

# Handling Nested Data in (Un)balanced Designs

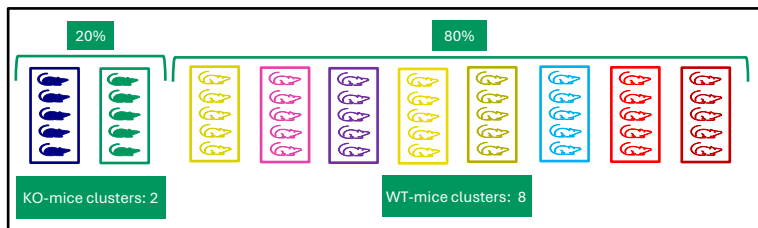


van Zyl, H.<sup>1</sup> Dolan, C.V.<sup>2</sup>, van der Sluis, S.<sup>1</sup>

<sup>1</sup>Department of Complex Trait Genetics, Amsterdam UMC/Vrije Universiteit Amsterdam; <sup>2</sup>Department of Biological Psychology, Vrije Universiteit Amsterdam

## Introduction

- Neuroscientific studies often implicate nested, unbalanced data designs (*Figure 1*)
- Such “clustering” compromises the validity of conclusions drawn from common statistical methods (standard *t*-test, ANOVA)
- Alternative, more appropriate clustering methods include:
  - ~ summary statistics *t*-test (SST)<sup>a</sup>
  - ~ linear mixed-effects models (LMEMs)<sup>b,c,d</sup>
- Above-mentioned two methods are only known to function well in balanced settings
  - ~ equal numbers of (observations per) clusters per group
- Aim: investigate the functioning of different clustering methods in nested, unbalanced (normally-distributed) data



*Figure 1: Repeatedly-recorded neuron measurements are nested within a KO/WT mouse (cluster). Each mouse icon represents one neuron measurement. In this scenario, the simulation settings (excluding ICC) are as follows: 10 total clusters, 20vs80 unbalancedness, and a cluster size of 5.*

## Methods

Normally-distributed data simulated in 400 settings (*Table 1*)

- Number of simulations = 1000 per setting
- Analyzed using:
  - ~ SST
  - ~ LMEMs (estimation: Maximum Likelihood/ML, Restricted ML/REML)
  - ~ Generalized Estimating Equations (GEEs) (correlation structures: autoregressive/AR, exchangeable/Ex, independent/Ind)
- Evaluation: Type 1 error rate

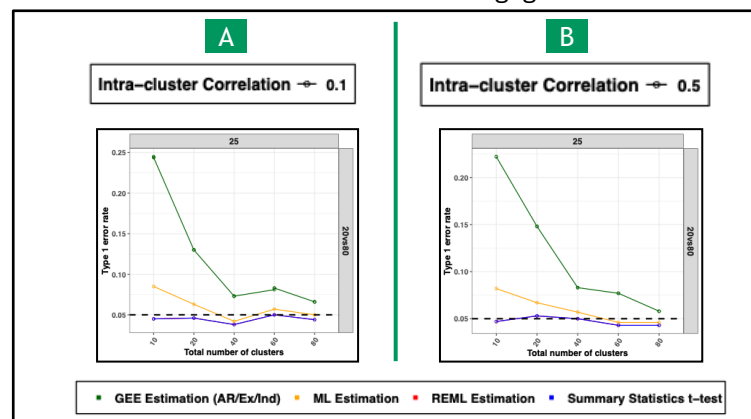
Total clusters	Unbalancedness (cluster level) ~ as percentages (%)	Cluster size (number of measurements)	ICC
10; 20; 40; 60; 80	50vs50; 40vs60; 30vs70; 20vs80	5 to 95 (in increments of 10)	0.1; 0.5

*Table 1: Simulation settings.*

## Results/Conclusion

Across all setting-scenarios (exemplified in *Figure 2*), Type 1 error rates are:

- very high for GEE models (all correlation structures)
- generally inflated for ML-LMEMs
- are at an acceptable level for SST and REML-LMEMs
  - ~ difference between two methods negligible



*Figure 2: Assuming 20vs80 unbalancedness, the change in Type 1 error rate is depicted for a cluster size of 25 as the number of clusters increases (10 to 80) with an ICC of 0.1 (A) and 0.5 (B).*

## Future directions

- Extending to non-normal data

For neuroscientists:

- General guideline on study design and data collection
- Pipeline statistical-analysis tool for (non)normal, nested data in (un)balanced designs

## References

- <sup>a</sup>McNabb et al., (2021), doi: 10.1016/j.crneur.2021.100024.
- <sup>b</sup>Aarts E et al., (2014), doi: 10.1038/nn.3648.
- <sup>c</sup>Aarts E et al., (2015), doi: 10.1186/s12868-015-0228-5.
- <sup>d</sup>Yu Z et al., (2022), doi: 10.1016/j.neuron.2021.10.030.