Optimal study formats for preclinical studies: leveraging manufacturing experience to overcome the reproducibility crisis.

Milana Filatenkova| Manager Statistics

Bruno Boulanger | CSO

NCS2018

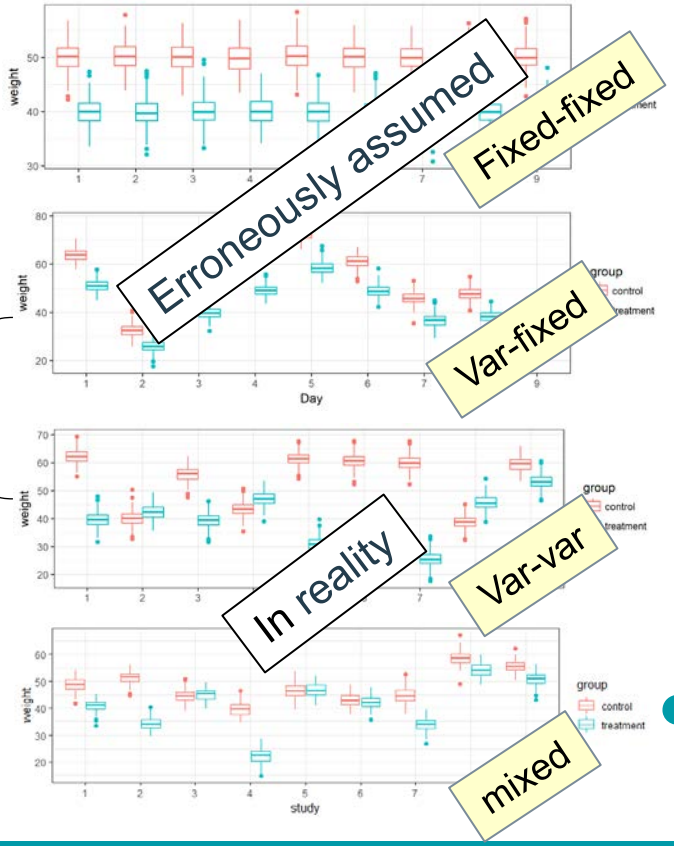
5 October 2018

Reproducibility crisis in pre-clinical pharmacology (PF)

- **The target**: reliable classification of a **treatment effect** using **animal model**
- New treatment is compared against a **control**
- **Multiple testing:** the same experiment (exposure to treatment) is repeated many times to estimate the effect of new treatments and classify new treatments
- Following a **precise experimental protocol ->** to minimize the estimate error
- **The conclusion of treatment effect significance is often not reproducible**
- **Study–study variability** due to factors that tend to alternate from study to study:
  - Animals
  - Reagents
  - Equipment
  - Lab conditions (temperature, humidity, etc.)
  - Analysts
  - Time
- Using control should help?

Study-to-study variability – WHAT are the possible scenarios?

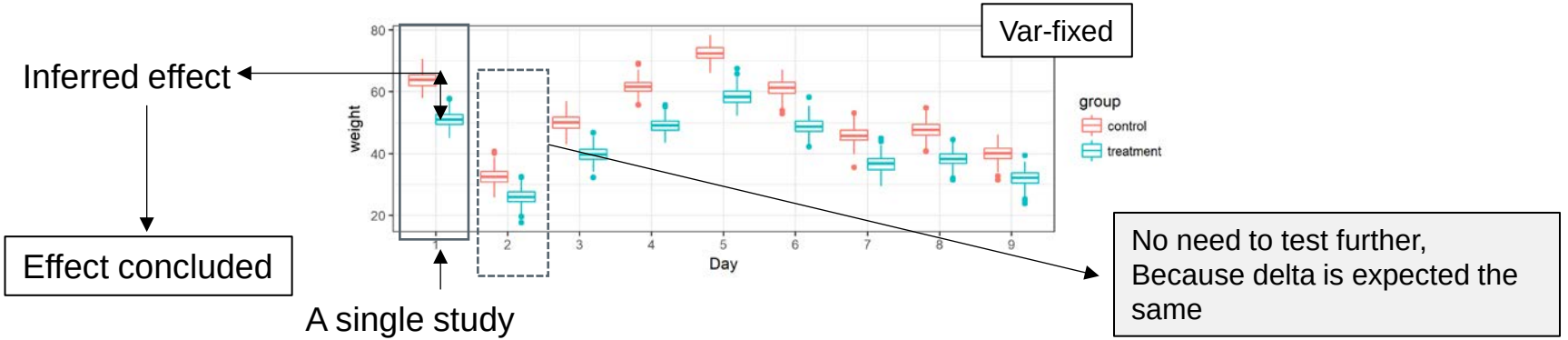
- Control: $Y_C = \mu_C + \boldsymbol{\delta_C} + \varepsilon$
- Treatment: $Y_T = \mu_C + \Delta_{treatment} + \boldsymbol{\delta_T} + \varepsilon$
- Scenario 1: Mean – **fixed**; group effect – **fixed**
- Scenario 2: Global mean – **var**ies from study to study; group effect – **fixed**
  - $\delta_C = \delta_T \sim N(0, \sigma^2_{study})$
- Scenario 3: Means – **var**y from study to study independently;
  - $\delta_C \sim N(0, \sigma^2_{study})$
  - $\delta_T \sim N(0, \sigma^2_{study})$
- Scenario 4: In reality, it is a mixture of the two scenarios (**mixed** run effect)



PF: Common approach to experimental design

- Erroneous assumption:
  - there is no study-study variability, or at least no variability in the effect size
- Homogeneous conditions -> “Better precision of effect estimate”

Commonly assumed variability scenario (Scenario 1 or 2)

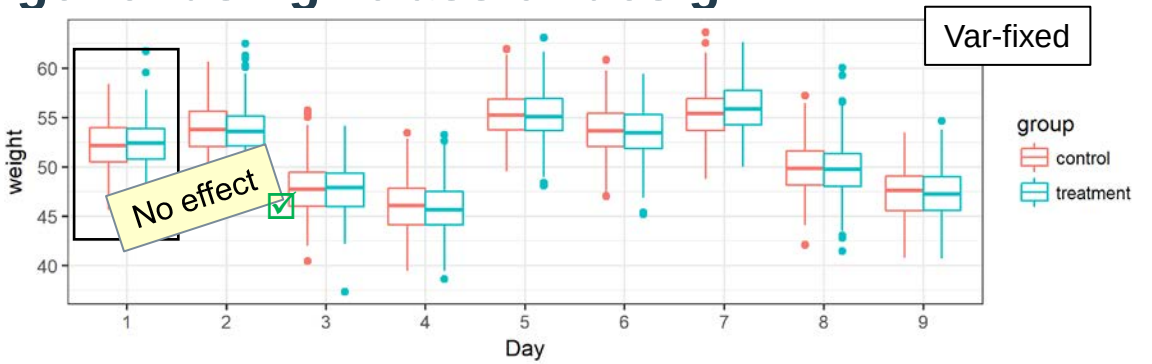


- **All the experiments are performed within controlled conditions (“classic” design)**

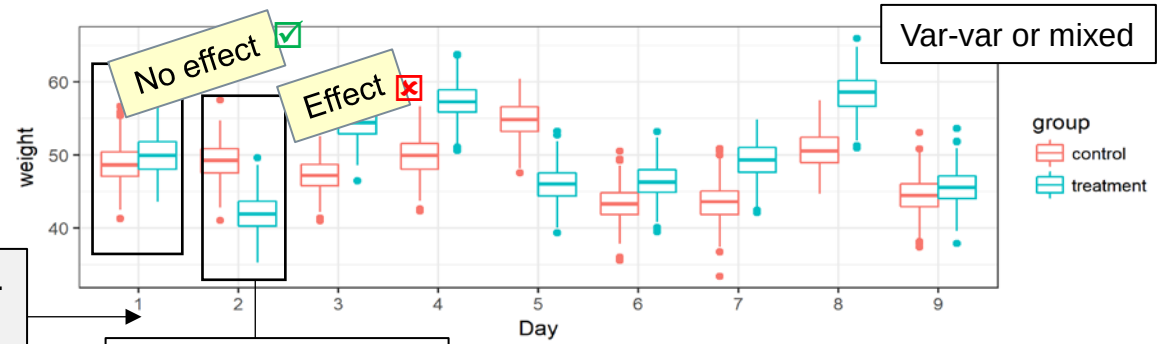
Attention: there is a problem with this approach!

Assuming no effect. Danger of using “classic” design

- $\Delta_{T-C} = \mathbf{0}$ -> simulate data
- Simulated data I (‘var-fixed’):
  - $\mu_C = \mu_T$



- Simulated data II (‘var-var’):
  - $\Delta\mu_{T-C} \sim N(0, 2\sigma^2_{study})$



Depending on the study - Day1, Day2, …
It is possible to observe an effect,
Induced by the study-study variability

‘Classic’ design

Classic design: There is a high chance of making a **biased** conclusion about the treatment effect !

Low reproducibility – WHY?

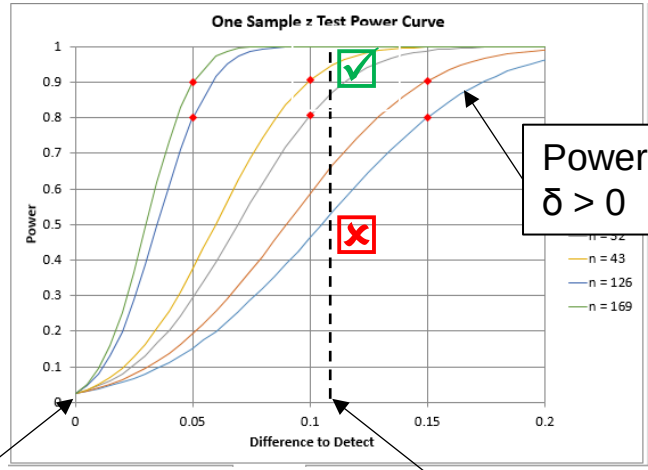
- Many reasons…
- **Inadequate study design is an obvious contributor to low reproducibility**
  - Ignoring the impact of study-to-study variability
  - Concentrating experiments within a single run

- Leading to a **biased estimation** of a treatment effect -> **not reproducible**

PF: bioassay performance assessment criteria

- Discriminative ability:
  - If $\delta > \delta^*$ – **maximize** the probability of detecting it (rejecting H_0 when H_0 is false)
    - $\Pr(H_1|H_1)$ - power
  - If $\delta = 0$ – **control** the probability of wrongly identifying it (rejecting H_0 when H_0 is true)
    - $\Pr(H_1|H_0) < \alpha$ - type I error
- Minimal number of samples to enable reliable identification of an effect
- Robustness to different variability scenarios, unknown in advance

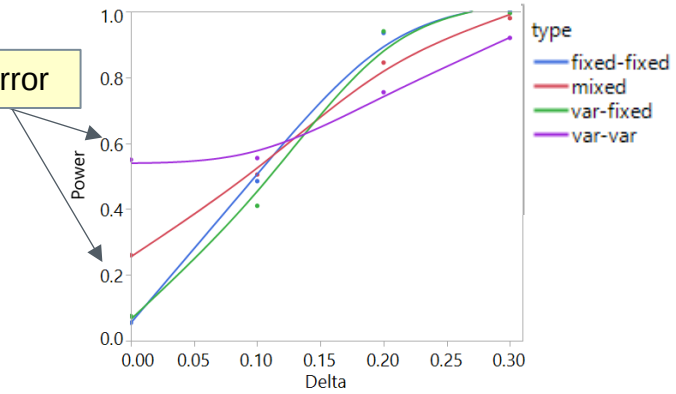
Hypothesis testing $H_0: \delta = 0; H_1: \delta > 0$ → Power curve



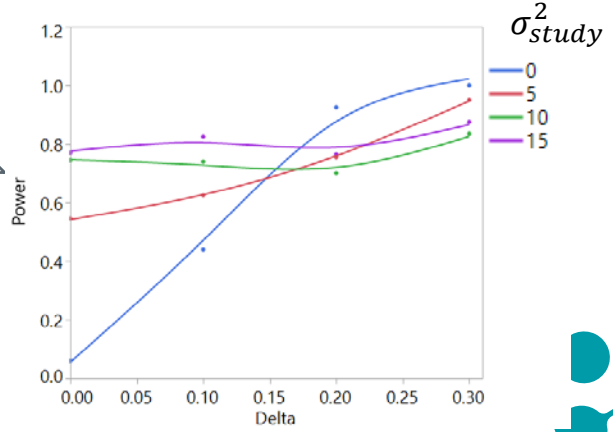
Performance of “classic design” in presence of study-study variability

- ‘Classic’ design (all –in a single study)
- Simulate 200 experiments
- Power curves - assuming variability scenarios:
  - Fixed-fixed
  - Var-fixed
  - Var-var
  - Mixed
- Power curves – assuming **‘var-var’** scenario for a range of **study random effect sizes**
  - $\delta_C \sim N(0, \sigma^2_{study})$
  - $\delta_T \sim N(0, \sigma^2_{study})$

Inflated Type 1 error



Type 1 error increases with the magnitude of the study effect



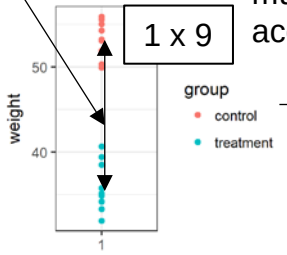
Attention: In the presence of significant study-study variability, there is an inflation of Type I error, when using ‘classic’ design !

Alternative study design

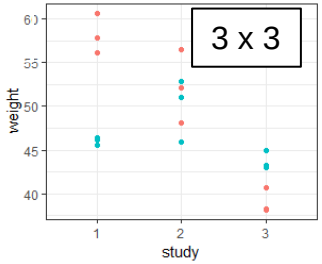
- Current approach -> ‘’classic” design (all in one study)
- Convention (USP 1032):
  - Convert **bias** into lack of precision of the estimate
  - Control **precision** with sample size
- Concept of study format:

  N(total) = X (runs) x Y (replicates)

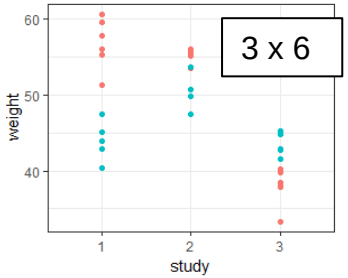
Δ_{C-T} - is likely to be **biased**



spread samples over multiple studies to account for the bias



Add extra samples to improve precision



Tested study formats

Classic (Common practice):

Model: $Y_{ij} = \mu + \Delta_i + \varepsilon_{ij}$

Optimal designs:

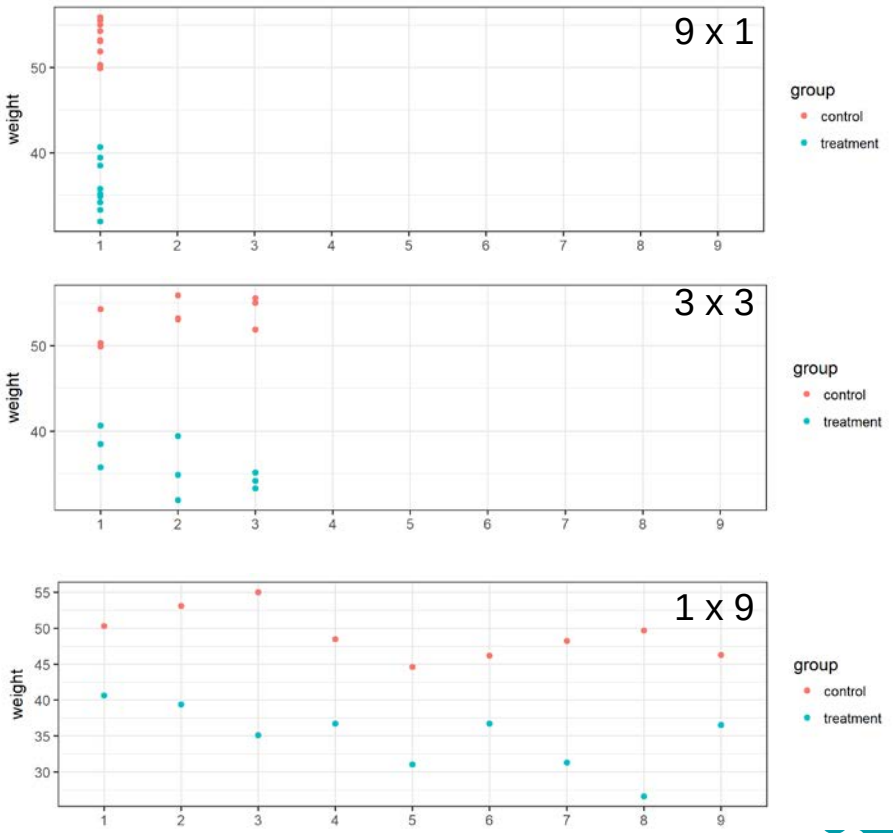
- Intermediate design

Model: $Y_{ij} = \mu + \Delta_i + \delta_j + \varepsilon_{ij}$

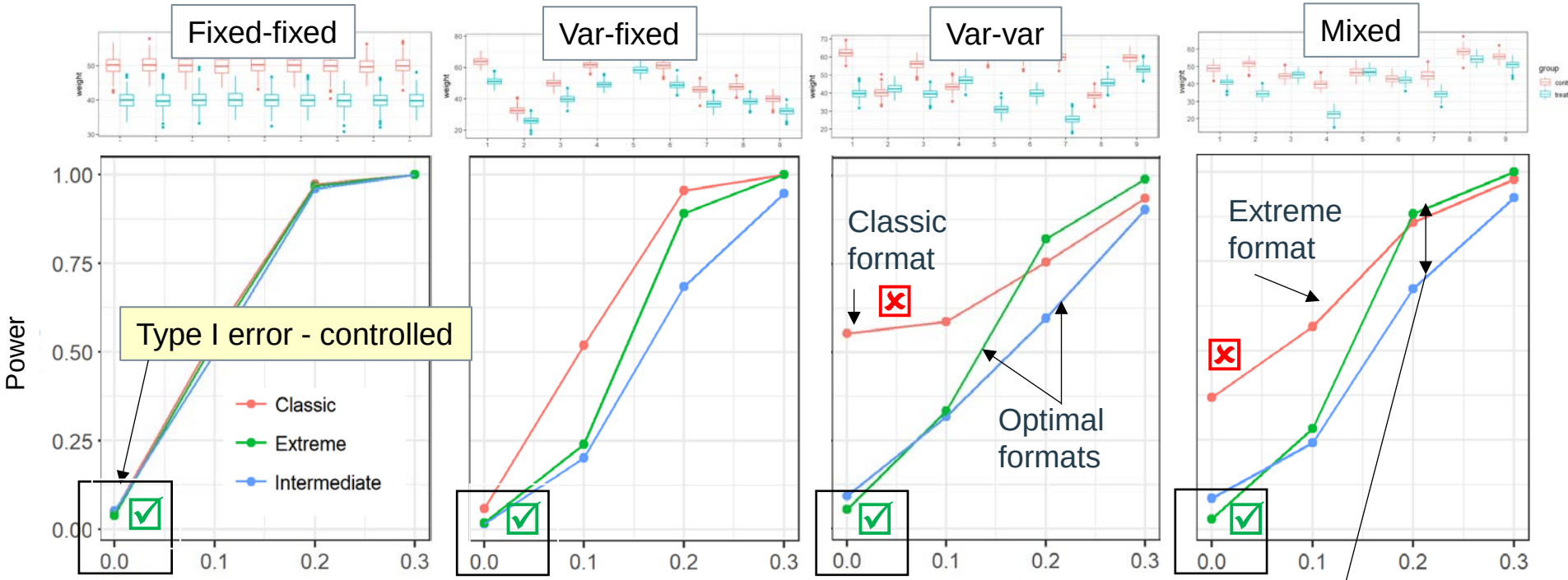
  - $\delta_{ij} \sim N(0, {\sigma_{study}}^2)$, random effect due to the jth study, Same for both groups

- Extreme design

Model: $Y_{ij} = \mu + \Delta_i + \varepsilon_{ij}$



Tested assay formats: performance comparison



Optimal formats allow to control for Type I error in all cases!

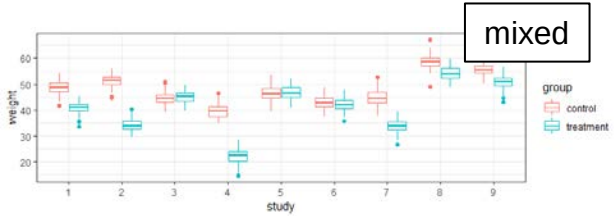
Extreme format has the best discriminative ability

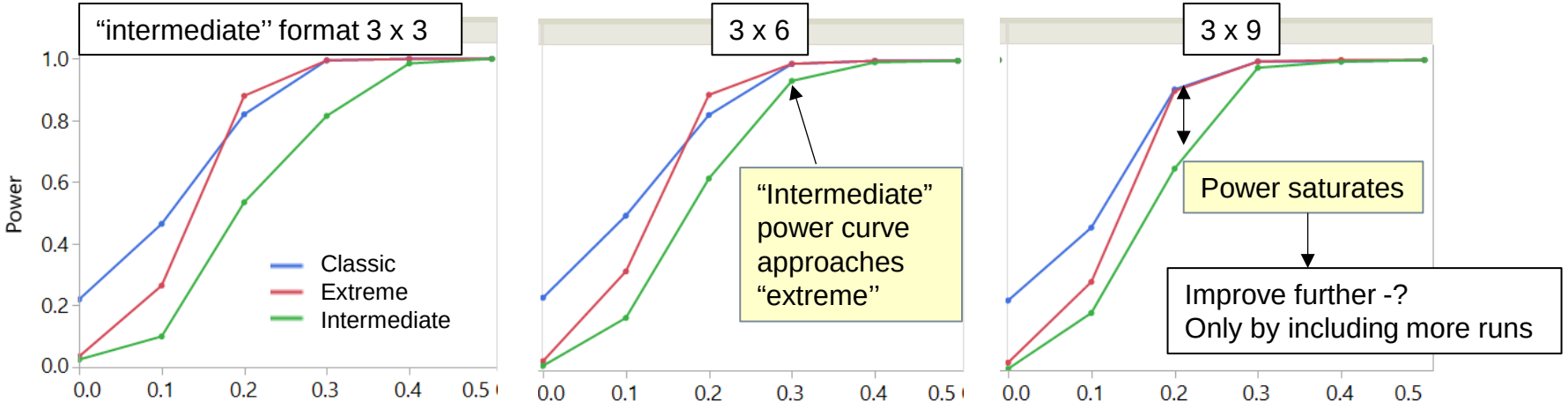
Proposition: leverage recommendations from manufacturing

	Pre-clinical	Manufacturing
Regulation	Poorly regulated (GRP)	Very well regulated (GMP, <USP>, …)
Objective	Control Type 1 error, Maximize power	Minimize the error on the estimate of the mean
Current practice	Experiments conducted **under homogenous conditions**	Experiments conducted under conditions of largest variance • Convert bias into lack of precision • Modulate precision by adding extra samples
Proposal	Adopt this practice from manufacturing	

PF: Application of USP approach in-silico

- Study-study variability - **‘mixed’** scenario
- <USP> approach:
  - **Convert bias into lack of precision**
  - **Modulate precision** – add more subjects per run to achieve “extreme” case performance





1. The power of “intermediate” assay improves
2. Type 1 error is not impacted
3. The power saturates quickly with addition of extra samples

Conclusions

- Current practice that avoids study-study variability leads to risk of high Type 1 error (false discovery rate)
- Good design of experiment practice recommends to spread experiments across highest sources of variability
- Optimal format of experiments allows to protect against false discovery rate in all cases, i.e. with or without important study-study variability
- Recommendations existing in manufacturing area since more than a decade – should be applied in preclinical pharmacology

Contacts

- **Milana Filatenkova** (Manager Statistics)
- milana.filatenkova@pharmalex.com

- Bruno Boulanger (CSO)
- bruno.boulanger@pharmalex.com

- PharmaLex Belgium
- 5 Rue Edouard Belin
- 1435 Mont-Saint-Guibert
- Belgium

Acknowledgements

- Anne-Cécile Kerangueven
- Mathilde Vidal

Follow us on social media

/pharmalex-gmbh

@PharmalexGlobal