Multi-center trials, heterogeneity of study samples and external validity

Bernhard Voelkl

University of Bern, Animal Welfare Division
Inference about Population Parameters

Central Limit Theorem:
Given sufficiently large random samples from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.
Inference about Population Parameters

Central Limit Theorem:
Given sufficiently large **random** samples from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.
**Bias:** where does it come from?

**Genotype**

**Environment**

**Genotype × Environment**
Bias: where does it come from?

Genotype

Environment

Genotype × Environment

Population: «Laboratory Mice»
**G × E Interactions:** Lab differences despite standardization

Genetics of Mouse Behavior: Interactions with Laboratory Environment


**Standardization:**
«*We went to extraordinary lengths to equate test apparatus, testing protocols, and all possible features of animal husbandry*»

Variables explicitly equated across laboratories included: apparatus, exact testing protocols, age of shipped and laboratory-reared mice, method and time of marking before testing, food, bedding, stainless steel cage tops, four to five mice per cage, light/dark cycle, cage changing frequency and specific days, male left in cage after births, culling only of obvious runts, postpartum pregnancy allowed, weaned at 21 days, specific days of body weight recording, and gloved handling without use of forceps. Unmatched variables included local tap water, requirement of filters over cage tops in Portland only, variation of physical arrangement of colonies and testing rooms across sites, different air handling and humidity, and different sources of batches of cocaine and alcohol.
**G × E Interactions:** Lab differences despite standardization

Genetics of Mouse Behavior: Interactions with Laboratory Environment

**G × E Interactions:** Lab differences despite standardization

Genetics of Mouse Behavior: Interactions with Laboratory Environment

G x E Interactions: Lab differences despite standardization

Genetics of Mouse Behavior: Interactions with Laboratory Environment

**G × E Interactions:** Lab differences despite standardization

Genetics of Mouse Behavior: Interactions with Laboratory Environment
Bias: where does it come from?

Genotype

Environment

Genotype $\times$ Environment

Reaction Norm:

*Function describing the relationship between a specific environmental parameter and the expected value for a specific trait for one genotype.*
**Multi-Laboratory Studies:** Incorporating environmental variation

**Norm of Reaction**

- **Trait**
- **Environmental parameter X**
- Lab0
- N=20
Multi-Laboratory Studies: Incorporating environmental variation

Norm of Reaction

Trait

Lab4
Lab1
Lab3
Lab2
Lab5

total N=20

Environmental parameter X
Multi-Laboratory Studies: Incorporating environmental variation

Norm of Reaction

Trait

Environmental parameter X

Lab10
Lab7
Lab8
Lab9
Lab6

total N=20
Multi-Laboratory Studies: Incorporating environmental variation

Norm of Reaction

Environmental parameter $X$

Trait

Lab0

Lab15

$N=20$
Multi-Laboratory Studies: how much can we gain?

Simulated sampling from real-world data:

<table>
<thead>
<tr>
<th>Set</th>
<th>Treatment</th>
<th>Measure</th>
<th># Studies</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>tPA</td>
<td>Infarct Volume</td>
<td>57</td>
<td>Stroke</td>
</tr>
<tr>
<td>2</td>
<td>Trastuzumab</td>
<td>RX since Xmm3</td>
<td>58</td>
<td>Breast Cancer</td>
</tr>
<tr>
<td>3</td>
<td>FK506</td>
<td>Infarct Volume</td>
<td>31</td>
<td>Stroke</td>
</tr>
<tr>
<td>4</td>
<td>Rosiglitazone2</td>
<td>Infarct Volume</td>
<td>21</td>
<td>Stroke</td>
</tr>
<tr>
<td>5</td>
<td>IL1-RA</td>
<td>Infarct Volume</td>
<td>36</td>
<td>Stroke</td>
</tr>
<tr>
<td>6</td>
<td>Hypothermia</td>
<td>Infarct Volume</td>
<td>98</td>
<td>Stroke</td>
</tr>
<tr>
<td>7</td>
<td>Cardiosphere DC</td>
<td>EF (%)</td>
<td>35</td>
<td>Myocard Infarct</td>
</tr>
<tr>
<td>8</td>
<td>Tirilazad</td>
<td>Infarct Volume</td>
<td>17</td>
<td>Stroke</td>
</tr>
<tr>
<td>9</td>
<td>Estradiol</td>
<td>Infarct Volume</td>
<td>24</td>
<td>Stroke</td>
</tr>
<tr>
<td>10</td>
<td>Human MSC</td>
<td>Infarct Volume</td>
<td>26</td>
<td>Stroke</td>
</tr>
<tr>
<td>11</td>
<td>MK801</td>
<td>Infarct Volume</td>
<td>30</td>
<td>Stroke</td>
</tr>
<tr>
<td>12</td>
<td>TMZ</td>
<td>Volume</td>
<td>26</td>
<td>Glioma</td>
</tr>
<tr>
<td>13</td>
<td>Ckit CSC</td>
<td>EF (%)</td>
<td>25</td>
<td>Myocard Infarct</td>
</tr>
<tr>
<td>14</td>
<td>Rat BMSC</td>
<td>Infarct Volume</td>
<td>25</td>
<td>Stroke</td>
</tr>
</tbody>
</table>
Multi-Laboratory Studies: how much can we gain?

Simulated sampling from real-world data:

Reproducibility of preclinical animal research improves with heterogeneity of study samples

Bernhard Voelkl¹, Lucille Vogt¹, Emily S. Sena², Hann o Würbel ¹*

¹ Division of Animal Welfare, VPH Institute, Vetsuisse Faculty, University of Bern, Bern, Switzerland
² Centre for Clinical Brain Sciences, Chancellor's Building, University of Edinburgh, Edinburgh, United Kingdom

* hanno.wuerbel@vetsuisse.unibe.ch

Abstract

Single-laboratory studies conducted under highly standardized conditions are the gold standard in preclinical animal research. Using simulations based on 400 preclinical studies across 13 different interventions in animal models of stroke, myocardial infarction, and breast cancer, we compared the accuracy of effect size estimates between single-laboratory and multi-laboratory study designs. Single-laboratory studies generally failed to predict effect size accurately, and larger sample sizes rendered effect size estimates even less accurate. By contrast, multi-laboratory designs including as few as 2 to 4 laboratories increased coverage probability by up to 12 percentage points without a need for larger sample sizes. These findings demonstrate that within-study standardization is a major cause of poor reproducibility. More representative study samples are required to improve the external validity and reproducibility of preclinical animal research and to prevent wasting animals and resources for inconclusive research.

Multi-Laboratory Studies: how much can we gain?

Result:

Multi-laboratory designs including as few as 2 to 4 laboratories increased coverage probability by up to 42 percentage points without a need for larger sample sizes.
The Standardization Fallacy

Resultst of highly standardized studies are less likely to be reproduced than results of heterogeneous (diversified) studies.

«Excessive standardization of every aspect of the testing environment will lead to idiosyncratic results that cannot be reproduced under even slightly different conditions.»

Multi-center trials, heterogeneity of study samples and external validity

Conclusion:

Highly standardized single-laboratory studies are a source of poor reproducibility because they ignore biologically meaningful variation.

Multi-laboratory studies (and other ways of creating more heterogeneous study samples) provide an effective means of improving the reproducibility of study results.

This is crucial to prevent wasting animals and resources for inconclusive research.