Introduction to Experimental Design at CRUK-CI

Chandra Chilamakuri
James Hadfield
Jing Su
Rory Stark

tinyurl.com/cruk-edesign

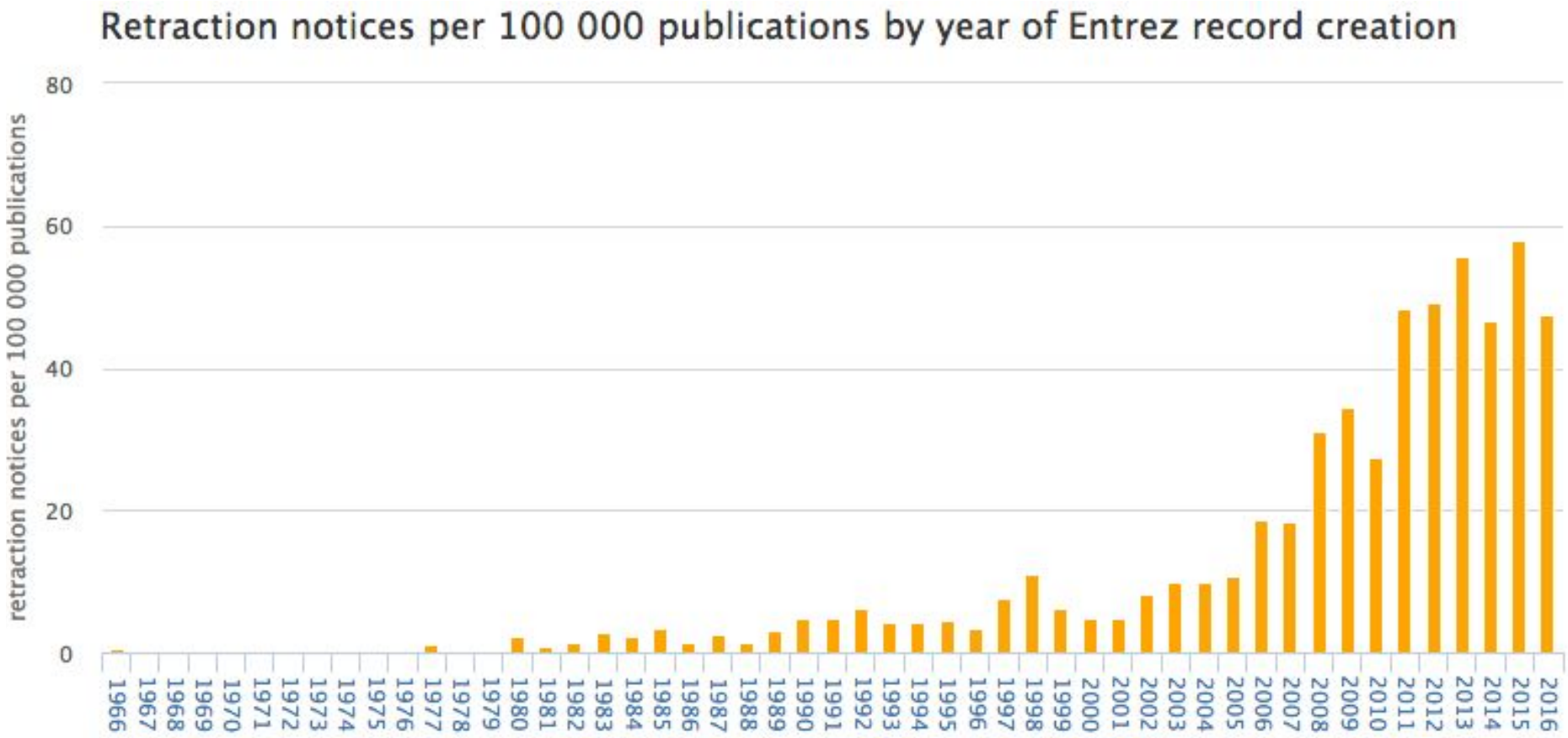
Agenda

- Why perform experiments?
- Why think about experimental design?
- What makes for a well designed experiment?
- Aspects of experimental design
  - Experimental variables
  - Power: variance and replicates
  - Bias: confounding factors, randomisation, and controls
- Experimental design types
- Experimental design at CRUK-CI

Why Perform Experiments?

Why Think About Experimental Design?

Crises in Reproducible Research!!



http://rpubs.com/neilfws/65778

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

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.

A Well-Designed Experiment

Should have

- Clear objectives
- Focus and simplicity
- Sufficient power
- Randomised comparisons

And be

- Precise
- Unbiased
- Amenable to statistical analysis
- Reproducible

Ronald A. Fisher(1890-1962)

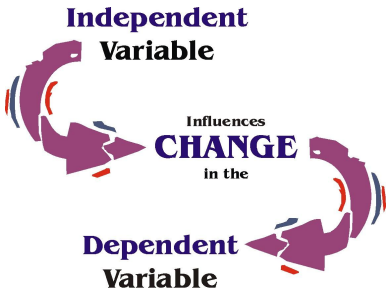
*"To consult the statistician **after** an experiment is finished is often merely to ask him to conduct a **post mortem** examination. He can perhaps say what the experiment died of." (1938)*

Aspects of Experimental Design

- Experimental Factors
- Power
  - Sources of Variance
  - Replicates
- Minimising Bias
  - Confounding factors
  - Randomisation wherever a decision is to be made
    - Controls for both measured and unmeasured factors
  - Controls

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)"

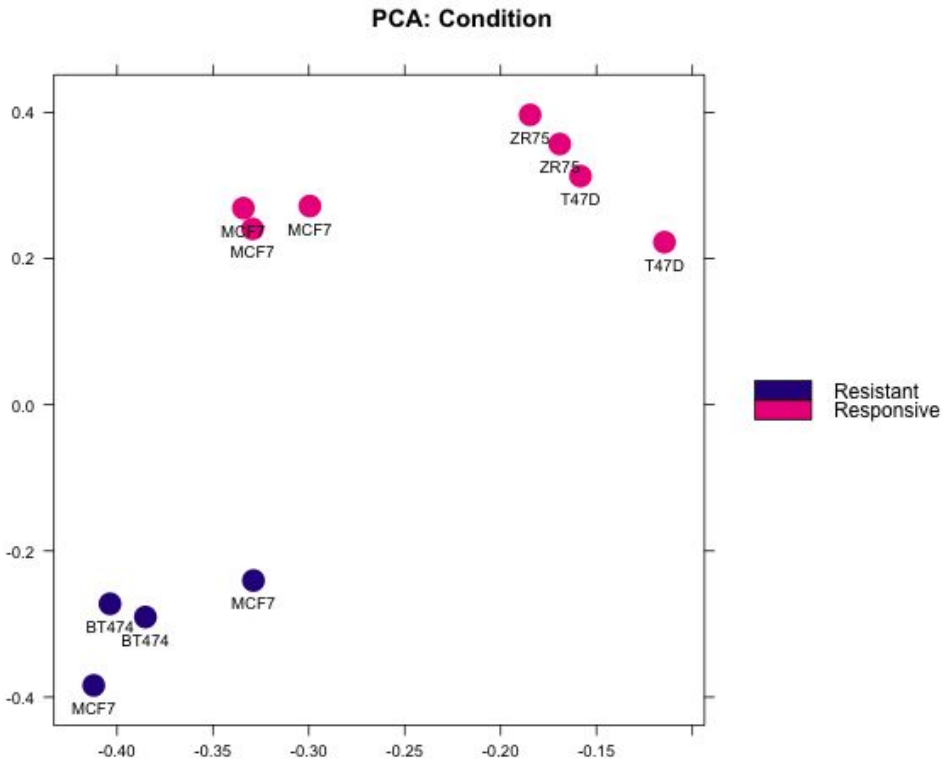


Sources of Variation

- 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

Types of Replication

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

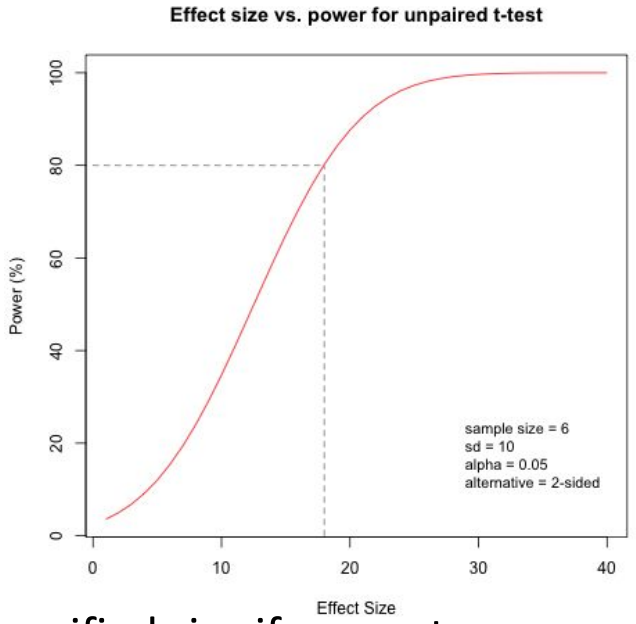


- Technical replication:
  - Experimental protocol
  - Measurement platform (i.e. sequencer)

How many samples?

- Why do you need replicates?

- Calculating appropriate sample sizes
  - Power calculations
  - Planning for precision
  - Resource equation

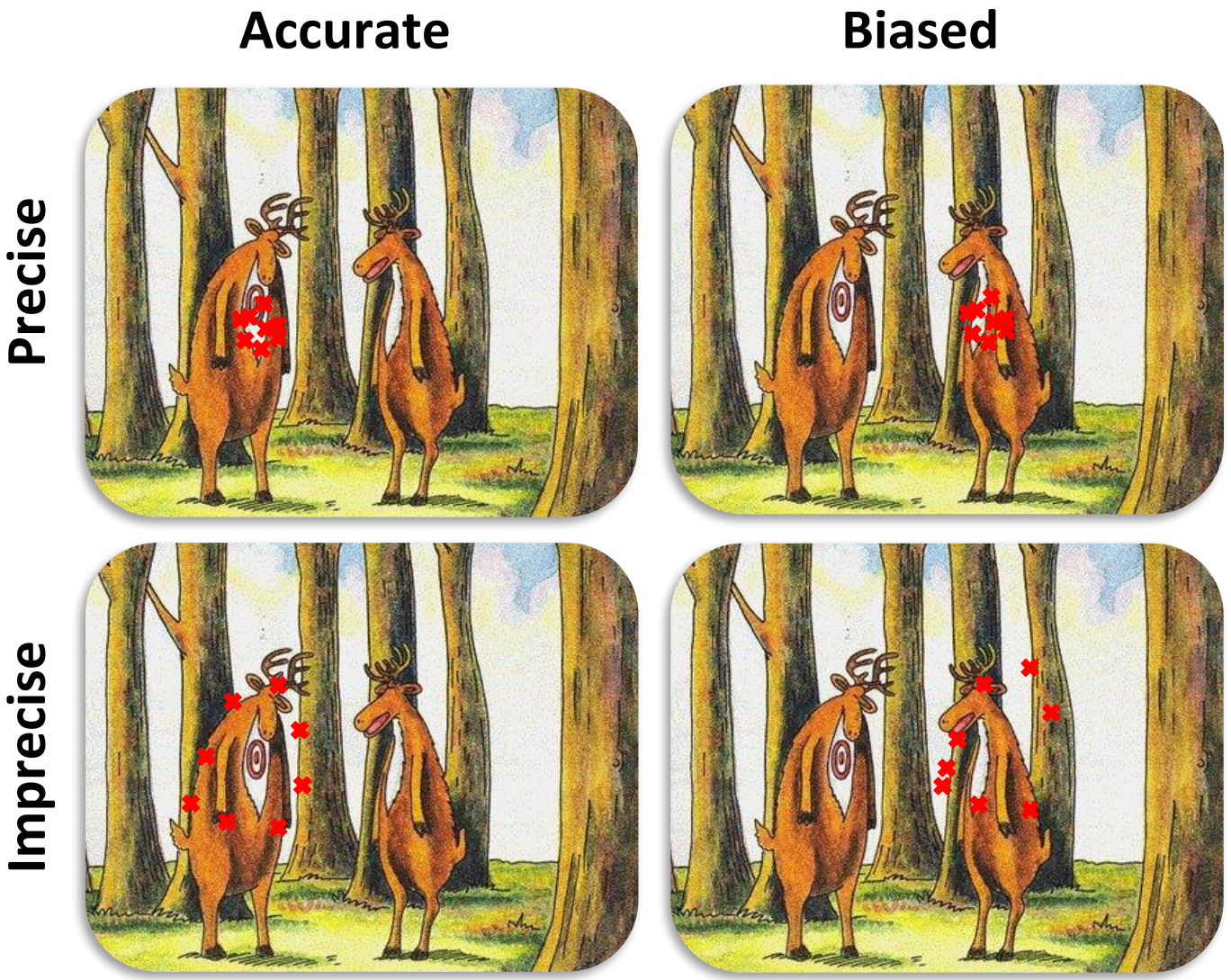


- Power: the **probability** of detecting an **effect** of a specified size if present.
  - Identify and control the **sources of variability**
    - Biological variability
    - Technical variability
  - Using **appropriate numbers** of samples (sample size/replicates)
  - Power calculations estimate 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!

Forms of Bias

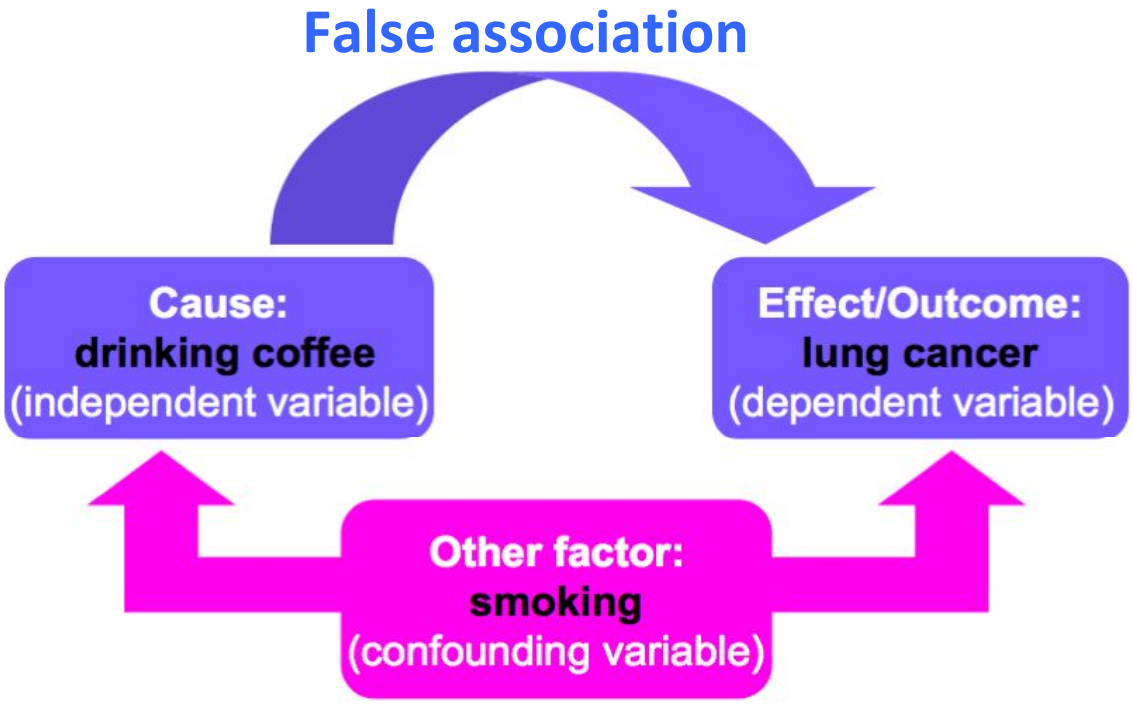
Type of Bias	Description
Selection bias	Systematic differences between **baseline characteristics** or **treatment groups** that are being compared.
Performance bias	Systematic differences between groups in exposure to factors other than the interventions of interest (aka ***confounding* or *extraneous* factors**).
Attrition bias	Systematic differences between groups due to samples being **withdrawn** from the study or **excluded** from the analyses.
Detection or Measurement bias	Systematic differences between groups in how outcomes are assessed or determined, e.g. **measurement errors** and inefficient use of data.
Reporting bias	Systematic differences between reported and unreported findings due to manipulation in the reporting of findings such as **selective or distorted reporting**, e.g. papers with more 'interesting results ' are more likely to be submitted and accepted for publication.

Precision, Accuracy & Bias



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.



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).
- **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!**

Sciencexpress Report

Genetic Signatures of Exceptional Longevity in Humans

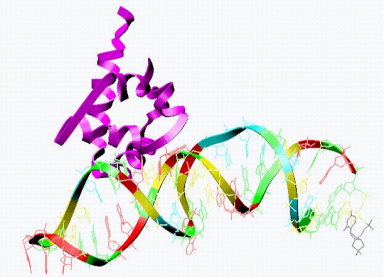
Paola Sebastiani,[1]* Nadia Solovieff,[1] Annibale Puca,[2] Stephen W. Hartley,[1] Efthymia Melista,[3] Stacy Andersen,[4] Daniel A. Dworkis,[3] Jemma B. Wilk,[5] Richard H. Myers,[5] Martin H. Steinberg,[6] Monty Montano,[3] Clinton T. Baldwin,[6,7] Thomas T. Perls[4]*

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

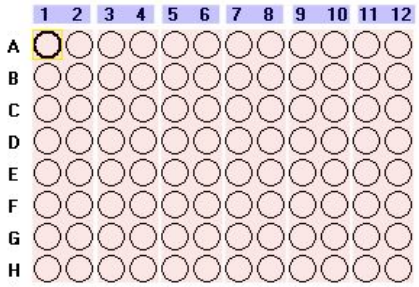
http://www.the-scientist.com/blog/display/57558/

Technical Confounding Factors: Batch Effects

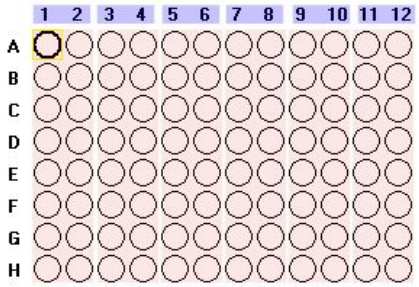


RNA Extraction

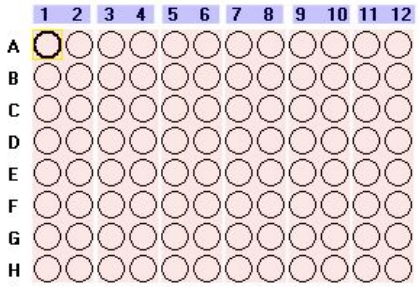
Day1, Plate 1



Day2, Plate 2



Day3, Plate 3



Control

Treatment 1

Treatment 2

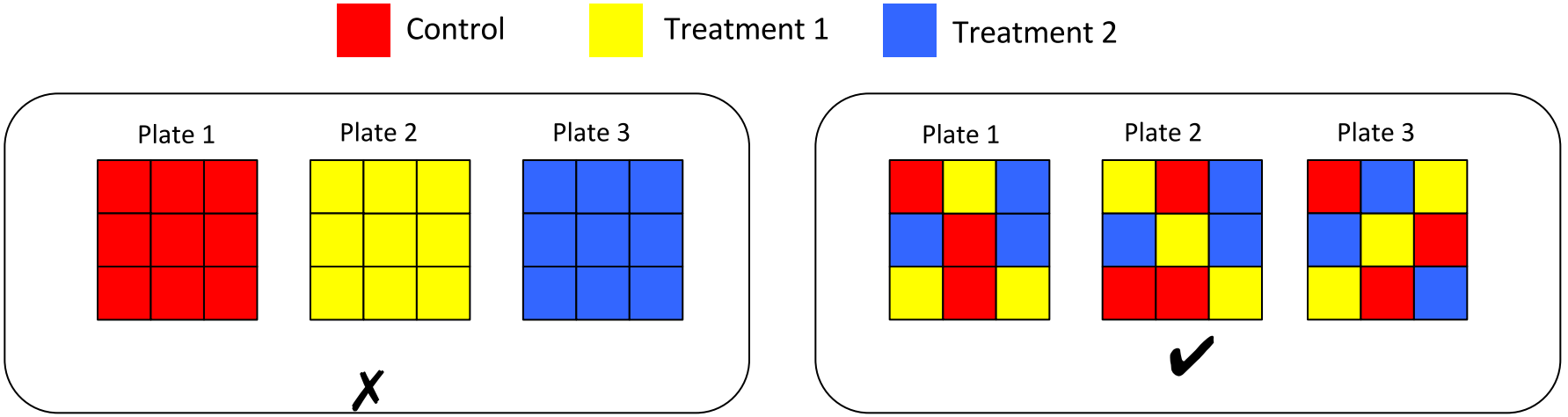
The difference between Control, Treatment 1 and Treatment 2 is confounded by **day** and **plate.**

Solutions

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

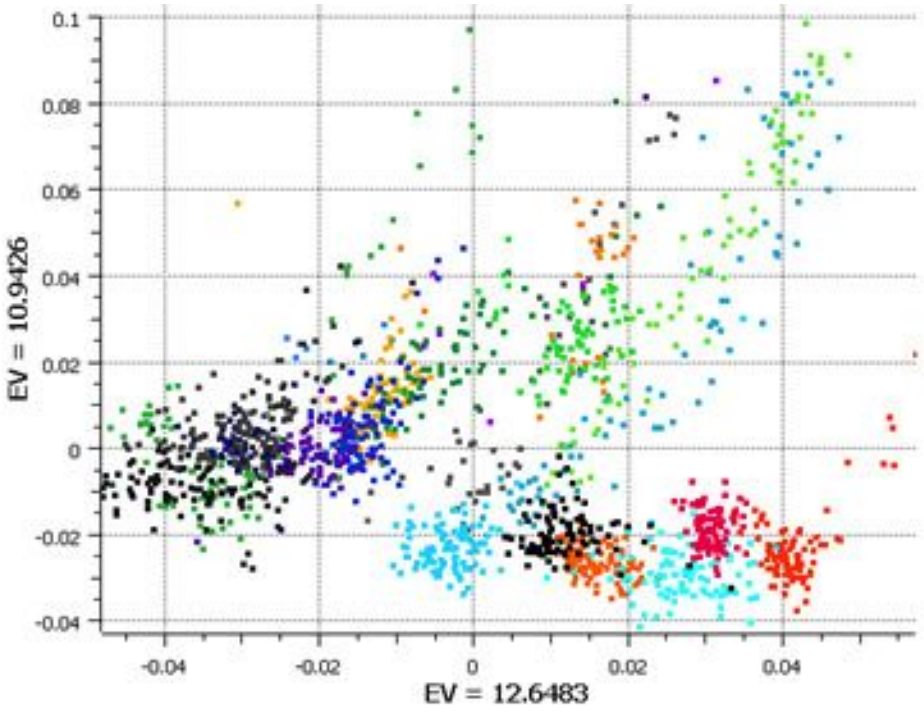
Good design example: Alzheimer's study from GlaxoSmithKline

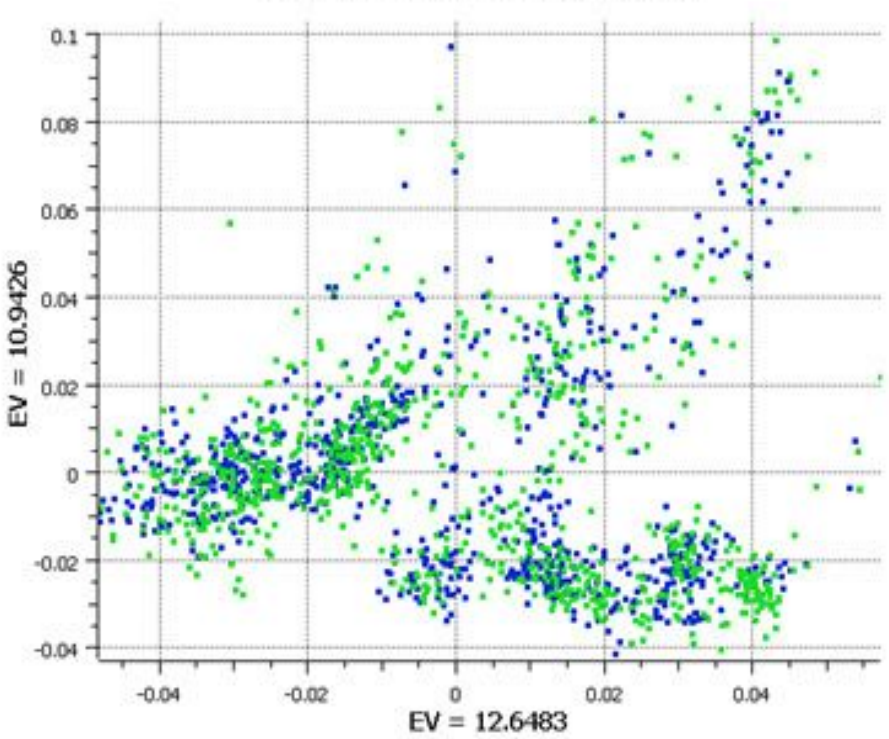
Plate effects by ***plate***

Left PCA plot show *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

Experimental Controls

- Controlling errors
  - Type I: FP
    - Negative controls: should have minimal or no effect
  - Type II: FN
    - Positive controls: known effect
- Technical controls
  - Detect/correct technical biases
  - Normalise measurements (quantification)

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 Issues: Sequencing Experiments

- Platforms (MiSeq, HiSeq, etc.)
- Library preps
- Multiplexing and pooling strategies
- Single-end vs paired end
- Sequencing depth
  - Coverage
  - Lanes
- Validation

Types of Experimental Designs

- Full random: compute associations after
- Block designs: randomisation
- Matched: tumour/normal
- Factorial/multifactorial designs: GLM
- Time series
- Hierarchical designs

CRI Experimental Design Meetings

- **Tuesday** 30 min slots (2:00-3:00pm) with Bioinformatics & Genomics Cores
- **Requirements:**
  - Email CRIExperimentalDesign@cruk.cam.ac.uk to request meeting
  - Fill in Experimental Design Form and return 1 week prior to meeting
  - **Your attendance**
  - Provide **project *background*** (a few slides from you)
- **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?

Experimental Design Guide

- https://rawgit.com/bioinformatics-core-shared-training/experimental-design/master/ExperimentalDesignManual.pdf

tinyurl.com/cruk-edesign

Practicals

1. Clinical: Identification of prognostic biomarkers in human prostate cancer patients (Rory)
2. Animal: Effects of mutant vs wildtype HHEX in liver and brain development (Jing)
3. RNA-seq: Effect of silencing different transcription factor on gene expression in a breast cancer cell line (James)
4. ChIP-seq: Transcription factor binding divergence in mice (James)

Investigate interactions between drug responsive and resistance conditions (Chandu)