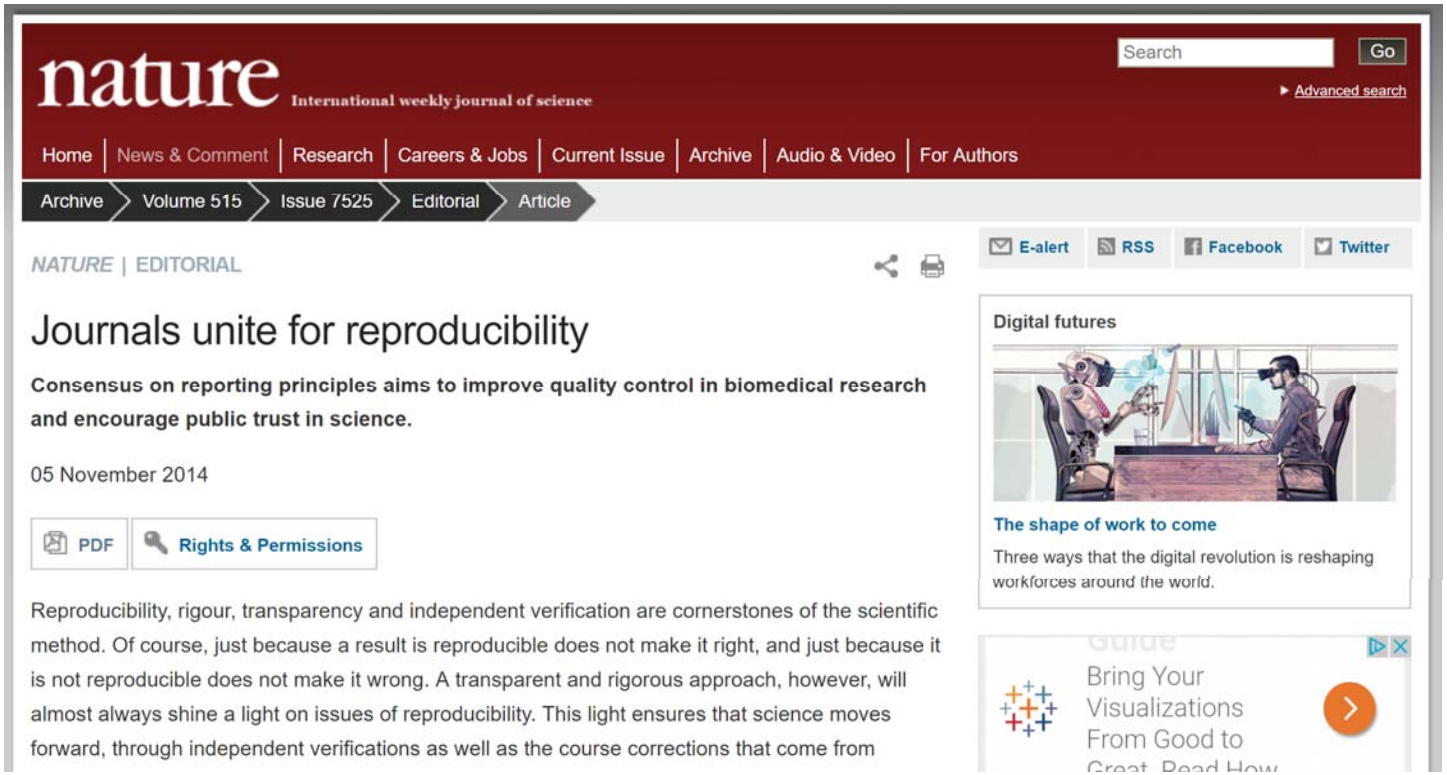


STATISTICAL ERRORS

P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

The editors endorse new rules



nature International weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Volume 515 > Issue 7525 > Editorial > Article

E-alert RSS Facebook Twitter

NATURE | EDITORIAL

Journals unite for reproducibility

Consensus on reporting principles aims to improve quality control in biomedical research and encourage public trust in science.

05 November 2014

PDF Rights & Permissions

Reproducibility, rigour, transparency and independent verification are cornerstones of the scientific method. Of course, just because a result is reproducible does not make it right, and just because it is not reproducible does not make it wrong. A transparent and rigorous approach, however, will almost always shine a light on issues of reproducibility. This light ensures that science moves forward, through independent verifications as well as the course corrections that come from

Digital futures

The shape of work to come

Three ways that the digital revolution is reshaping workforces around the world.

Bring Your Visualizations From Good to

EDITORIAL

Raising the bar

Marcia McNutt is Editor-in-Chief of Science.

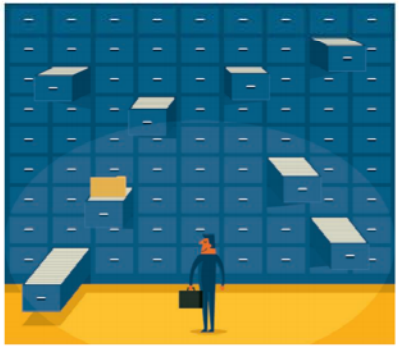
Numbers. Lots and lots of numbers. It is hard to find a paper published in *Science* or any other journal that is not full of numbers. Interpretation of those numbers provides the basis for the conclusions, as well as an assessment of the confidence in those conclusions. But unfortunately, there have been far too many cases where the quantitative analysis of those numbers has been flawed, causing doubt about the authors' interpretation and uncertainty about the result. Furthermore, it is not realistic to expect that a technical reviewer, chosen for her or his expertise in the topical subject matter or experimental protocol, will also be an expert in data analysis. For that reason, with much help from the American Statistical Association, *Science* has established, effective 1 July 2014, a Statistical Board of Reviewing Editors (SBoRE), consisting of experts in various aspects of statistics and data analysis, to provide better oversight of the interpretation of observational data.

For those familiar with the role of *Science*'s Board of Reviewing Editors (BoRE), the function of the SBoRE will be slightly different. Members of the BoRE perform a rapid quality check of manuscripts and recommend which should receive in-depth review by technical specialists. Members particularly when sophisticated approaches are needed. But even when taking added precautions, no review system is infallible, and no combination of reviewers can be expected to uncover all of the ways in which the interpretation of results may have gone wrong. In particular, it is difficult for reviewers to detect whether authors have approached the study with a lack of bias in their data collection and presentation.

I recall a study that I conducted years ago involving a global analysis of some oceanographic features that I was modeling to understand the rheology of oceanic plates on million-year time scales. I had only a handful of data points—perhaps a dozen or so—and the fit to my model failed a significance test. Clearly, throwing out a few of the data points by declaring them "outliers" would have improved the fit dramatically, and in fact I even recall a reviewer of the paper commenting: "Can't you make the data fit the model better?"

Really?

The editor published the paper despite the lousy fit of the model to the data. It was not too long before it was realized that those "outliers" were the key to a more complete understanding of the long-term rheological behavior of the oceanic plates. Although the model in the earlier paper needed an overhaul, the



"Readers must have confidence in the conclusions published in our journal."


U.S. Department of Health & Human Services
NIH National Institutes of Health
Turning Discovery Into Health
Search NIH
NIH Employee Intranet | Staff Directory | En Español
Health Information
Grants & Funding
News & Events
Research & Training
Institutes at NIH
About NIH
Home » Research & Training » Rigor and Reproducibility
RIGOR AND REPRODUCIBILITY
Rigor and Reproducibility
Principles and Guidelines
Expanded Guidelines
Application Instructions
Training
Funding Opportunities
Meetings and Workshops
Announcements
Publications
Principles and Guidelines for Reporting Preclinical Research
NIH held a joint workshop in June 2014 with the Nature Publishing Group and Science on the issue of reproducibility and rigor of research findings, with journal editors representing over 30 basic/preclinical science journals in which NIH-funded investigators have most often published. The workshop focused on identifying the common opportunities in the scientific publishing arena to enhance rigor and further support research that is reproducible, robust, and transparent.
Related Links
Nature Editorial: Journals Unite for Reproducibility
Science Editorial: Journals Unite for Reproducibility
NIH Office of Research on Women's Health
Further information regarding NIH expectations for the consideration of sex as a biological variable

2015
21st Century Cures Act US Congress Bill

A BILL

To accelerate the discovery, development, and delivery of 21st century cures, and for other purposes.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

This Act may be cited as the "21st Century Cures Act".

TITLE III—MODERNIZING CLINICAL TRIALS

Subtitle A—Clinical Research Modernization

Subtitle B—Broader Application of Bayesian Statistics and Adaptive Trial Designs

Raising concerns about use of p-values

… implying “that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.”

-Sir Harold Jeffreys (Astronomer, Geophysicist, Mathematician), 1939

“… surely the most bone-headedly misguided procedure ever institutionalized in the rote training of science students.”

-William Rozeboom (philosopher of science), **1960**

“ . . . dangerous nonsense (dressed up as the ‘scientific method’) and will cause much trouble before it is widely appreciated as such.”

-A.W.F. Edwards, FRS (statistician, geneticist, Fisher protege´), 1972

… misunderstood? “If you use p=0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time.”

-David Colquhoun, FRS (British pharmacologist), 2014

… banned from the journal *Basic and Applied Social Psychology*

“… prior to publication, authors will have to remove all vestiges of the NHSTP (p-values, t-values, F-values, statements about ‘significant’ differences or lack thereof, and so on).”

-David Trafimow and Michael Marks (journal editors), 2015

From Dave LeBlond, 2015

As the American Statistical Association officially reminded in March 2016….

REPRODUCIBILITY

Statisticians issue warning on *P* values

Statement aims to halt missteps in the quest for certainty.

BY MONYA BAKER

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warned on 8 March. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the *P* value was being misapplied, in ways that cast doubt on statistics generally, he adds.

cannot indicate the importance of a finding; for instance, a drug can have a statistically significant effect on patients' blood glucose levels without having a therapeutic effect.

Giovanni Parmigiani, a biostatistician at the Dana Farber Cancer Institute in Boston, Massachusetts, says that misunderstandings about what information a *P* value provides often crop up in textbooks and practice manuals. A course correction is long overdue, he adds. "Surely if this happened twenty years ago, biomedical research could be in a better place now."

FRUSTRATION ABOUNDS

Criticism of the *P* value is nothing new. In 2011, researchers trying to raise awareness about false positives gamed an analysis to reach a statistically significant finding: that listening to music by the Beatles makes undergraduates younger

Finally in June 2016, the ASA reminded in a press realse….

- P-values can indicate how incompatible the data are with a specified statistical model.
- P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency.
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

New York Times June 6th 2017

Le Monde, 2 octobre 2017

SCIENCES | Vidéos | Archéologie | Supplément partenaire : Les Prix EDF Pulse | Affaire de logique

ÉDITION ABONNÉS

Dans les labos, des petits arrangements avec la science

L'impératif de productivité scientifique augmente le risque de mauvaises pratiques. Ce sont le plus souvent les images et les statistiques qui sont manipulées par les chercheurs.

LE MONDE SCIENCE ET TECHNO | 02.10.2017 à 17h48 • Mis à jour le 03.10.2017 à 15h24 |

Par **David Larousserie**

Abonnez vous à partir de 1 € | Réagir | Ajouter | Partager (504) | Tweeter

ASA SYMPOSIUM ON
STATISTICAL
INFERENCE

OCTOBER 11-13, 2017 BETHESDA, MARYLAND

Scientific Method for the 21st Century: A World Beyond $p < 0.05$

A world beyond p-values

- "The most important task before us in developing statistical science is to demolish the P-value culture, which has taken root to a frightening extent in many areas of both pure and applied science and technology."
  Nelder, J. A. 1999. Statistics for the millennium. Statistician 48:257–269.

- “Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.”
  Ron Wasserstein, President American Statistical Association, March 2016

- “… we recommend abandoning the null hypothesis significance testing paradigm entirely, leaving p-values as just one of many pieces of information with no privileged role in scientific publication and decision making.”
  McShane, Gal, Gelman, Robert & Tackett, 21SEP2017

The European Quality In Preclinical Data (EQIPD) project

Members of the EQIPD consortium have been pivotal in producing substantial evidence which suggests that the **robustness, rigor and validity of preclinical research is limited** and that this provides a barrier to the effective and efficient development of new drugs.
We believe there is a need for simple, sustainable solutions that facilitate data quality without impacting innovation and freedom of research.


EQIPD
THE UNIVERSITY OF EDINBURGH
janssen
concentris
research management
Imperial College London
PASP
Arlenda
UNIVERSITY OF ABERDEEN
u b
b UNIVERSITÄT BERN
JG|U
JOHANNES GUTENBERG UNIVERSITÄT MAINZ
EBERHARD KARLS UNIVERSITÄT TÜBINGEN
CHARITÉ
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
ECNP
european college of neuropsychopharmacology
Sylics
Science EXCHANGE
Porsolt
UMC
Noldus
Information Technology
abbvie
Boehringer Ingelheim
NOVARTIS PHARMACEUTICALS
ORION
Pfizer

HOW TO SURVIVE THE REPRODUCIBILITY CRISIS ?

Proposal

- Embrace the **Lifecycle** vision in research
- Be inspired by the **Bayesian** Statistics in decision making
- Always apply **Design of Experiments**
  - Stop thinking analysis of data, think **modeling**
  - It goes beyond blinding, randomization….
  - Think about robustness and generalysability
- Evaluate the "**capability**" of the assay to achive objectives
- Continuously **Control** and assess performance and improve
- Be **transparent**

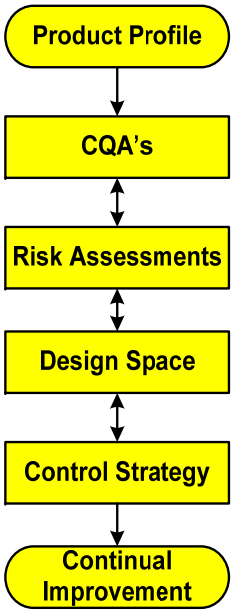
Quality by Design and

LIFECYCLE APPROACH

Quality by Design approach

1. Define objectives and criteria of success
2. Identify biological/animal model and relevant quantifiable quality attributes (end-points) linked to the objectives
3. Develop jointly a modeling strategy (or analyses) of the data to generate
    1. Relevant with the objectives
    2. Aligned with the MBDD strategy
4. Optimize design of assays and studies to probability of success and cost/time effectivness
5. Validate, control and improve the capability of the assays/studies

Q8(R2) - Example QbD Approach

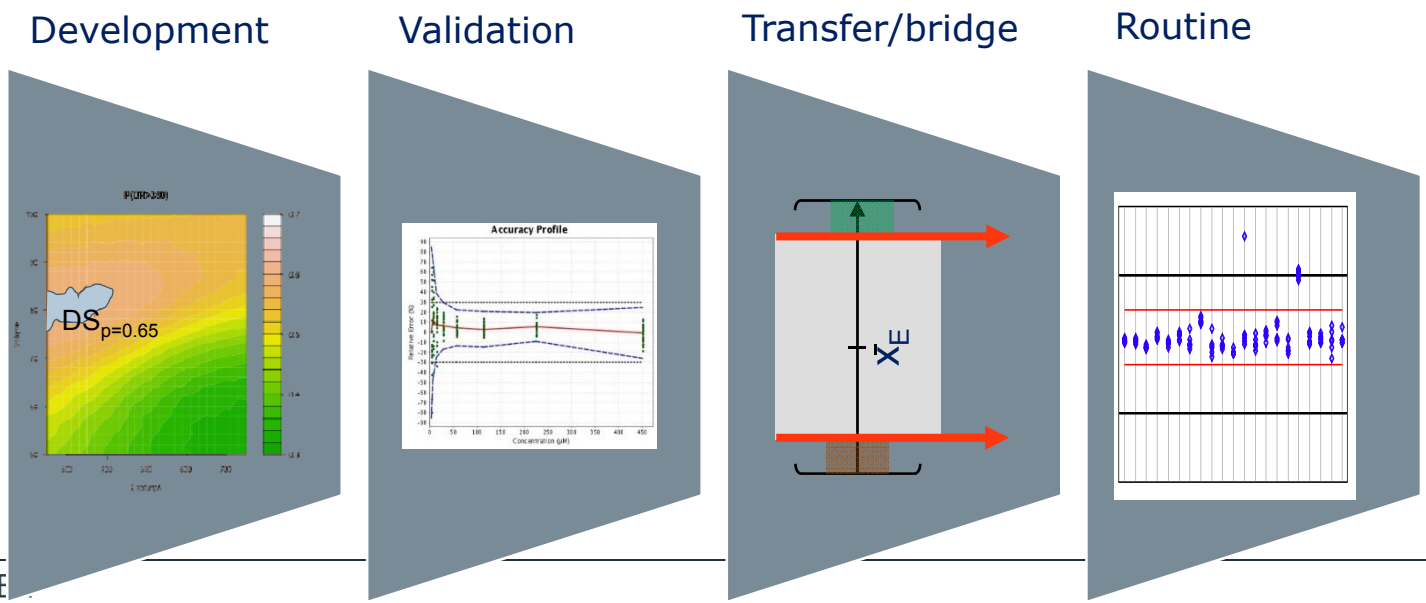


- Analytical Target Product Profile (aTPP)
- Determine "potential" critical quality attributes (CQAs)
- Identify link of assay parameters to CQAs and perform risk assessment
- Develop a Design of Experiment and a **design space**
- Design and implement a control strategy
- Manage assay lifecycle, including continual improvement

Lifecycle vision

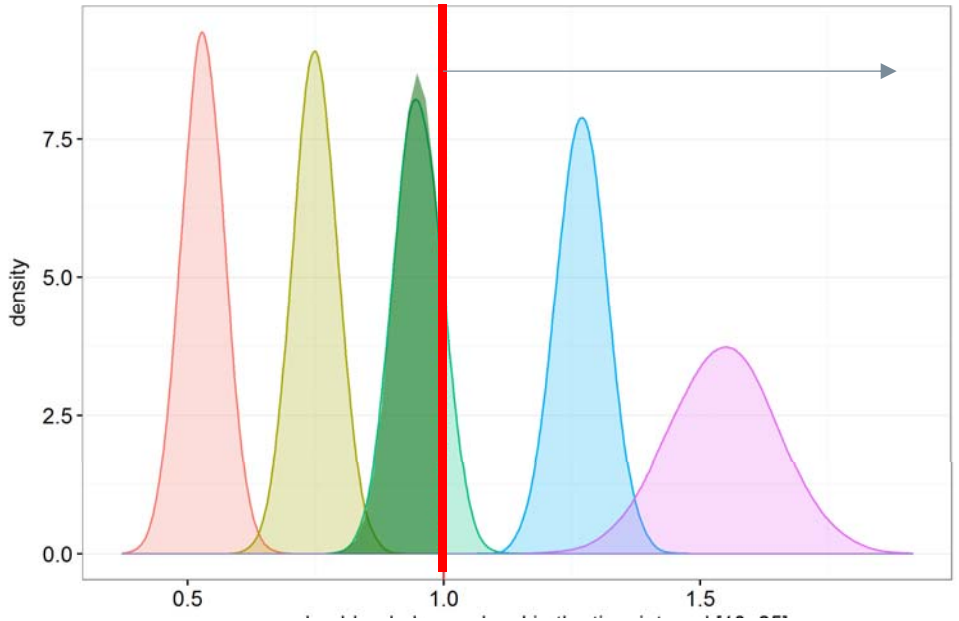
How to develop, validate, transfer and maintain a procedure to ensure it will continuously produce results that are fit-for-use?

How to keep the risk low and maintained along the value chain?



Bayesian statistics and learning process

- Evaluate if you reached your objectives
  - Given the data
  - Given the knowledge
- ➔ Bayesian statistics



HOW TO MAKE A DECISION

The objective: is my treatment effective ?

How to make a decision ?

A

What is the probability of obtaining the observed data, if the treatment is not effective?

B

What is the probability that the treatment is effective, given the observed data?

Two different ways to make a decision based on

A

Pr(**observed data** | **treatment is not effective**)

- Better known as the **p-value** concept
- Used in the **null hypothesis** test (or decision)
- This is the likelihood of the data assuming an hypothetical explanation (eg the “null hypothesis”)
- **Classical statistics** perspective (Frequentist)

B

Pr(**treatment effective** | **observed data**)

- **Bayesian** perspective
- It is the probability of efficacy given the data

The Bayesian perspective allows to directly address the question of interest.

The diagnostic test example

Cancer ? → diagnostic test → result

A problem of decision making

The accuracy of a diagnostic test is assessed as follows:

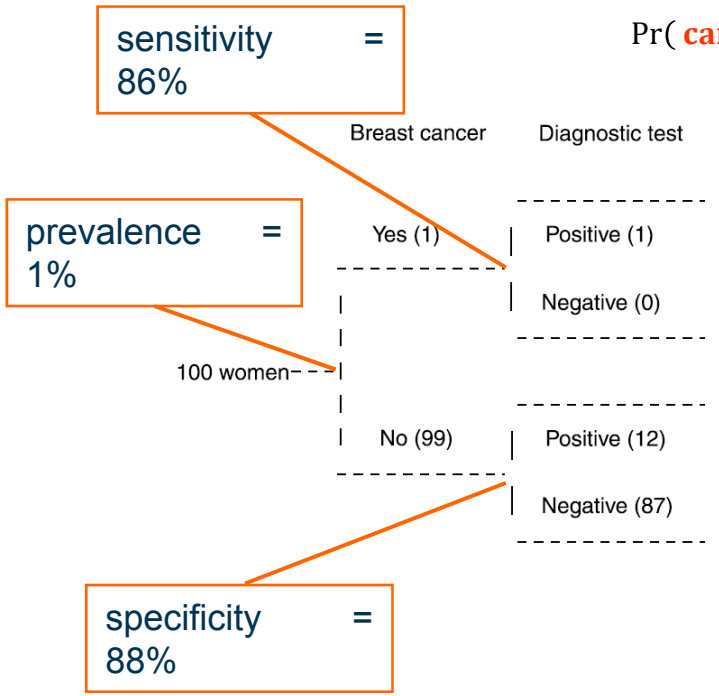
- Sensitivity: $\Pr(\textbf{positive result} \mid \textbf{cancer})$
- Specificity: $\Pr(\textbf{negative result} \mid \textbf{no cancer})$

In practice:

Given that the diagnostic test result is positive, what is the probability you truly have cancer?

$$\Pr(\textbf{cancer} \mid \textbf{positive result}) = ?$$

Example



$$\Pr(\textbf{cancer} \mid \textbf{positive result}) = \frac{1}{12+1} = 0.077$$

How can that be so low?
The small proportion of errors for the large majority of women who do not have breast cancer swamps the large proportion of correct diagnoses for the few women who have it.

The probability of interest dependents on the underlying prevalence of the disease.

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* Wiley, 2nd ed.
Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of *p*-values. *R. Soc. Open sci.* **1**(3): 140216

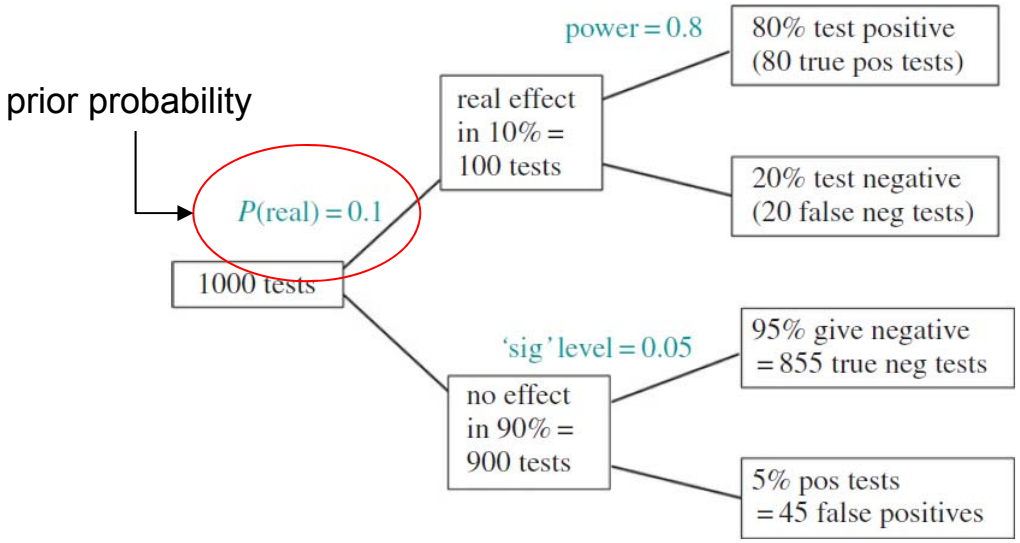
The animal study analogy

effective? → Animal study → data

$$\Pr(\textbf{drug effective} \mid \textbf{data}) = ?$$

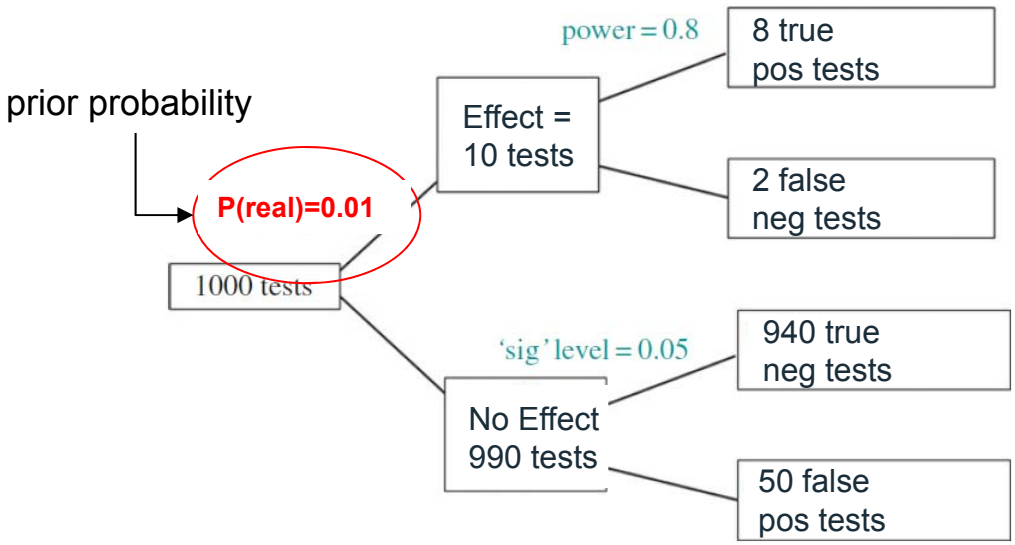
depends largely on **prior probability** that there is a real effect

"If you use p = 0.05 to suggest that you have made a discovery, you will be wrong at least 30% of the time."



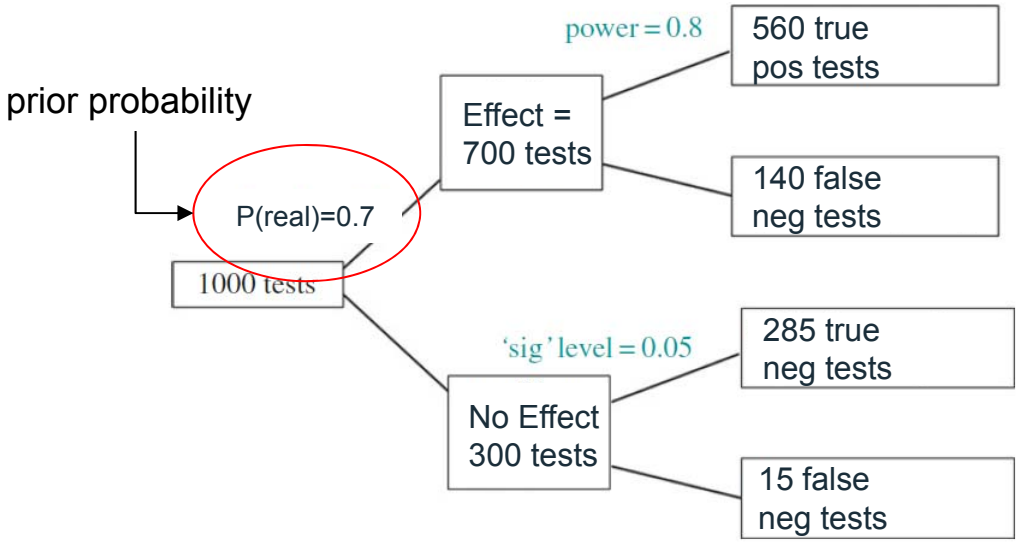
$$\Pr(\textbf{real effect} \mid \mathbf{p} < 0.05) = \frac{80}{80 + 45} = 0.64$$

"If you use p = 0.05 "….when you are in early discovery



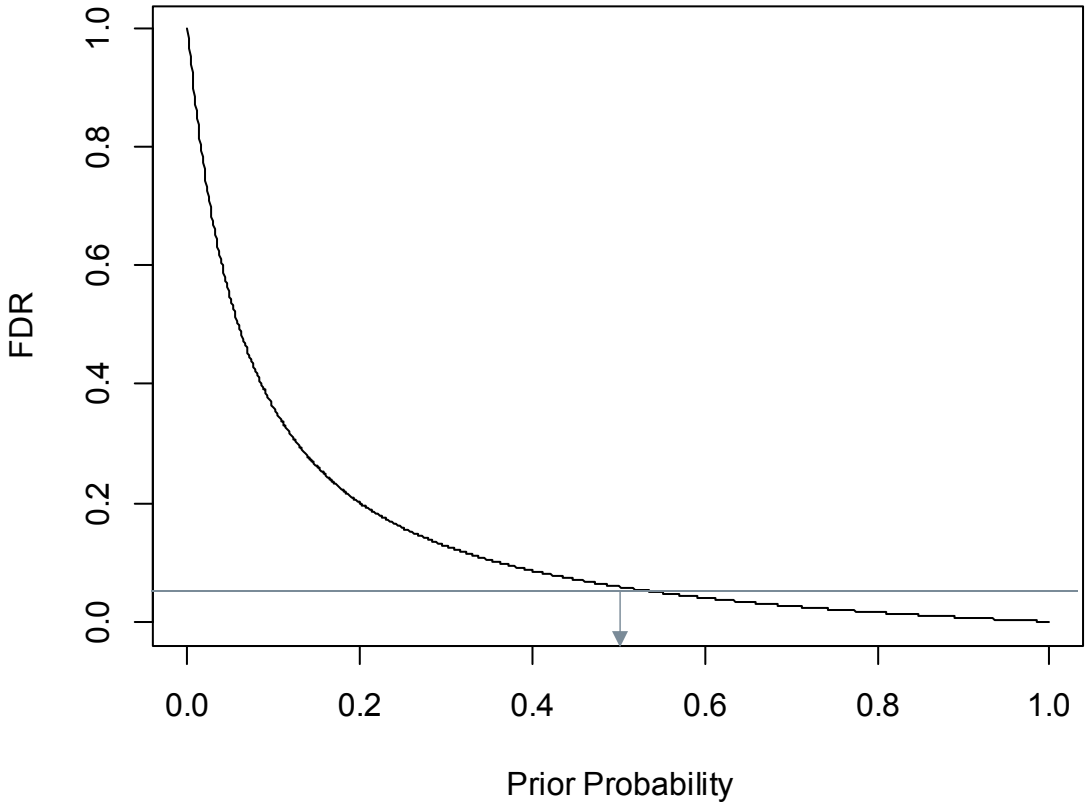
$$\Pr(\textbf{real effect} \mid \mathbf{p} < 0.05) = \frac{8}{8+50} = \mathbf{0.14}\ !!!!$$

“If you use p = 0.05 “…. if you have a good prior as before starting a Phase III

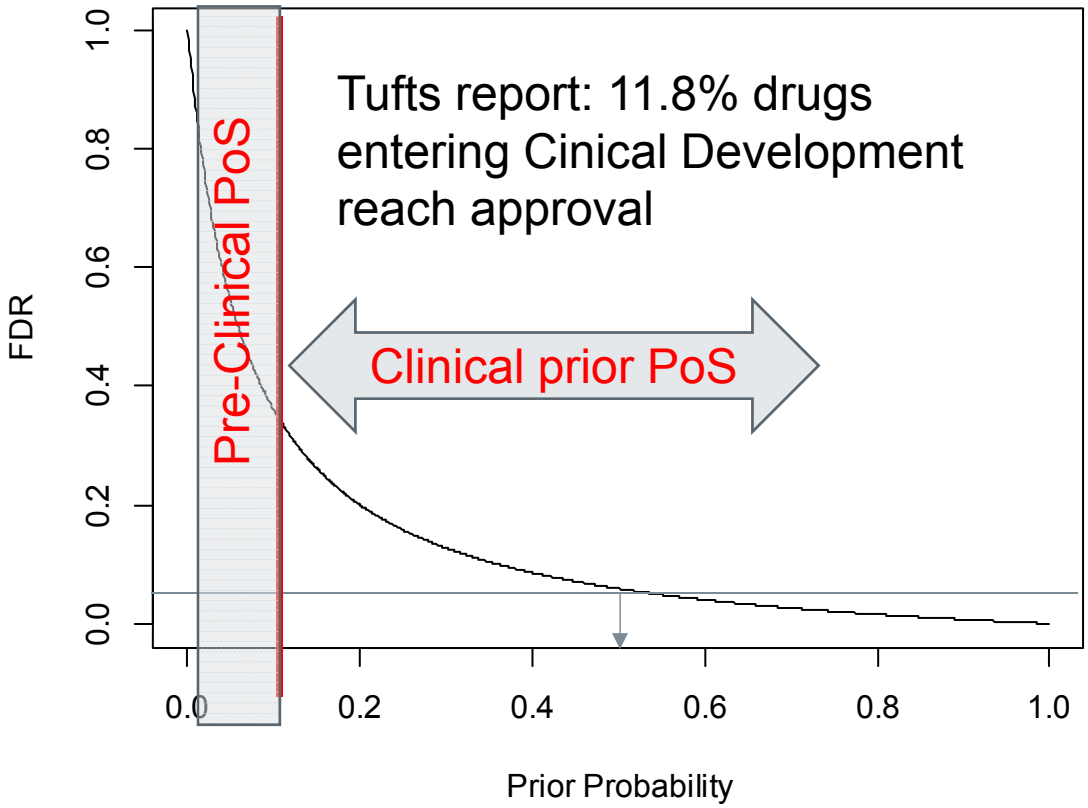


$$\Pr(\textbf{real effect} \mid \mathbf{p} < 0.05) = \frac{560}{560 + 15} = \mathbf{0.97}$$

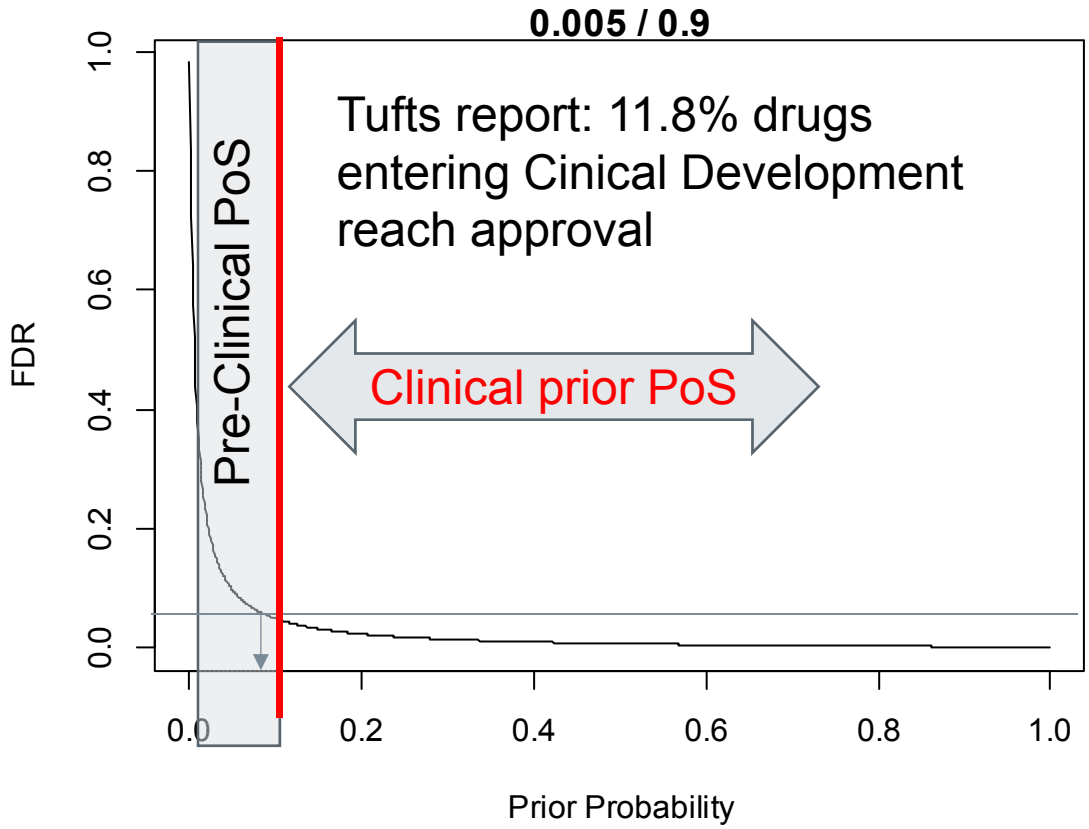
False Discovery Rate for p<0.05, power=0.8 as function of Prior Probability



False Discovery Rate for p<0.05, power=0.8 as function of Prior Probability



False Discovery Rate for p<0.005, power=0.9 as function of Prior Probability



P<0.005 is just a Quick Fix

- 70 Statisticians publish a paper in July 2017:

  "We propose to change the default P-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005.“

- But few weeks later, Gelman et al said:

  “… we recommend abandoning the null hypothesis significance testing paradigm entirely, leaving p-values as just one of many pieces of information with no privileged role in scientific publication and decision making.”
  McShane, Gal, Gelman, Robert & Tackett, 21SEP2017

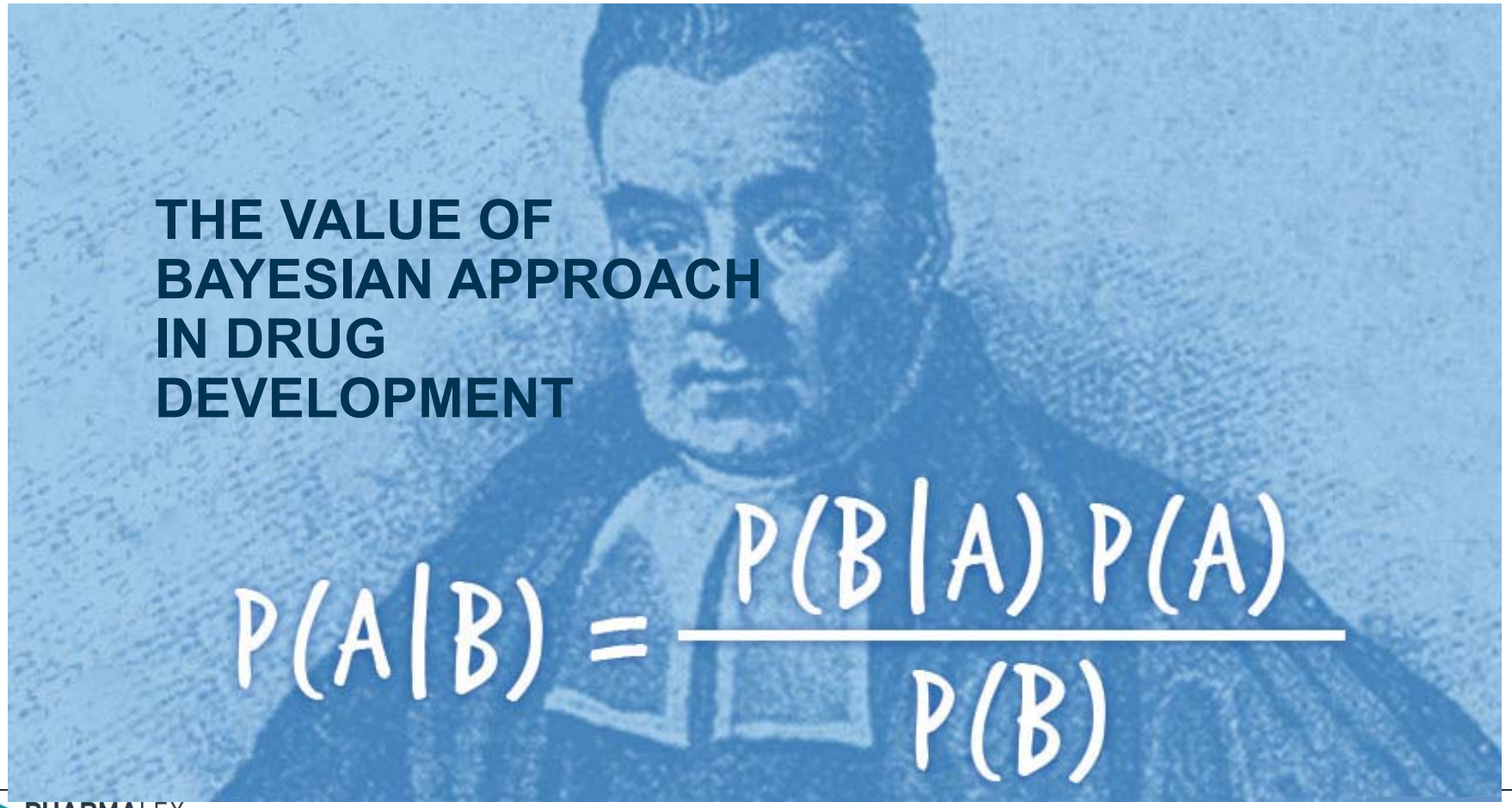
Some critics about using p-values and Frequentist approach

- Efficacy is not a hypothesis; it is a matter of degree
- Would you rather know the chance of making an assertion of efficacy when the drug has no effect, or the chance the drug is effective?
- Need a formal way to insert extra-study information
  - skepticism
  - trustworthy evidence / past data
- Frequentist paradigm requires a certain design rigidity
- Frequentist approach conservative when want to learn continuously

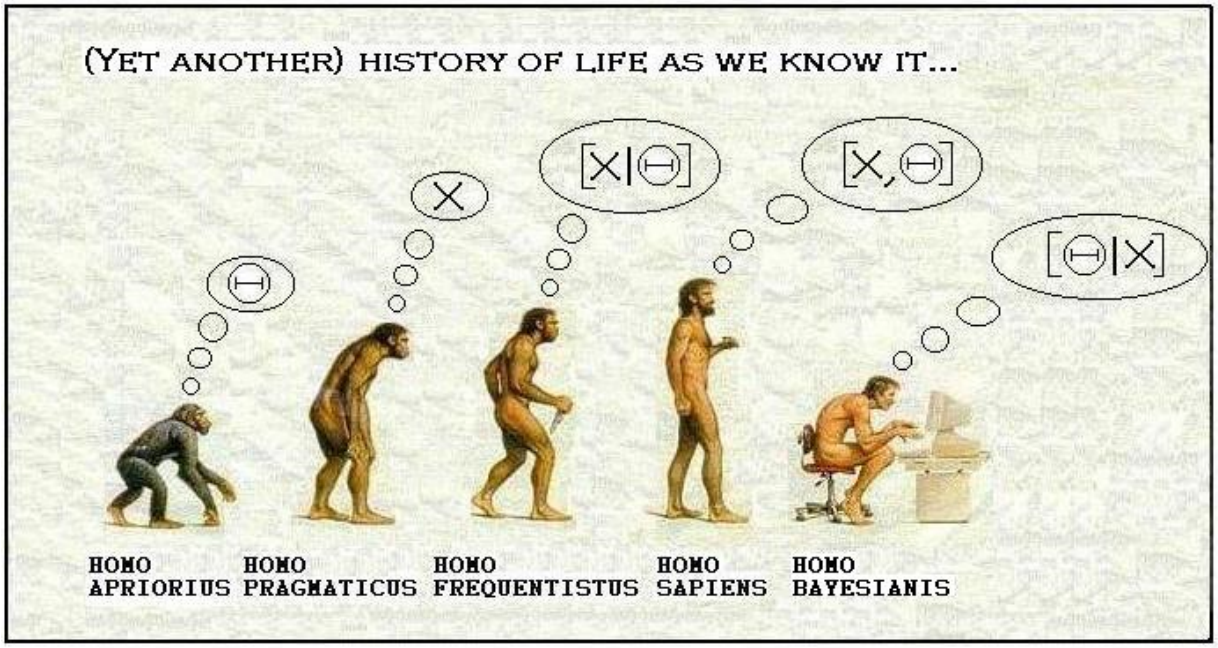
From Frank Harrel, 2017

- Frequentist mainly provide study specific conclusions (no learning)

THE VALUE OF BAYESIAN APPROACH IN DRUG DEVELOPMENT



The raise of Bayesian statistics



Bayes' Rule: $$[\Theta \mid X] = \frac{[X, \Theta]}{[X]} = \frac{[X \mid \Theta] \cdot [\Theta]}{[X]}$$

Agenda

- Bayesian principles
- Posterior computation
- Predictive Distribution
- Comparison Bayesian-Frequentist
- Prior elicitation

Bayesian principle

- Example: clinical trial to collect evidence of an unkown treatment effect
  - **Frequentist analysis:**
    - point estimate and confidence intervals as summaries of size of treatment effect
    - Asks: what this trial tells us about the treatment effect
  - **Bayesian analysis:**
    - Before the trial: a priori opinion on the treatment effect
    - Asks: how should this trial change our opinion about the treatment effect?
- Motivations for adopting Bayesian approach:
  - Natural and coherent way of thinking about science and learning

Bayesian principle

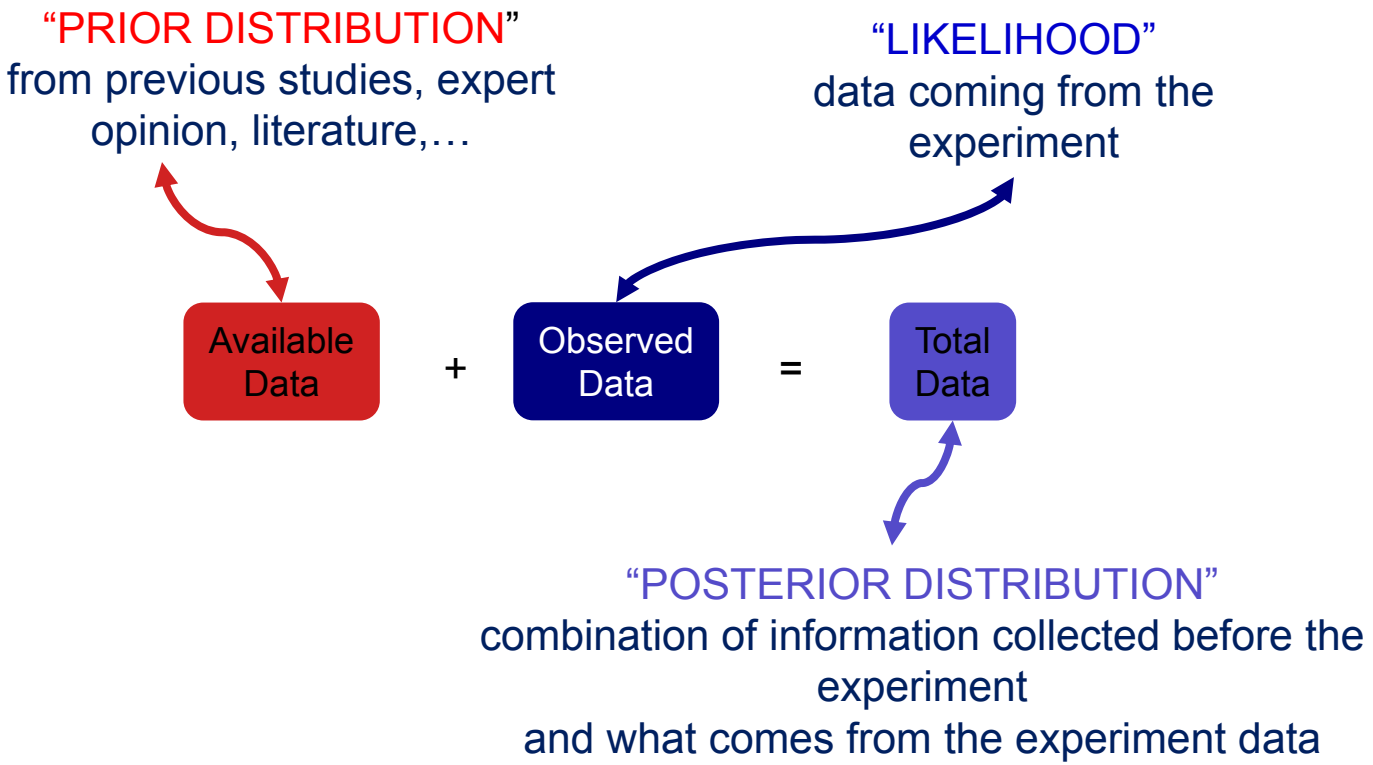
- After having observed the data of the study, the prior distribution of the treatment effect is updated to obtain the posterior distribution

PRIOR distribution *STUDY data* *POSTERIOR distribution*

- Instead of having a point estimate (+/- standard deviation), we have a complete distribution for any parameter of interest

P(treatment effect > 5.5)= P(success)

Bayesian principle



Bayesian principle

- Let's consider that θ is the parameter of interest (ex: treatment effect)

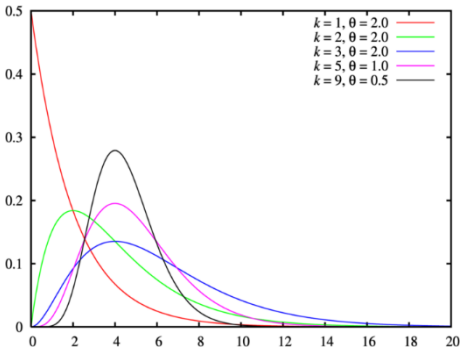
θ is treated as random variables

1. ***Prior distribution*** of parameter θ : $p(\theta)$
   - Distribution of θ before any data are observed
   - Reasonable opinion concerning the plausibility of different values of θ
   - Ideally based on all available evidence/knowledge (or belief)
   - Or deliberately select a non-informative prior

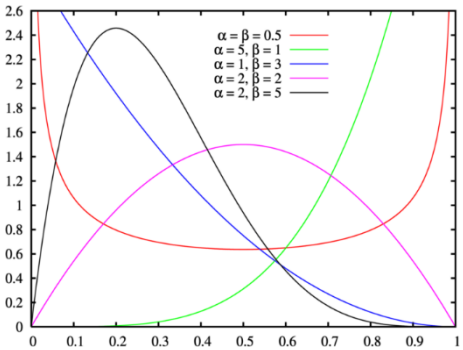
Bayesian principle

Examples of prior distributions

Gamma distributions



Beta distributions



- Prior **distribution** -> Specify the domain of plausible values

  -> Specify the weights given to these values
- Prior distributions do not have to be a Normal (not only prior mean and prior variance)
- Prior distributions ≠ initial values.

Bayesian principle

2. ***Likelihood***:
   - Conditional probability of the data given θ: p(y| θ)
   - Based solely on data

3. ***Posterior distribution***:
   - Distribution of θ after observed data have been taken into account: p(θ|y)
   - Final opinion about θ

4. ***Predictive distribution***:
   - Given the model and the posterior distribution of its parameters, what are the plausible values for a future observation y*?
     p(y*| θ)

Bayesian principle

- The posterior distribution of θ is obtained by the Bayes' rule:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)d\theta}$$

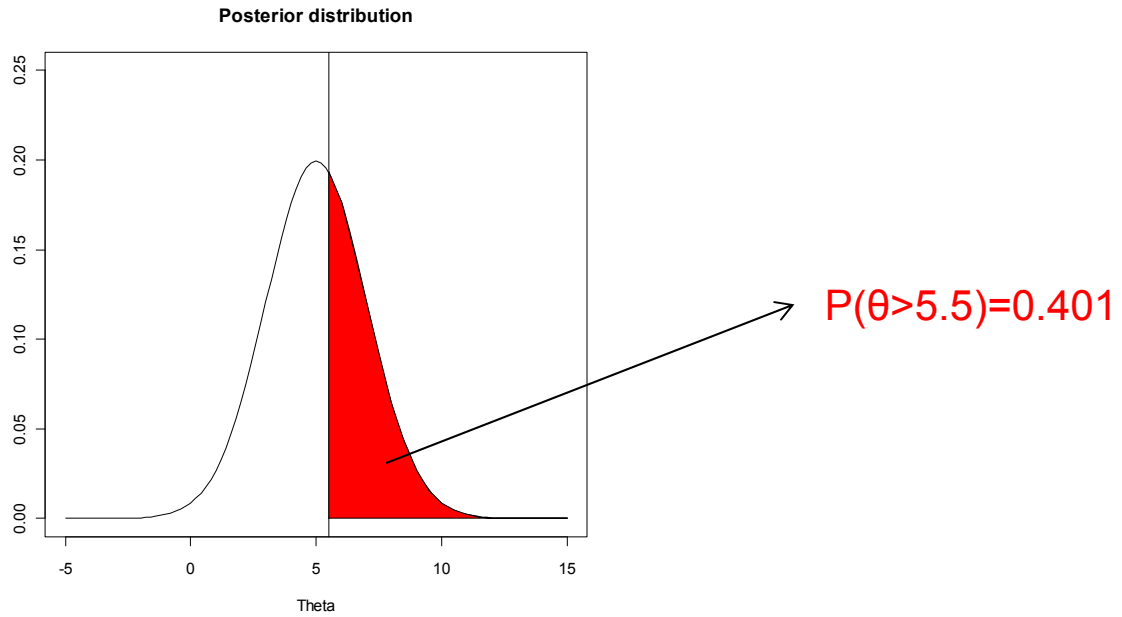
$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

- Results reflect the combined evidence of data and prior knowledge or belief

- The posterior distribution is used for inference regarding parameter

Bayesian principle

- Uncertainty is described in terms of probability :



Posterior computation

- The posterior distribution contains everything that can be said about θ.
- To summarize its information content:
  - Measures of location:posterior mode, posterior median, posterior mean
  - Measures of spread: Posterior variance
  - Bayesian credibility interval:
    - Get the quantiles of the distribution (2.5% and 97.5%)
    - An interval that contains 95% of the posterior probability for θ, i.e. 95% most plausible/credible values
  - Any probability on the values of θ or on a function of θ

Bayesian Predictive Distribution

The Bayesian theory provides a definition of the
Predictive Distribution of a new observation given past data.

$$p(\tilde{x}|data) = \int\int_{\sigma^2\,\mu} p(\tilde{x}|\mu,\sigma^2,data) \times p(\mu,\sigma^2|data)\; d\mu d\sigma^2$$

Integrate over parameter distribution

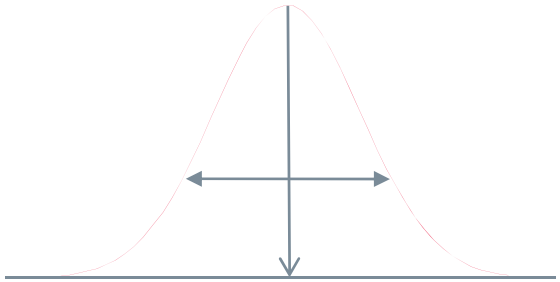
Model

Joint posterior

Difference Simulations/Predictions

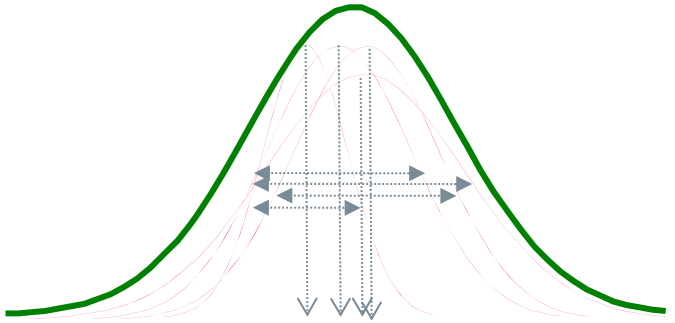
Simulations

the “new observations” are drawn from distribution “centered” on estimated location and dispersion parameters (treated as “true values”).



Predictions

the uncertainty of parameter estimates (location and dispersion) is taken into account before drawing “new observations” from relevant distribution



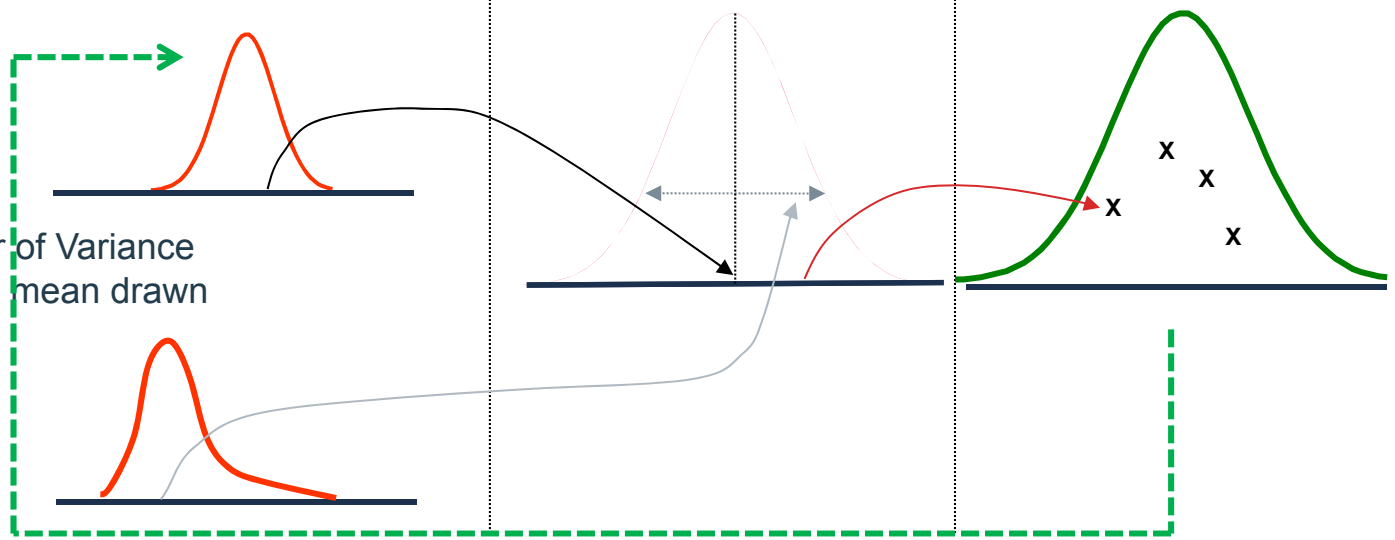
Practically, how to make predictions

1st , draw a mean and a variance from:

- Posterior of mean μ_i
- Posterior of Variance σ^2_i given mean drawn

2nd , draw an observation from the resulting distribution Y~ Normal(μ_i, σ^2_i)

3rd , repeat this operation a large number of time to obtain the predictive distribution



Comparison Bayesian-Frequentist

1. **Random vs fixed:**
   - Bayesian: probability of parameters given observed data
   - Frequentist: probability of observed data given parameters

2. **Evidence used (in the analysis):**
   - Bayesian: all available (relevant) information/knowledge
   - Frequentist: specific to experiment

Comparison Bayesian-Frequentist

3. Inference

- Bayesian : examine the probability of θ given the data.
- Frequentist : tests of significance are performed by supposing that a hypothesis is true (the null hypothesis) and then computing the probability of observing a statistic at least as extreme as the one actually observed during hypothetical future repeated trials. (This is the *P-value*).

(p-value=probability to observe something more disadvantageous for H0 than what we have observed, if H0 is true)

Comparison Bayesian-Frequentist

4. Intervals

- Bayesian : *credible interval* : 95% most plausible/credible values

- Frequentist : *Confidence interval*: “If we repeat the same experiment a large number of times, the confidence interval will contain the true value in 95% of the cases.”

Comparison Bayesian-Frequentist

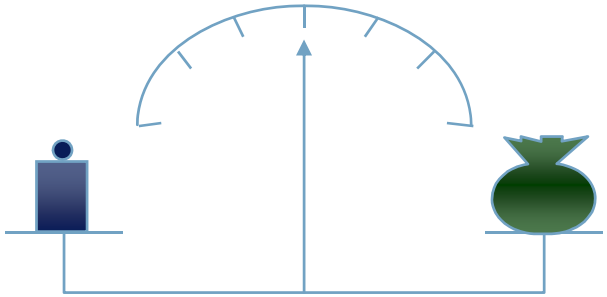
5. Design flexibility

- Bayesian : May adapt trial design as evidence accumulates
  - Sample size does NOT need to be pre-specified
  - Interim analysis may be conducted anytime and at any frequency

- Frequentist: Interim analyses possible but restricted
  - Must be pre-specified
  - Number and timing affect the analyses

THE VALUE OF DESIGN OF EXPERIMENTS

The Weighing Problem



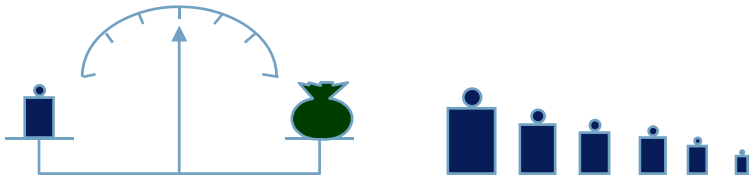
The problem and the hardware

- The problem

  Find the weights of 3 objects A, B, and C, in 4 weighings and with the best precision

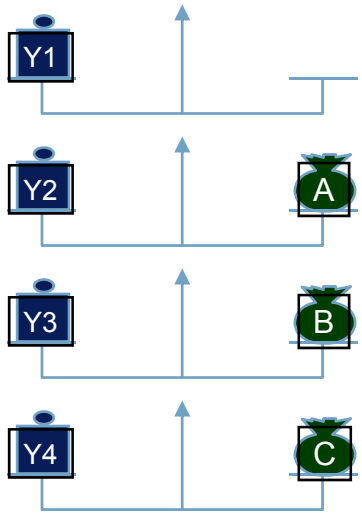
- The hardware

  A pair of scales to be equilibrated with weights



Strategy 1

- Weigh one object at a time



Design matrix

	A	B	C
Exp 1	0	0	0
Exp 2	1	0	0
Exp 3	0	1	0
Exp 4	0	0	1

0 : the object is not on the scales
1 : the object is on the right scale
-1 : the object is on the left scale

What's the precision on the estimations ?

- The variance of the error on each weighing Y1, Y2, Y3, and Y4, is $V(\varepsilon) = \sigma^2$

Weight estimators

$M_0 = Y1$
$M_A = Y2 - Y1$
$M_B = Y3 - Y1$
$M_C = Y4 - Y1$

Variance properties

$V(X+Y) = V(X) + V(Y) + 2^* Cov(X,Y)$
$V(X-Y) = V(X) + V(Y) - 2^* Cov(X,Y)$
$V(aX+b) = a^2 V(X)$

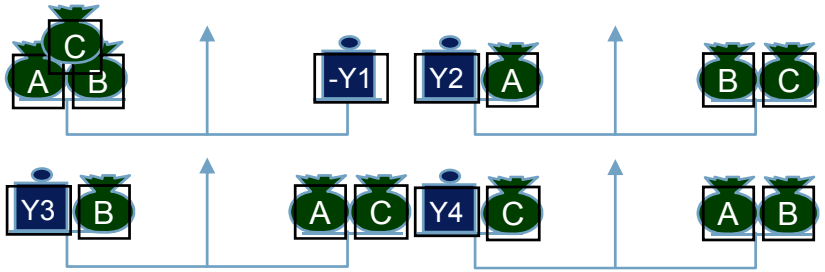
Variances of the estimators

$V(M_0) = \sigma^2$

$V(M_A) = V(M_B) = V(M_C) = \mathbf{2\ \sigma^2}$

How to get more precise estimations ?

Strategy 4



Design matrix

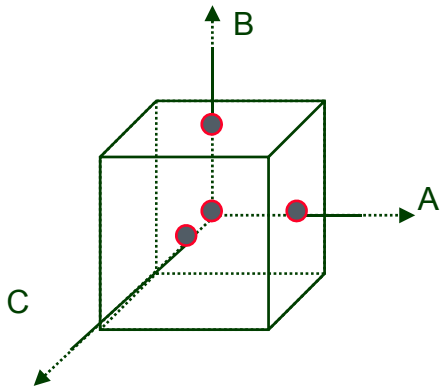
	A	B	C
Exp 1	-1	-1	-1
Exp 2	-1	1	1
Exp 3	1	-1	1
Exp 4	1	1	-1

$V(MA) = V(MB) = V(MC) = (\sigma^2 + \sigma^2 + \sigma^2 + \sigma^2) / 16 = \mathbf{\sigma^2 / 4}$ Cost = 4,000 $

Why has precision been improve by a factor 8 ?

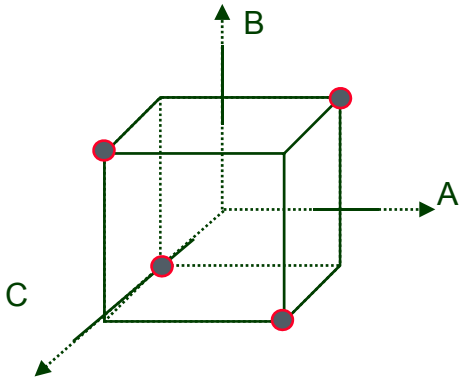
Strategy 1

	A	B	C
Exp 1	0	0	0
Exp 2	1	0	0
Exp 3	0	1	0
Exp 4	0	0	1



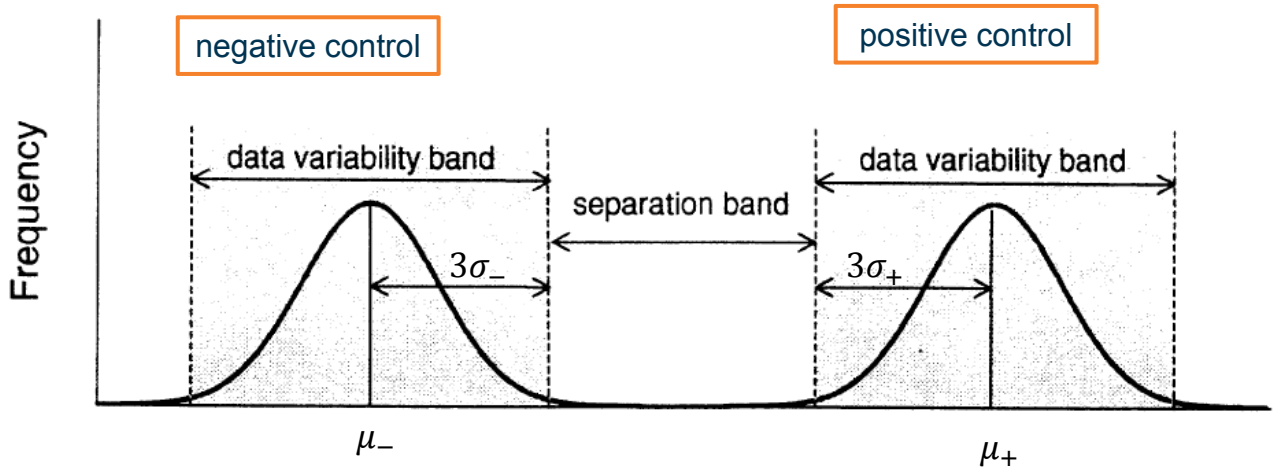
Strategy 4

	A	B	C
Exp 1	-1	-1	-1
Exp 2	-1	1	1
Exp 3	1	-1	1
Exp 4	1	1	-1



FORMAT OF ASSAYS

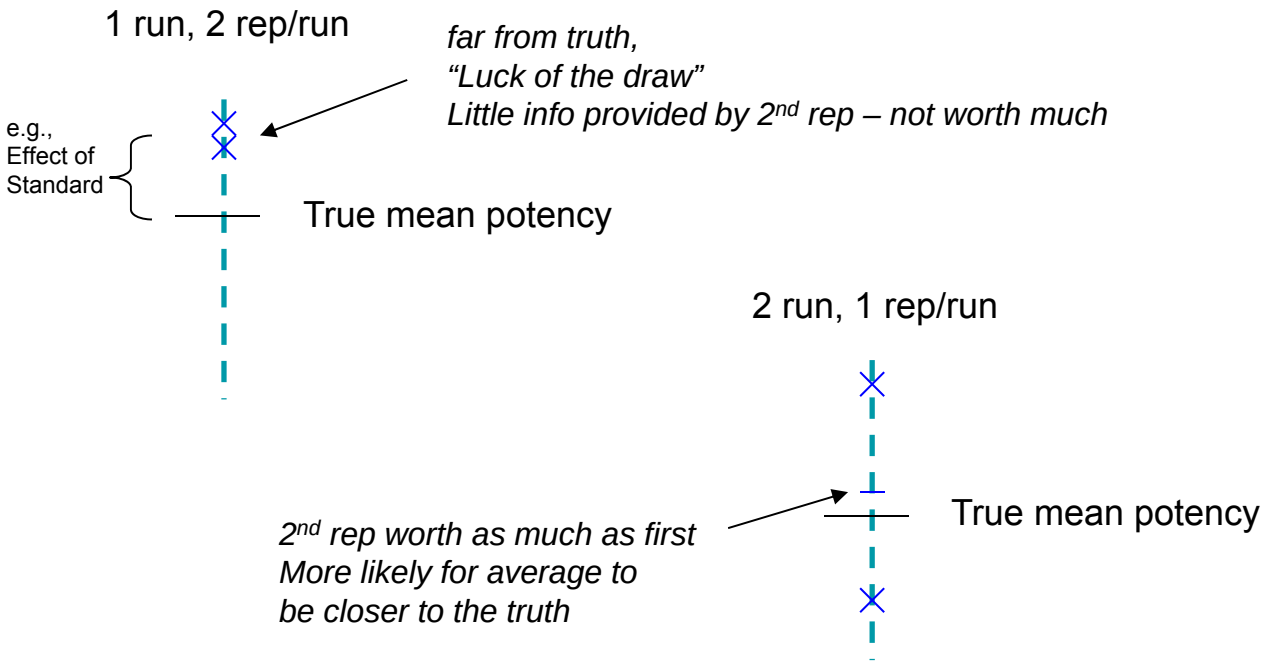
Z-factor (1/2)



- **separation band:**
  $(\mu_+ + 3\sigma_+) - (\mu_- - 3\sigma_-) - (6\sigma_+ + 6\sigma_-) = (\mu_+ - \mu_-) - (3\sigma_+ + 3\sigma_-)$
- **dynamic range:** $\mu_+ - \mu_-$

Zhang J.-H., Chung T. D. Y. & Oldenburg K. R. (1999). A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *Journal of Biomolecular Screening* **4**, 67–73.

Example of Increasing mean precision



$\sqrt{\sigma^2_{run}/1+\sigma^2_{rep}/2}$ ⟶ Std. Err of Estimate ⟶ $\sqrt{\sigma^2_{run}/2+\sigma^2_{rep}/2}$

Common error encountered: standardization vs generalizability

- Standardization: Usually scientists try to obtain the results in condition with smallest variance

➔ They introduce biais in results

- Generalizability: The experimental units should be spread across the conditions with the greatest variance

➔ This eliminate the biais linked to conditions

➔ The Precision can be improved by the sample size

➔ The conclusions should be “whatever day, strain, ….”

- This is key to Reproducibility

$$SE = \sqrt{\frac{s^2_{Exp}}{r} + \frac{s^2_{Fami}}{r \cdot p} + \frac{s^2_{Anim}}{r \cdot p \cdot n}}$$

Example 1

OBJECTIVE AND END-POINT

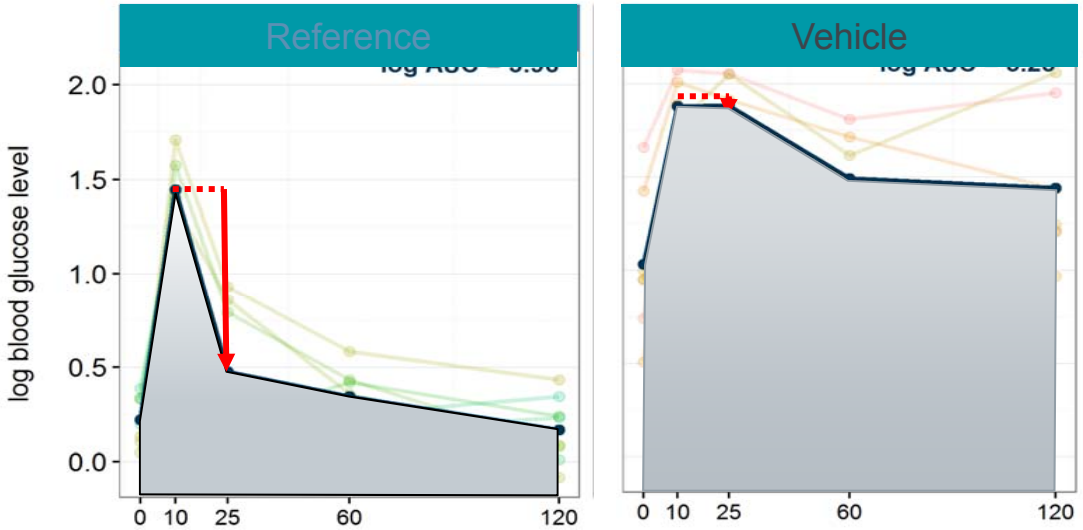
An example (1) - objectives

Define and share objectives and criteria of success

- Metabolic desease:
- Find a compound that **rapidly** control the level of glycemia in diabetes for **several** hours .
- The new compound should have high chances not to be inferior to the Reference product
- Margin of non-inferiority is 80% of reference

An example (2) – end-points

- Identify animal model and end-points linked to objective
  - Lower the 2hour blood glucose level in OGTT
  - Decrease rapidly glucose from 10’ to 25’



An example (3) - Modeling

Modelling
Translate the objective
Allow precise estimate of criterions
Based on literature knowledge
Use logitudinal data (or individual trajectories)

Model the data to ….

```
proc mixed data=data;
   class study mouse_id treatment time;
   model y = time treatment treatment*time;
   random study / group=treatment;
   repeated / subject=mouse_id(study)
              type=ar(1)
              group=treatment;
run;
```

…estimate global effect…..

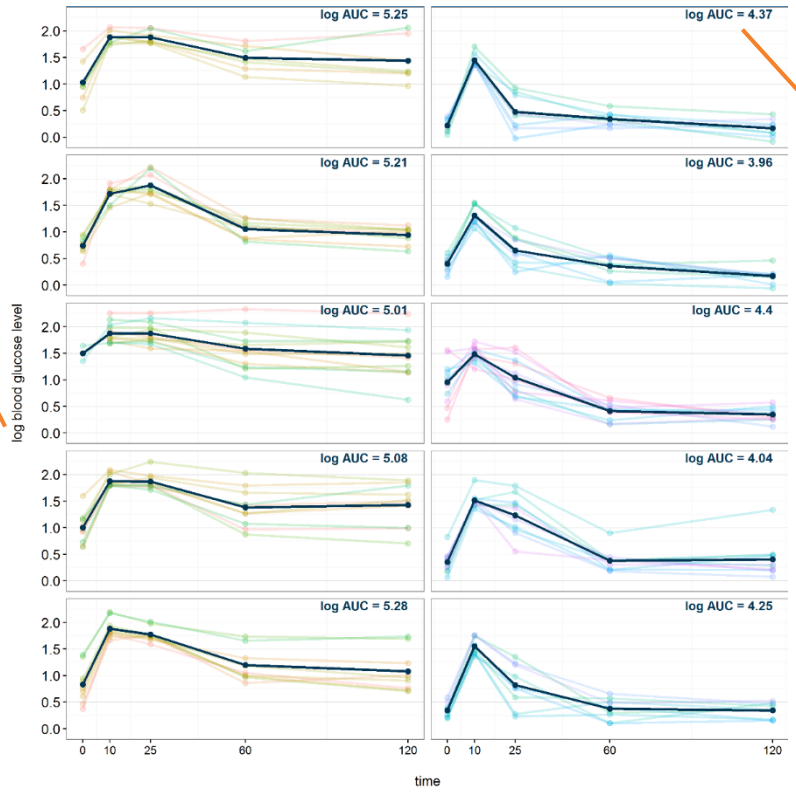
```
lsmeans treatment;
```

…or speed of onset…..

```
        estimate "product A dose 2 - t2-t3"
            intercept        2
    time              0  1  1  0  0
    treatment         0  0  0  2  0
    treatment*time    0  0  0  0  0
                         0  0  0  0  0
                      0  0  0  0  0
                         0  1  1  0  0
                      0  0  0  0  0 / divisor = 2;
```

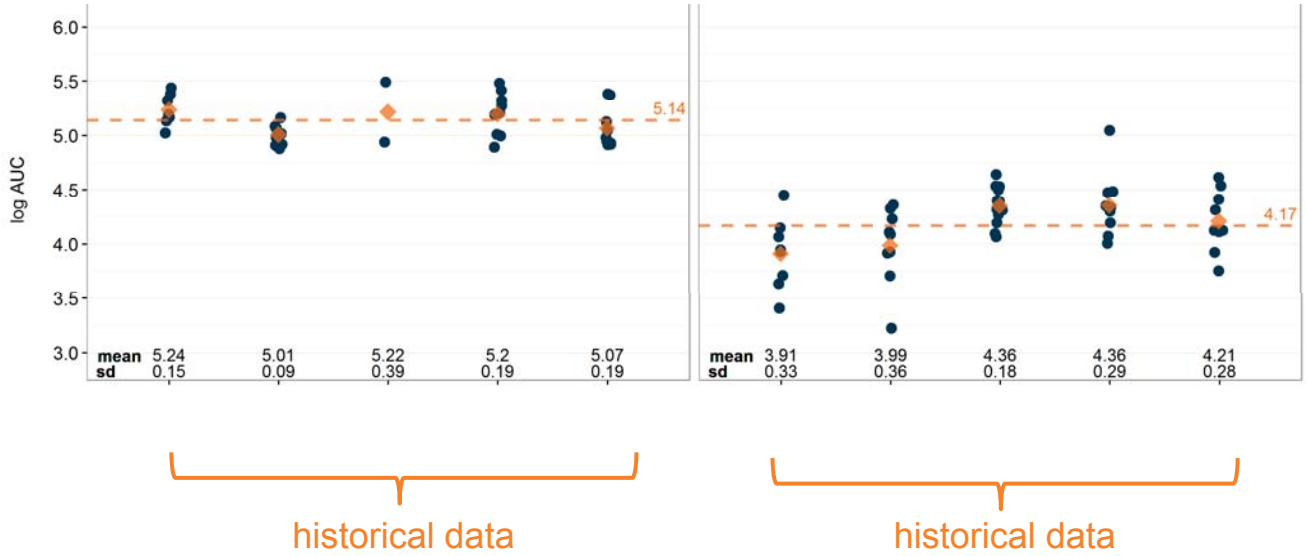
Individual and average profiles of log glucose levels

log transformation



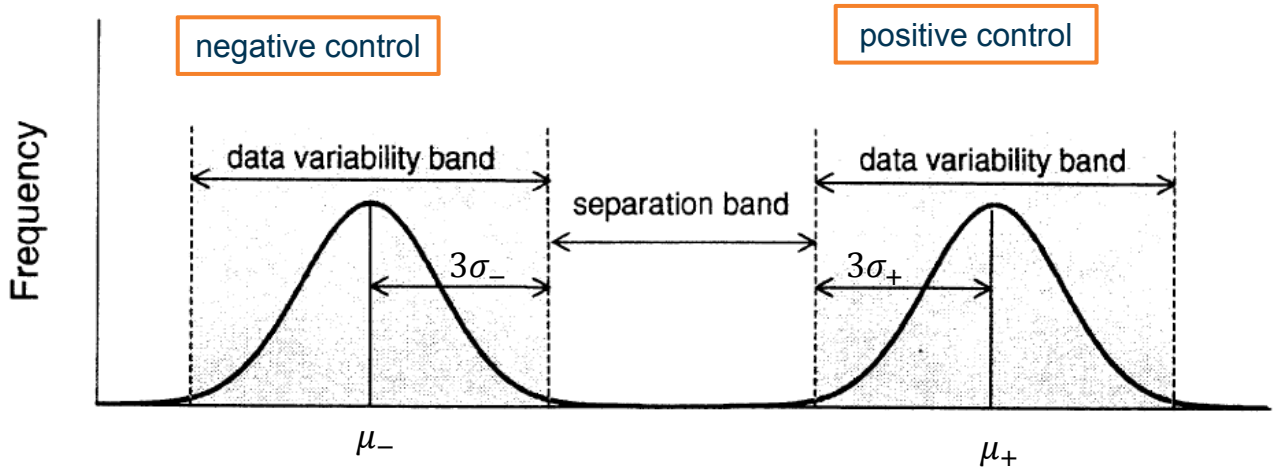
log area under average profile

Log areas under individual profiles



Within each treatment group, the diamonds indicate the positions of the means within a study while the dashed line indicates the position of the overall mean across all studies.

Z-factor (1/2)



- **separation band:**
  $(\mu_+ + 3\sigma_+) - (\mu_- - 3\sigma_-) - (6\sigma_+ + 6\sigma_-) = (\mu_+ - \mu_-) - (3\sigma_+ + 3\sigma_-)$
- **dynamic range:** $\mu_+ - \mu_-$

Zhang J.-H., Chung T. D. Y. & Oldenburg K. R. (1999). A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *Journal of Biomolecular Screening* **4**, 67–73.

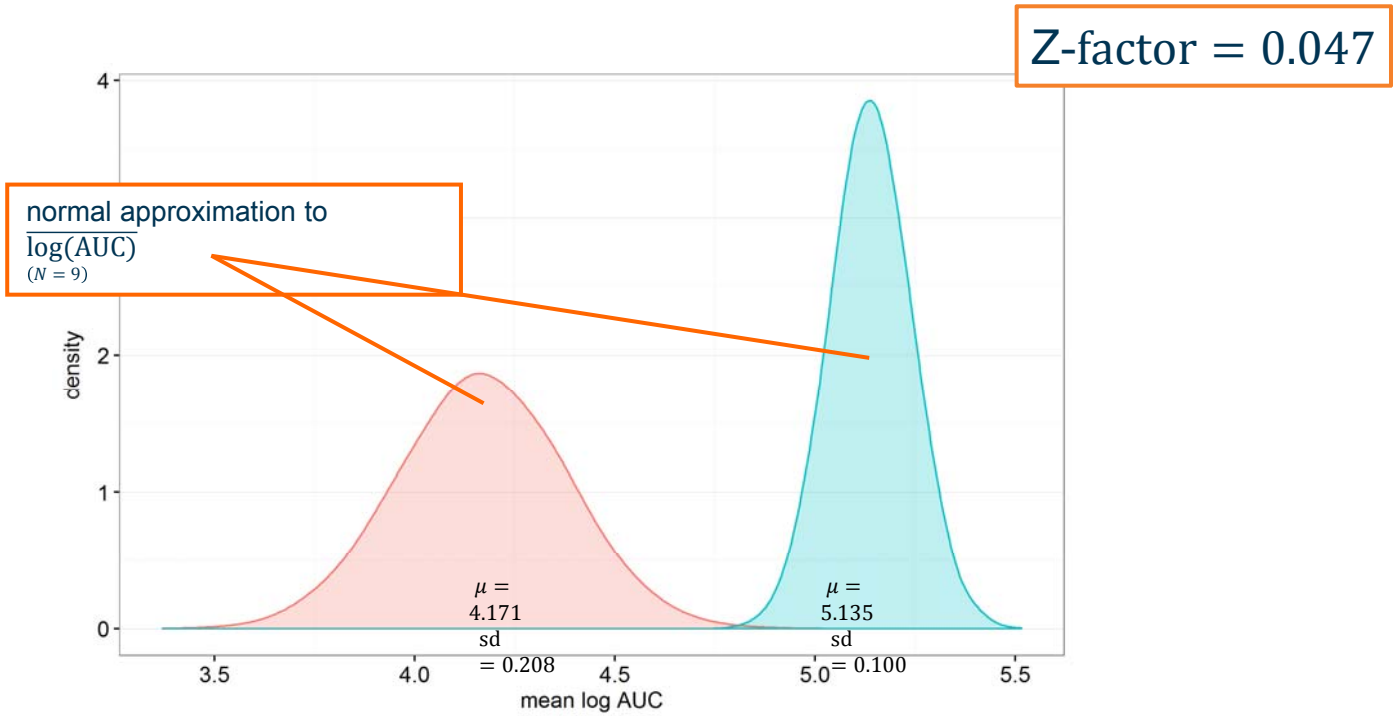
Z-factor (2/2)

$$\text{Z-factor} = \frac{\text{separation band}}{\text{signal dynamic range}} = 1 - \frac{3\sigma_+ + 3\sigma_-}{\mu_+ - \mu_-}$$

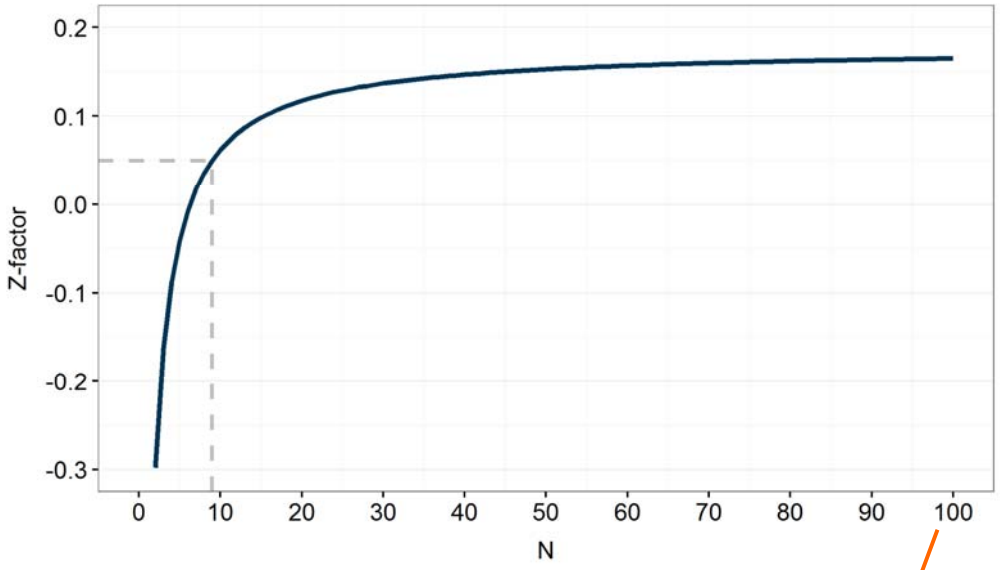
- $\text{Z-factor} = 1$: ideal assay; as $(3\sigma_+ + 3\sigma_-)$ approaches zero, i.e. very small standard deviations, or as $\mu_+ - \mu_-$ approaches infinity.
- Z-factor between 0.5 and 1: excellent assay; separation band is large.
- Z-factor between 0 and 0.5: separation band is small.
- $\text{Z-factor} < 0$: no separation band; there is too much overlap between the positive and negative controls for the assay to be useful.

Zhang J.-H., Chung T. D. Y. & Oldenburg K. R. (1999). A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *Journal of Biomolecular Screening* **4**, 67–73.

Overall z-factor (1/3)



Overall z-factor (2/3)



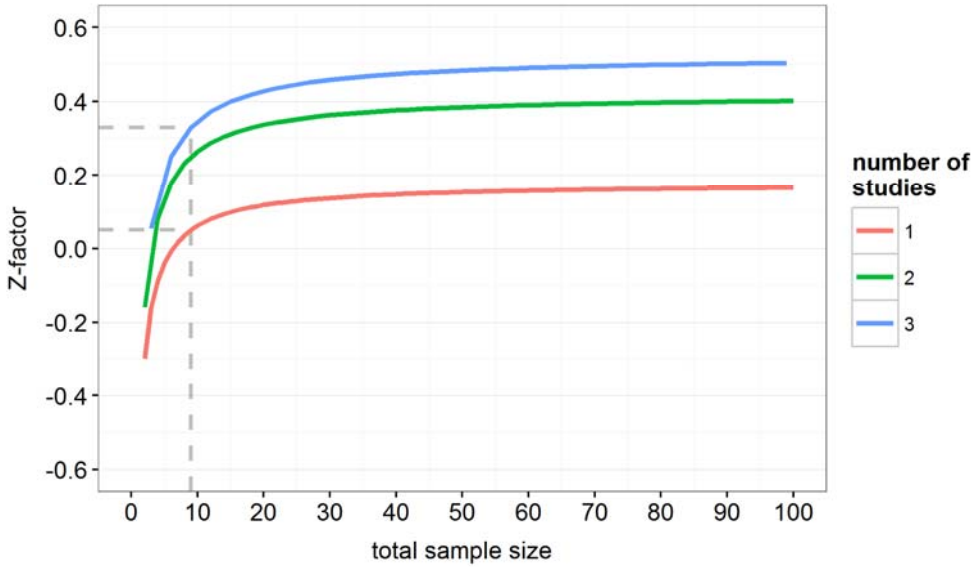
only $\sigma^2_{\text{between}}$ plays a role

Changing the design can be a remedy

$$\mathrm{Var}(\bar{Y}_{\cdot\cdot}) = \frac{\sigma^2_{\text{between}}}{S} + \frac{\sigma^2_{\text{within}}}{S \times n}$$

S = number of studies (“runs”) n = sample size within a study

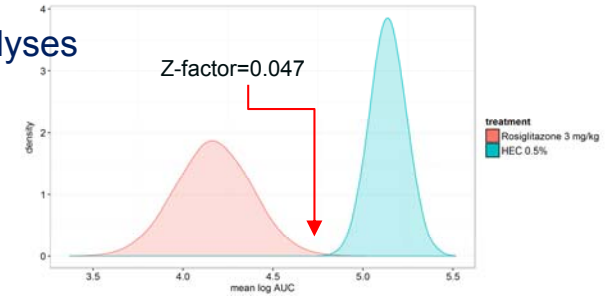
Overall z-factor (3/3)

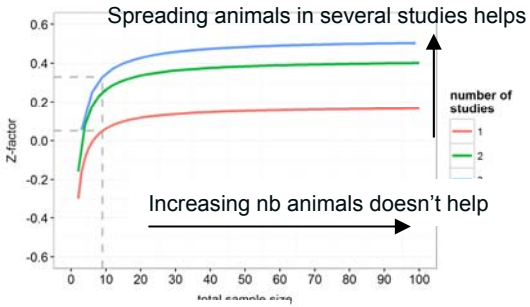


An example (4) – optimize assay to support model

Classical analyses

Classical AUC

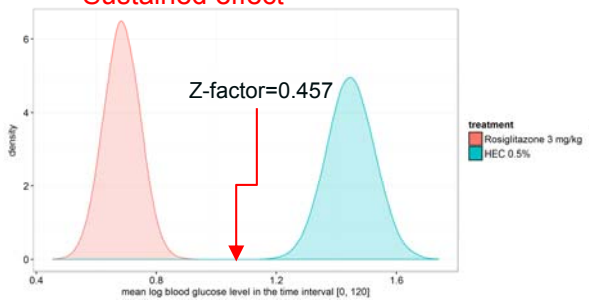




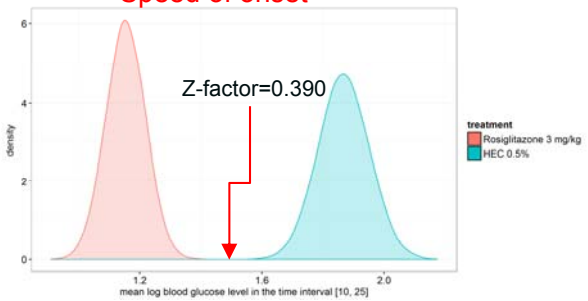
With classical analyses, the capability of the assay is not satisfactory

Longitudinal model

Sustained effect



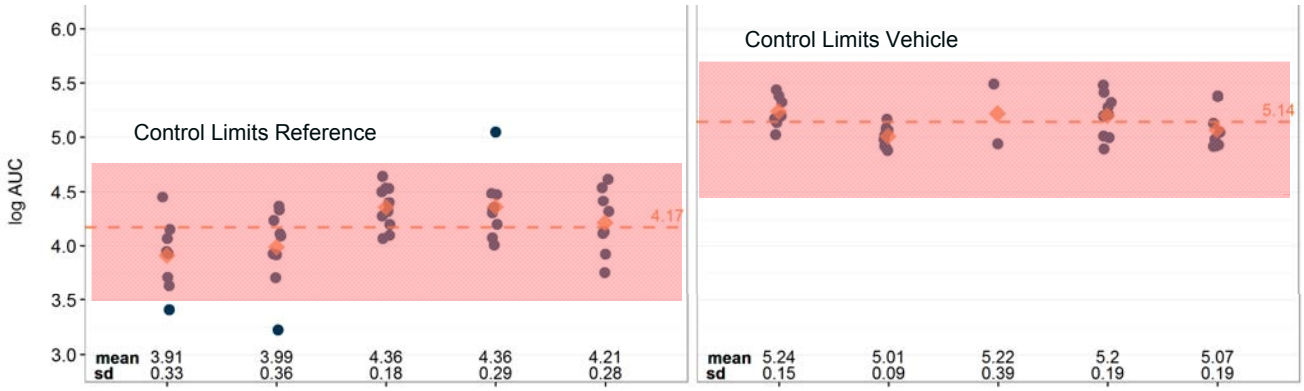
Speed of onset



With appropriate modeling, the capability of the assay is fit-for-purpose

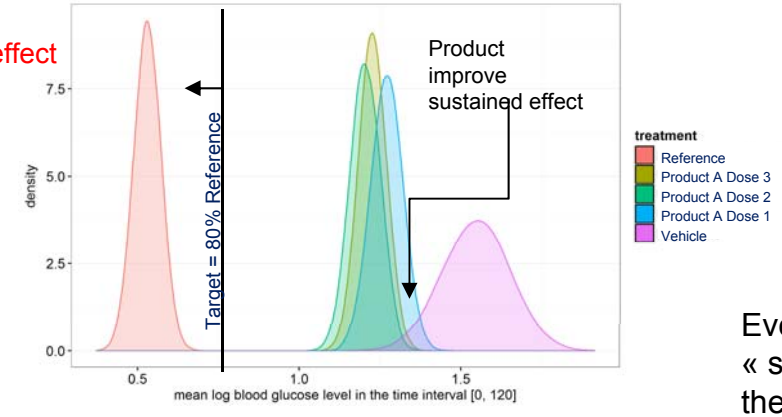
An example (5) – Control and improve capability

To guarantee Capability, future study results should fall within the control limits

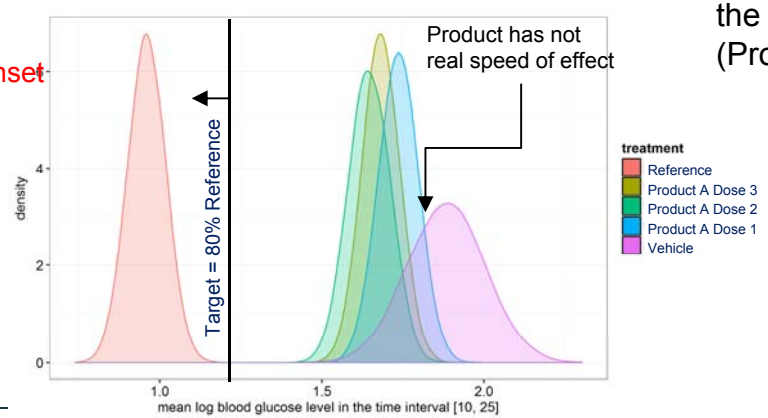


An example (6) – Allow riskless decision making

Sustained effect



Speed of onset



Even if results are « statistically significant » the probability to achieve the target is very low (Probability of Success)

Impact

- Improve overall efficiency
  - Use Prior knowledge to improve precision
  - Use historical controls to reduce # of animals
  - Provide easy to interpret statistics linked to the objective
- De-risk decision making
  - Progressively get away from p-values
  - Currently: attempt to assemble a clear picture based on a patchwork of p-values
  - Future: Provide the predictive probability of success of the objective
  - Provide interpretable (graphical) results
  - Understand and keep the uncertainty along the value chain