



Designing Functional Genomics Experiments for Successful Analysis

Mark Fernandes CRUK-CI
(Original deck: RORY STARK 18/9/2017)

Agenda

WHY PERFORM EXPERIMENTS?

WHY THINK ABOUT EXPERIMENTAL DESIGN?

WHAT MAKES FOR A WELL DESIGNED EXPERIMENT?

KEY ASPECTS OF EXPERIMENTAL DESIGN

- Experimental variables
- Power: variance and replicates
- Bias: confounding factors, randomisation, and controls

**DESIGN PARAMETERS FOR FUNCTIONAL SEQUENCING
EXPERIMENTS**

EXPERIMENTAL DESIGN PROCESS AT CRUK-CI



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Why Perform Experiments?

Why Think About Experimental Design?



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Reproducible Research



UNIVERSITY OF
CAMBRIDGE

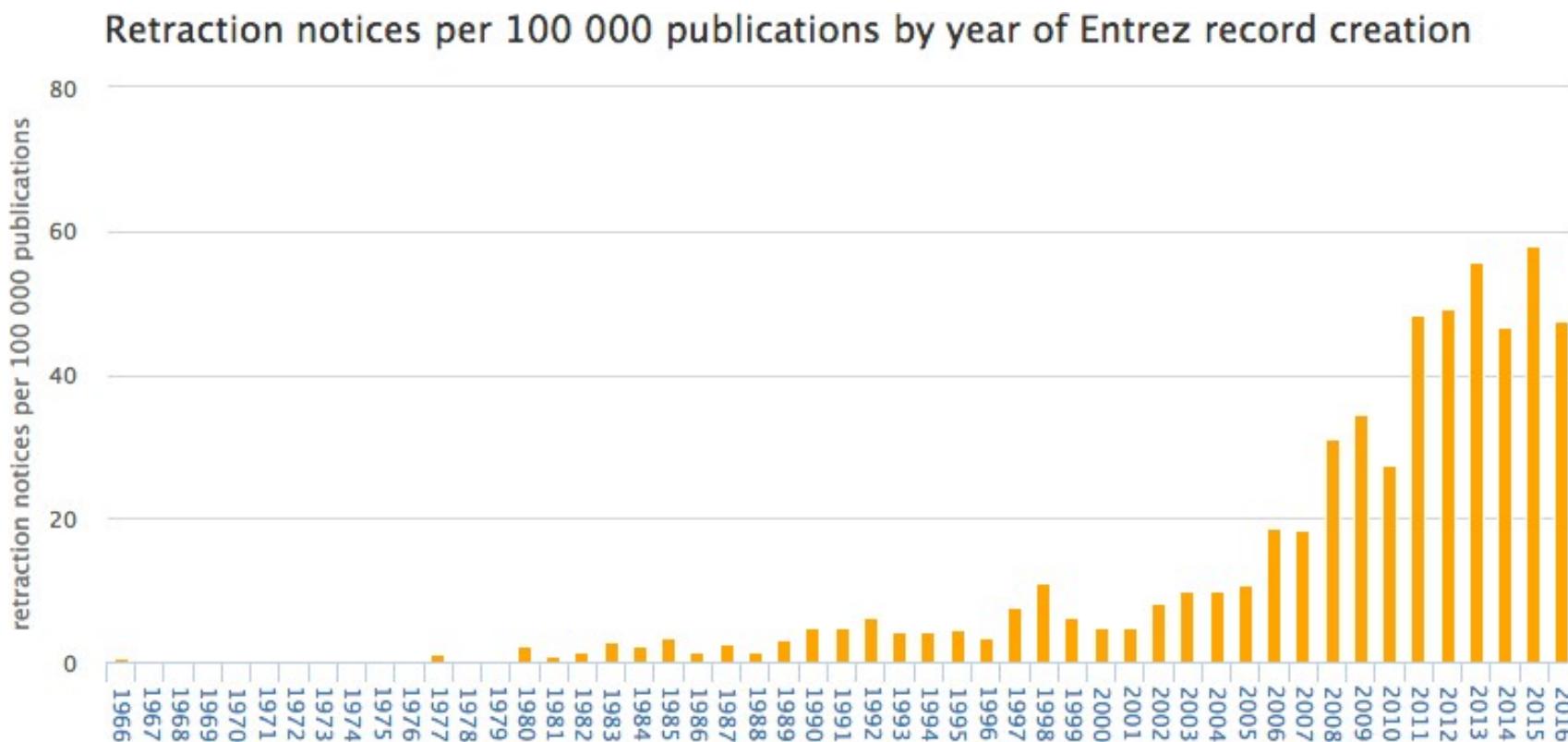


CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Crisis in Reproducible Research

http://rpubs.com/neilfws/6577_8



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Do not end up here! <https://retractionwatch.com>

47 of 53 high-profile cancer studies were not reproducible!

The screenshot shows the header of the Nature journal website. The main title "nature" is in large white letters, with "International weekly journal of science" in smaller text below it. A navigation bar includes links for Home, News & Comment, Research, Careers & Jobs, Current Issue, Archive, Audio & Video, and a partially visible "F" link. Below the navigation is a breadcrumb trail: Archive > Volume 483 > Issue 7391 > Comment > Article. The page content starts with "NATURE | COMMENT" and features the title "Drug development: Raise standards for preclinical cancer research" by C. Glenn Begley & Lee M. Ellis. It also includes sections for "Affiliations" and "Corresponding author". At the bottom, it shows the publication details: "Nature 483, 531–533 (29 March 2012) | doi:10.1038/483531a" and "Published online 28 March 2012". There are social media sharing icons at the top right.



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Need for Good Design



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Consequences of Poor Experimental Design...

- **Cost** of experimentation. We have a responsibility to CRUK donors!
- **Limited & Precious** material, esp. clinical samples.
- **Immortalization** of data sets in public databases and methods in the literature. Our bad science begets more bad science.
- **Ethical concerns** of experimentation: animals and clinical samples.



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

A Well-Designed Experiment: Should have

CLEAR OBJECTIVES

FOCUS AND SIMPLICITY

SUFFICIENT POWER (i.e. detect expected degree of change)

RANDOMISED COMPARISONS

And be

PRECISE

UNBIASED

AMENABLE TO STATISTICAL ANALYSIS

REPRODUCIBLE



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Obligatory Statistician quote: Ronald A. Fisher(1890-1962)



“TO CONSULT THE STATISTICIAN AFTER AN EXPERIMENT IS FINISHED IS OFTEN MERELY TO ASK HIM TO CONDUCT A POST MORTEM EXAMINATION. HE CAN PERHAPS SAY WHAT THE EXPERIMENT DIED OF.” (1938)



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Aspects of Experimental Design

EXPERIMENTAL FACTORS

VARIABILITY

- Sources of Variance
- Replicates

BIAS

- Confounding factors
- Randomisation wherever a decision is to be made
 - Controls for both measured and unmeasured factors
- Controls



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Experimental Factors



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Experimental Factors

FACTORS: ASPECTS OF EXPERIMENT THAT CHANGE AND INFLUENCE THE OUTCOME OF THE EXPERIMENT

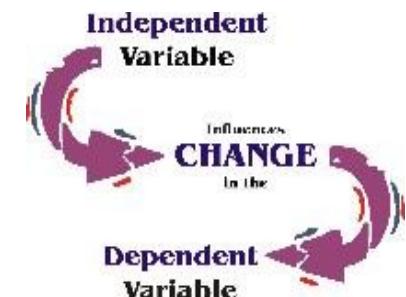
- e.g. time, weight, drug, gender, ethnicity, country, plate, cage etc.

VARIABLE TYPE DEPENDS ON TYPE OF MEASUREMENT:

- Categorical (nominal) , e.g. gender
- Categorical with ordering (ordinal), e.g. tumour grade
- Discrete, e.g. shoe size, number of cells
- Continuous, e.g. body weight in kg, height in cm

INDEPENDENT AND DEPENDENT VARIABLES

- Independent variable (IV): what you change
- Dependent variable (DV): what changes due to IV
- “If (**independent** variable), then (**dependent** variable)”



Capturing Variance



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Sources of Variation

BIOLOGICAL “NOISE”

- Biological processes are inherently stochastic
- Single cells, cell populations, individuals, organs, species....
- Timepoints, cell cycle, synchronized vs. unsynchronized

TECHNICAL NOISE

- Reagents, antibodies, temperatures, pollution
- Platforms, runs, operators

CONSIDER IN ADVANCE AND CONTROL

REPLICATION REQUIRED TO CAPTURE VARIANCE



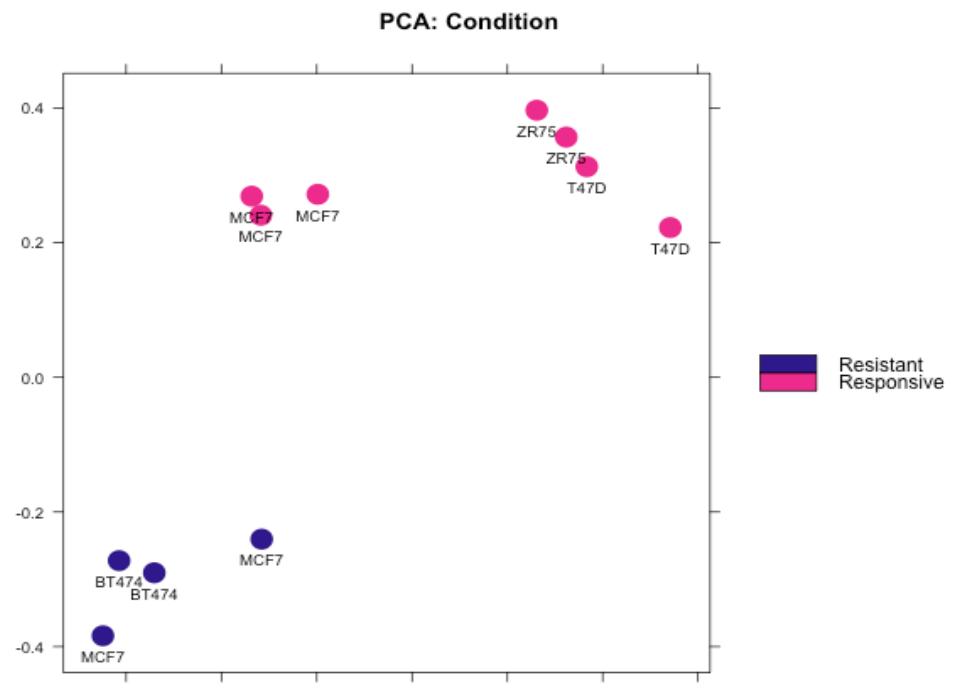
Types of Replication

BIOLOGICAL REPLICATION:

- *In vivo*:
 - Patients
 - Mice
- *In vitro*:
 - Different cell lines
 - Re-growing cells (passages)

TECHNICAL REPLICATION:

- Experimental protocol
- Measurement platform (i.e. sequencer)

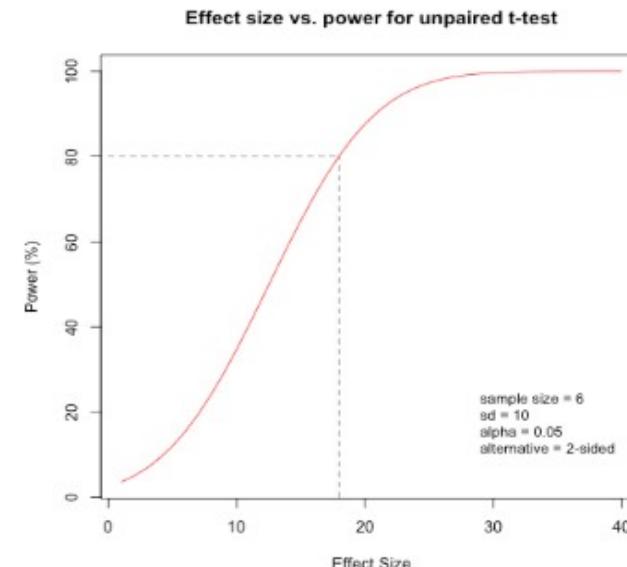


How many samples?

WHY DO YOU NEED REPLICATES?

CALCULATING APPROPRIATE SAMPLE SIZES

- Power calculations
 - Planning for precision
 - Resource equation
-
- Power: the probability of detecting an effect of a specified size if present.
 - Identify and control the sources of variability
 - Biological variability
 - Technical variability
 - Using appropriate numbers of samples (sample size/replicates)
 - Power calculations estimate sample size required to detect an effect *if degree of variability is known*
 - Depends on δ , n, sd, α , H_A
 - If adding samples increases variability, that alone won't add power!



Confounding Factors and Bias



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

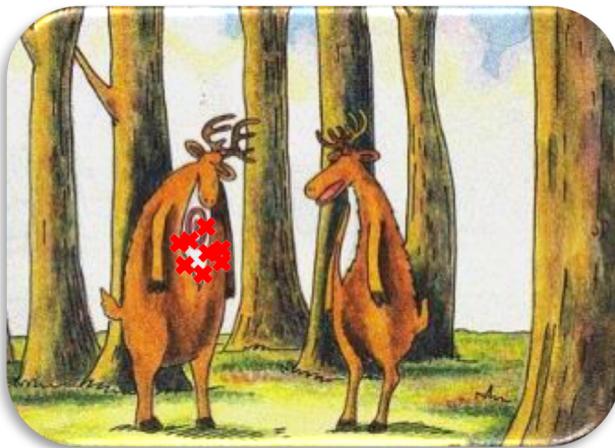
Precision, Accuracy & Bias

(according to Gary Larson's Far Side)

Accurate

Biased

Precise



Imprecise



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

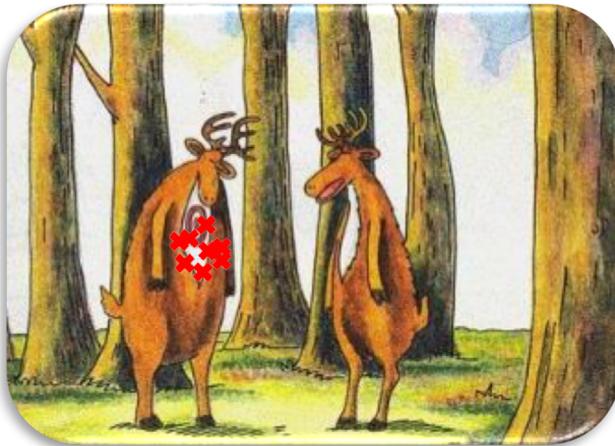
Precision, Accuracy & Bias

(according to Gary Larson's Far Side)

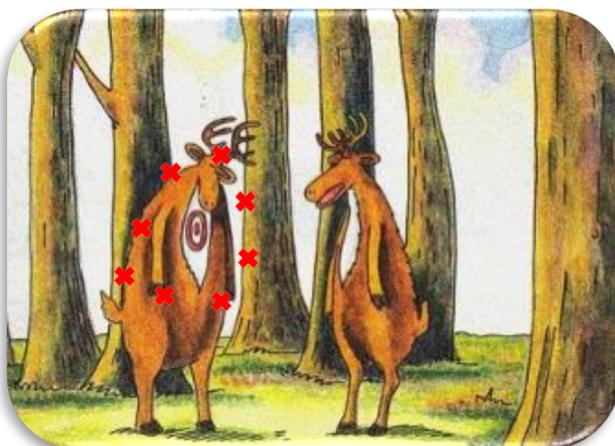
Accurate

Biased

Precise



Imprecise



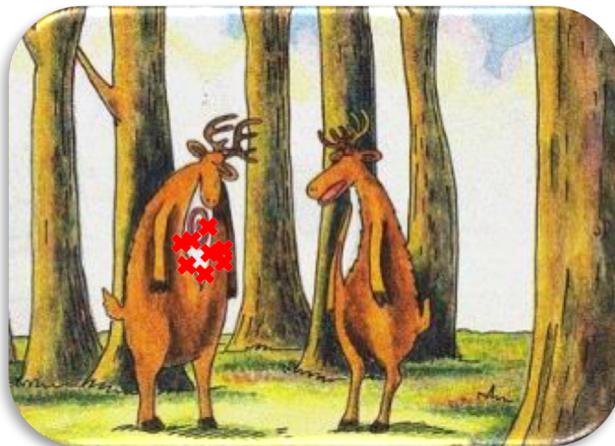
CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

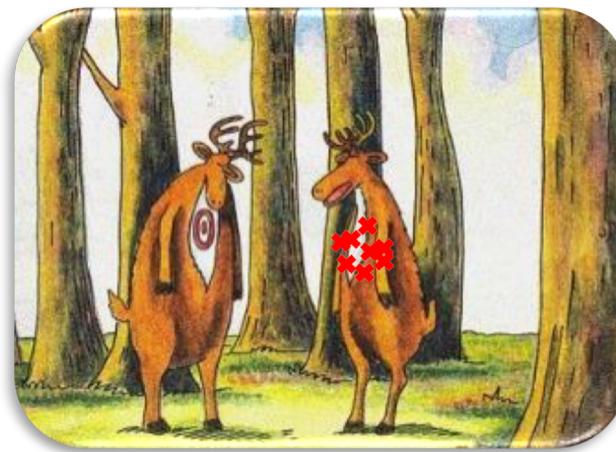
Precision, Accuracy & Bias

(according to Gary Larson's Far Side)

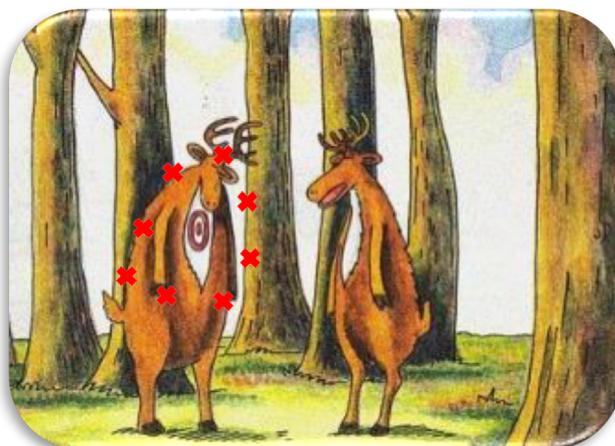
Accurate



Biased



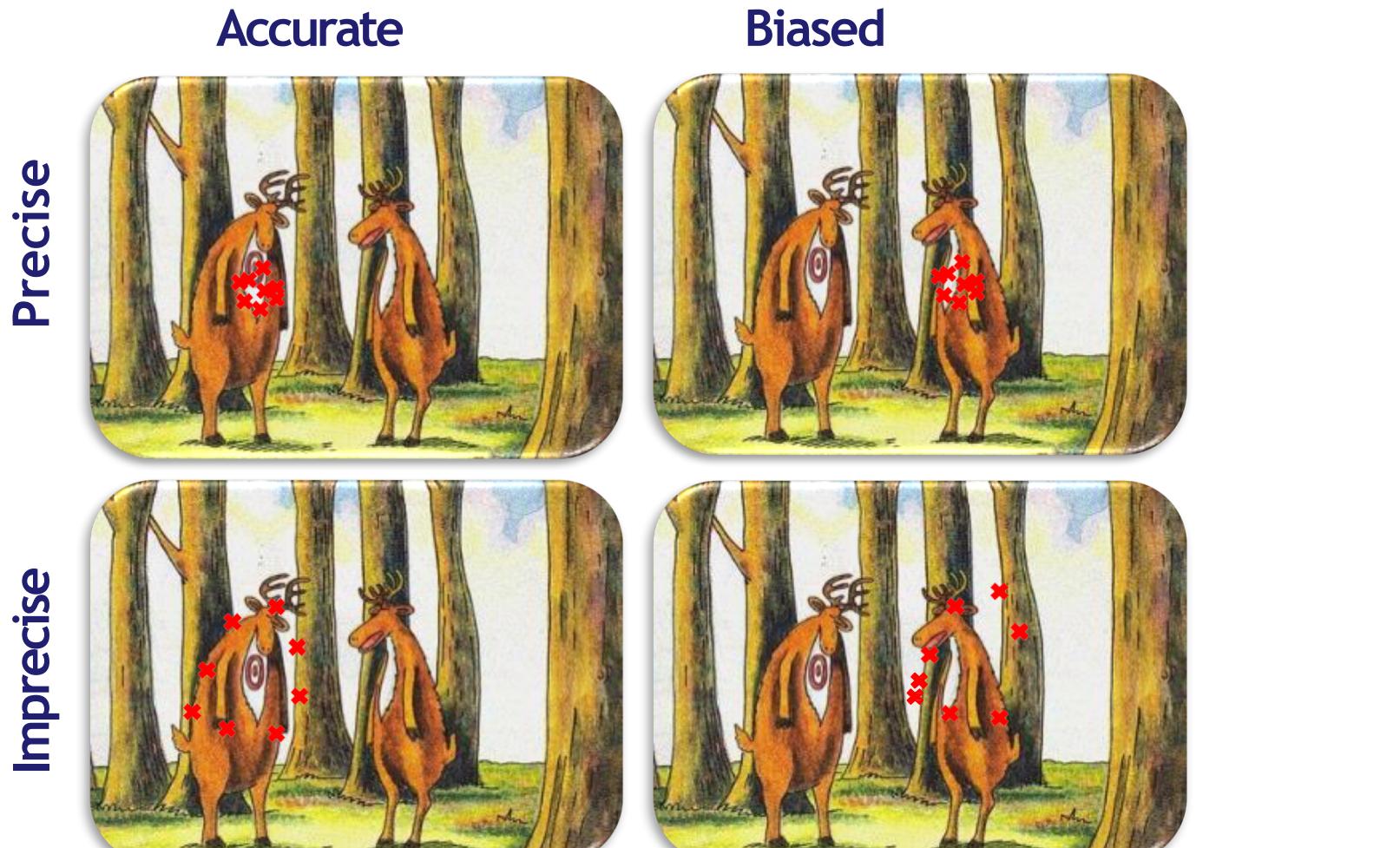
Precise



Imprecise

Precision, Accuracy & Bias

(according to Gary Larson's Far Side)

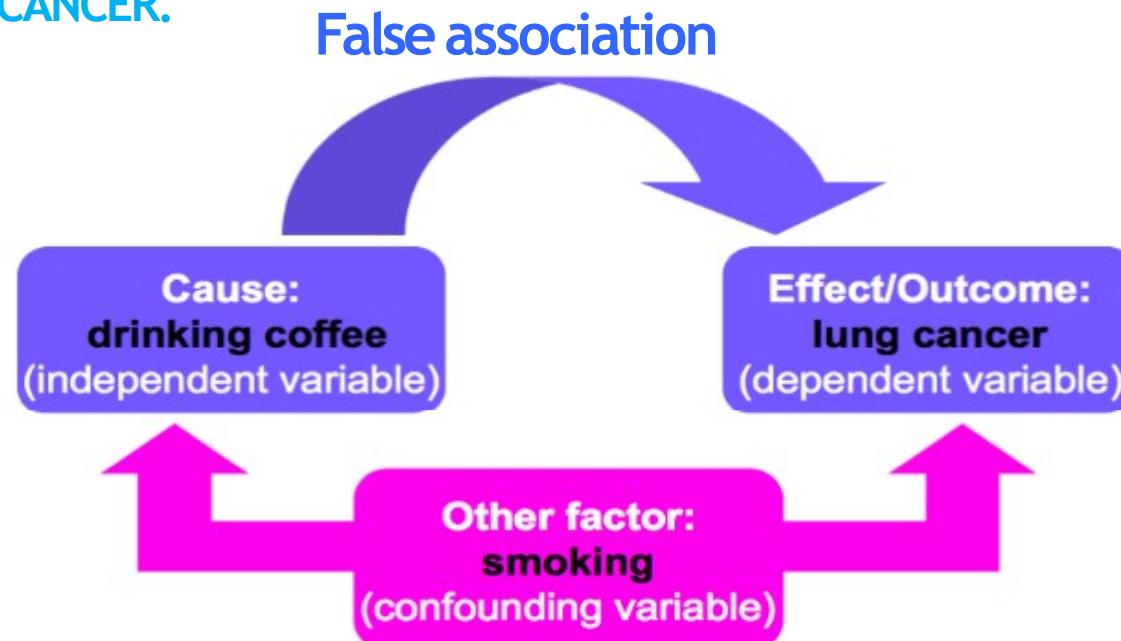


Confounding Factors

ALSO KNOWN AS EXTRANEous, HIDDEN, LURKING OR MASKING FACTORS,
OR THE THIRD VARIABLE OR MEDIATOR VARIABLE.

MAY MASK AN ACTUAL ASSOCIATION OR FALSELY DEMONSTRATE AN APPARENT
ASSOCIATION BETWEEN THE INDEPENDENT & DEPENDENT VARIABLES.

HYPOTHETICAL EXAMPLE WOULD BE A STUDY OF COFFEE DRINKING AND LUNG
CANCER.



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Confounding Factors

OTHER EXAMPLES:

- Democrats were less satisfied with their sex lives than Republicans. (ABC poll report).
- Slightly overweight people live longer than thin people (US Centre for Disease Control).
[\(https://www.nhs.uk/news/obesity/overweight-people-live-longer-study-claims/ \)](https://www.nhs.uk/news/obesity/overweight-people-live-longer-study-claims/) Use of BMI, existing medical observation, quality of life

INADEQUATE MANAGEMENT AND MONITORING OF CONFOUNDING FACTORS

- one of the most common causes of researchers wrongly assuming that a correlation leads to a causality.

IF A STUDY DOES NOT CONSIDER CONFOUNDING FACTORS, DON'T BELIEVE IT!



Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani,^{1,*} Nadia Solovieff,¹ Annibale Puca,² Stephen W. Hartley,¹ Efthymia Melista,³ Stacy Andersen,⁴ Daniel A. Dworkis,³ Jemma B. Wilk,⁵ Richard H. Myers,⁵ Martin H. Steinberg,⁶ Monty Montano,³ Clinton T. Baldwin,^{6,7} Thomas T. Perls^{4,*}

¹Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA. ²IRCCS Multimedica, Milano, Italy; Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate, 20122, Italy. ³Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA. ⁴Section of Geriatrics, Department of Medicine, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. ⁵Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA. ⁶Departments of Medicine and Pediatrics, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. ⁷Center for Human Genetics, Boston University School of Medicine, Boston, MA 02118, USA.

- **GWAS STUDY: 800 CENTENARIANS VS. CONTROLS**
- **FOUND 150 SNPS PREDICTING CENTENARIANS WITH 77 % ACCURACY**
- **PROBLEM: THEY USED DIFFERENT SNP CHIPS FOR CENTENARIANS AND CONTROLS**
- **RETRACTED IN 2011 FOLLOWING INDEPENDENT REVIEW AND QC OF DATA**

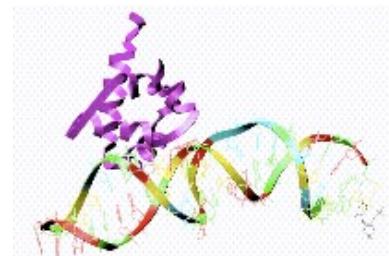
<http://www.the-scientist.com/blog/display/57558/>



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Technical Confounding Factors: Batch Effects



RNA Extraction

Day1, Plate 1

	1	2	3	4	5	6	7	8	9	10	11	12
A	●											
B												
C												
D												
E												
F												
G												
H												

Control

Day2, Plate 2

	1	2	3	4	5	6	7	8	9	10	11	12
A	●											
B												
C												
D												
E												
F												
G												
H												

Treatment 1

Day3, Plate 3

	1	2	3	4	5	6	7	8	9	10	11	12
A	●											
B												
C												
D												
E												
F												
G												
H												

Treatment 2

The difference between Control, Treatment 1 and Treatment 2 is confounded by day and plate.

Solutions

RANDOMISATION

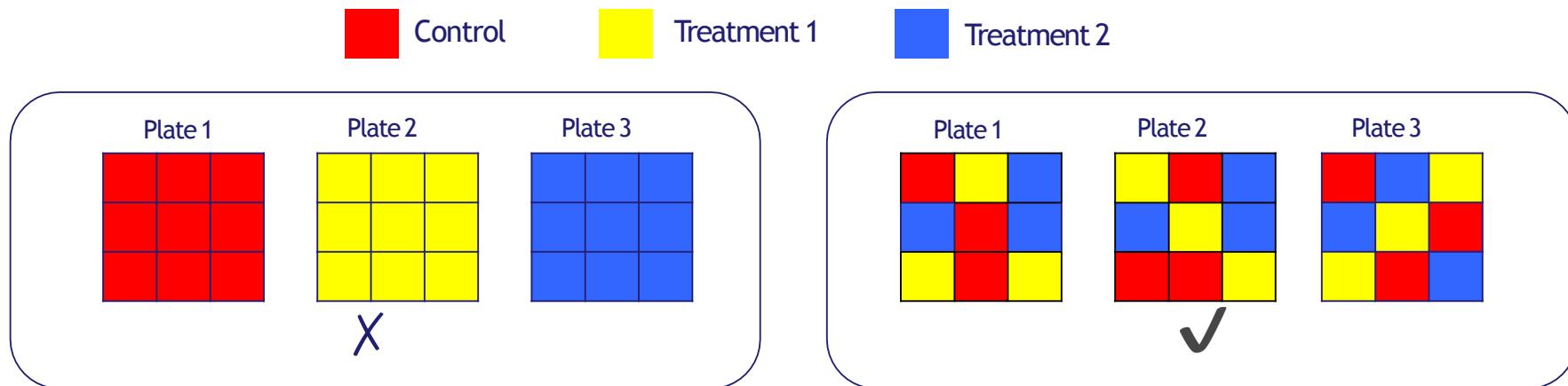
- Statistical analysis assume randomised comparisons
- May not see issues caused by non-randomised comparisons
- Make every decision random not arbitrary

BLINDING (removing the influence of condition knowledge)

- Especially important where subjective measurements are taken
- Every experiment should reach its potential degree of blinding

Randomised Block Design

Blocking is the arranging of *experimental units* in groups (blocks) that are similar to one another.



RBD across plates so that each plate contains spatially randomised equal proportions of:

- Control
- Treatment 1
- Treatment 2

controlling plate effects.

Randomised Block Design

Good design example: Alzheimer's study from GlaxoSmithKline

Plate effects by plate

Left PCA plot shows *large plate effects*.

Each colour corresponds to a different plate

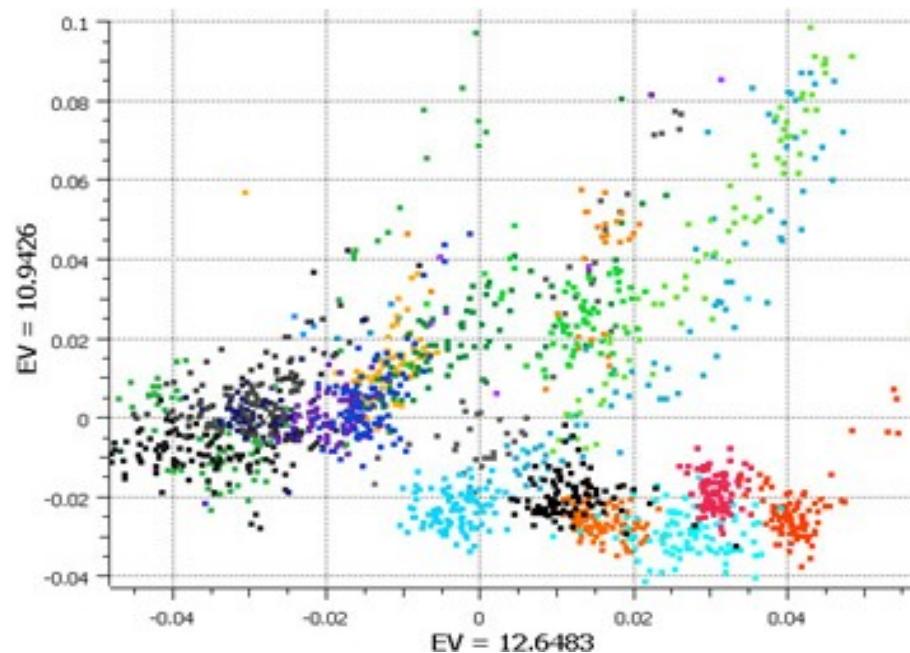
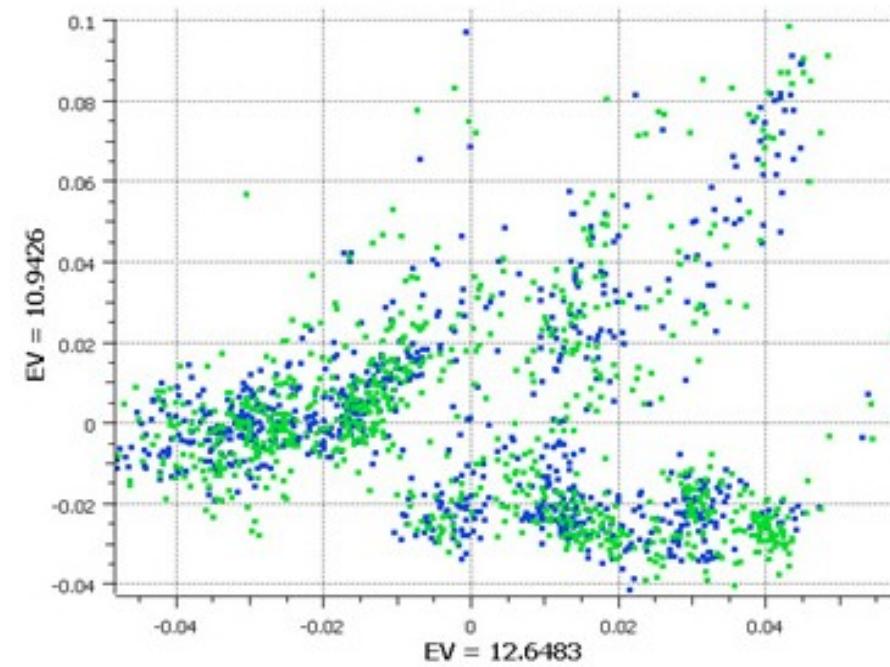


Plate effects by case/control

Right PCA plot shows each plate cluster contains *equal proportions* of cases (blue) and controls (green).



<http://blog.goldenhelix.com/?p=322>

Despite a plate effect being in existence, each plate has the same proportions of cases & controls

Experimental Controls



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Experimental Controls

CONTROLLING ERRORS

- Type I: False Positives (reject true H_0)
 - Use Negative controls: A group that should have minimal or no effect
- Type II: False Negative (fail to reject a false H_0)
 - Use Positive controls: A group where known response expected

TECHNICAL CONTROLS

- Detect/correct technical biases
- Normalise measurements (quantification) e.g. RNA spike-in

Examples of Experimental Controls

WILD-TYPE ORGANISM (KNOCKOUTS)

INACTIVE siRNA (SILENCING)

VEHICLE (TREATMENTS)

INPUT: FRAGMENTED CHROMATIN (CHIP)

SPIKE-INS (QUANTIFICATION/NORMALISATION)

“GOLD STANDARD” DATAPOINTS

MULTI-LEVEL CONTROLS

- e.g. contrast Vehicle/Input vs. Treatment/Input

Design Parameters for Sequencing Experiments



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Design Issues: Sequencing Experiments

PLATFORMS

LIBRARY PREPS

MULTIPLEXING AND POOLING STRATEGIES

SINGLE-END VS PAIRED END

SEQUENCING DEPTH

- Coverage
- Lanes

VALIDATION

- Knock-downs
- Pull-downs



Experimental Design process at CRUK-CI



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Establishing an experimental design process

- Students required to take (this) Experimental Design class
- All sequencing and proteomics experiments require experimental design review meeting
 - Simple form: [EDM Form.docx](#)
 - Attended by Scientists, Genomics/Proteomics Core, Bioinformatics Core, Statistician
 - Project opened in LIMS afterwards
- Randomisation and Layouts
 - Checkpoint for experiment
 - Project cleared for sample submission
- Keys:
 - Form and meeting not onerous
 - (Currently) not chargeable
 - Scientists agree process improves experiments!

Acknowledgements

Others who contributed to
these lecture notes:

Rory Stark

Sarah Vowler

Chandra Chilamakuri

James Hadfield

Jing Su



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

CRI Experimental Design Meetings

TUESDAY 30 MIN SLOTS (2:00-3:00PM) WITH BIOINFORMATICS & GENOMICS/PROTEOMICS CORES

DISCUSSION:

- Planning, time-scale, cost, aims, scope, questions
- Choosing the correct technology
- Technical issues e.g. what sequencing depth?
- Sample collection and processing methods
- Sample information (meta-data) collection
- Randomisation, Blocking and Replication issues
- Analyst?
- Pilot study?
- Effect size & Sample-size calculation?



Practical: Investigation into the effect of RAR α on transcription in breast cancer tissue treated with estrogen

- RAR α is a transcription factor that appears to interact with estrogen (E2) in ER+ breast cancer.
- We are interested in characterising this interaction by looking at how gene expression changes in breast cancer cells treated with estrogen when RAR α is not present (using a siRNA in cultured cells).
- We wish to identify which estrogen- induced and estrogen- repressed genes are impacted by the presence or absence of RAR α , and to analyse the key pathways involved.

Experimental Design Practical Questions I

1. What are your objectives?
2. What are you measuring?
3. What are your primary sample groups of interest?
4. What controls will you use each type of sample group?
5. What constitutes a replicate in this experiment? Are they biological or technical? How many samples/replicates should be collected?
6. Sketch out the design as a matrix, with sample numbers
7. What sample group comparisons (contrasts) will you make with the data? Which gene set(s) will you use for pathway analysis?
8. What are possible confounding factors and sources of bias?



Experimental Design Practical Questions I

9. How will you confirm effective silencing?
10. What information about your experiment should be recorded to help identify any problems should there be any?
11. Will you be multiplexing samples? How will you assign barcodes? Will you use pooled libraries? How many pools? How will samples be assigned to pools?
12. What are the sequencing parameters you need to be aware of (e.g. sequencing type and depth)?
13. What other types of data might be useful to assay, and how might the sequencing parameters need to change to accommodate this?
14. Can you think of any other design related issues that could/should be addressed?

