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

Bernhard Voelkl

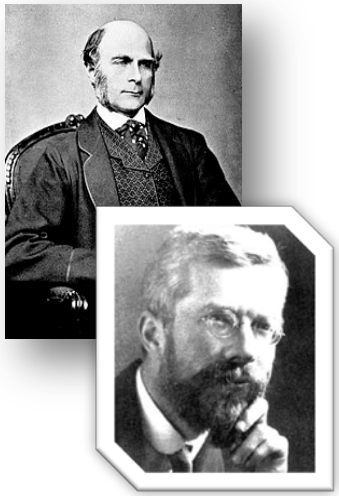
University of Bern, Animal Welfare Division

EQIPD

imi innovative medicines initiative

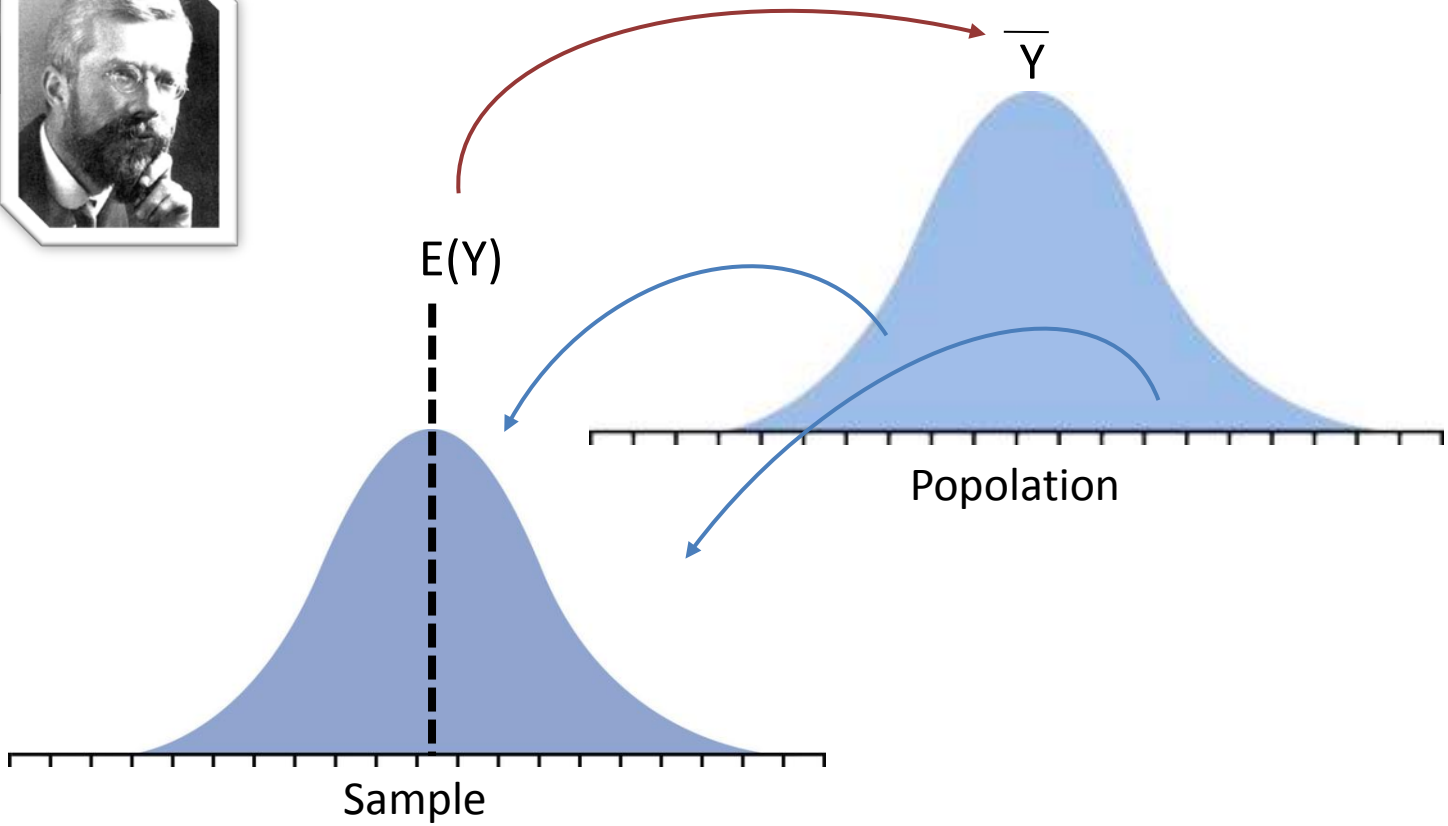
3R Reduce Refine Replace

Inference about Population Parameters

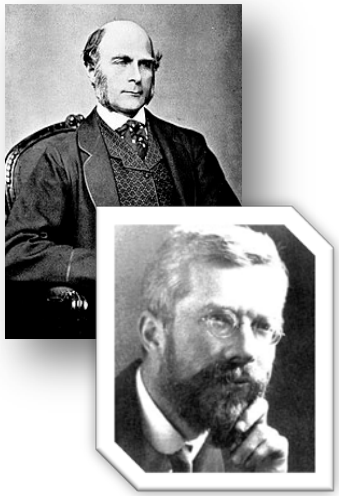


Central Limit Theorem:

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.

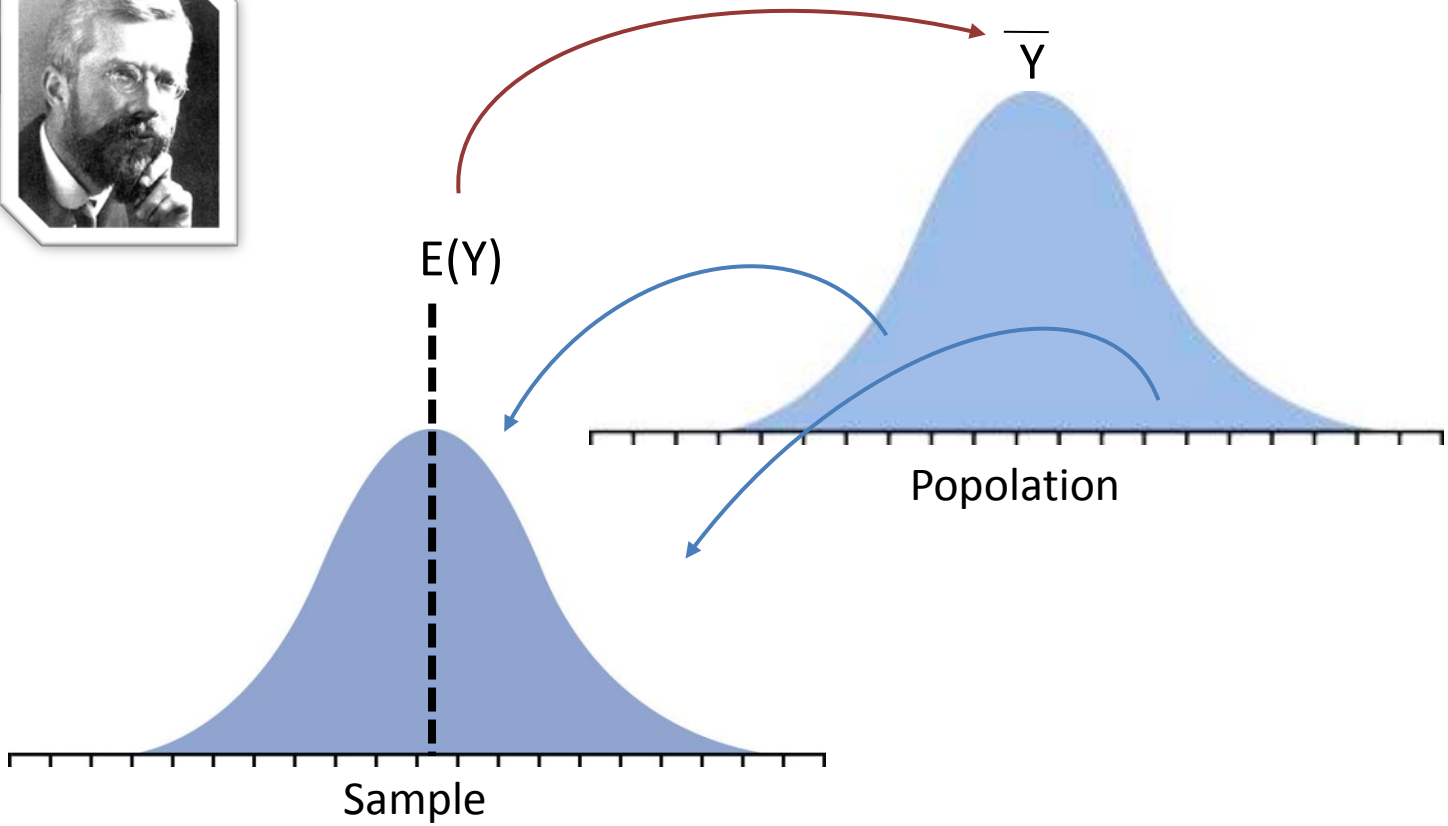


Inference about Population Parameters



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.

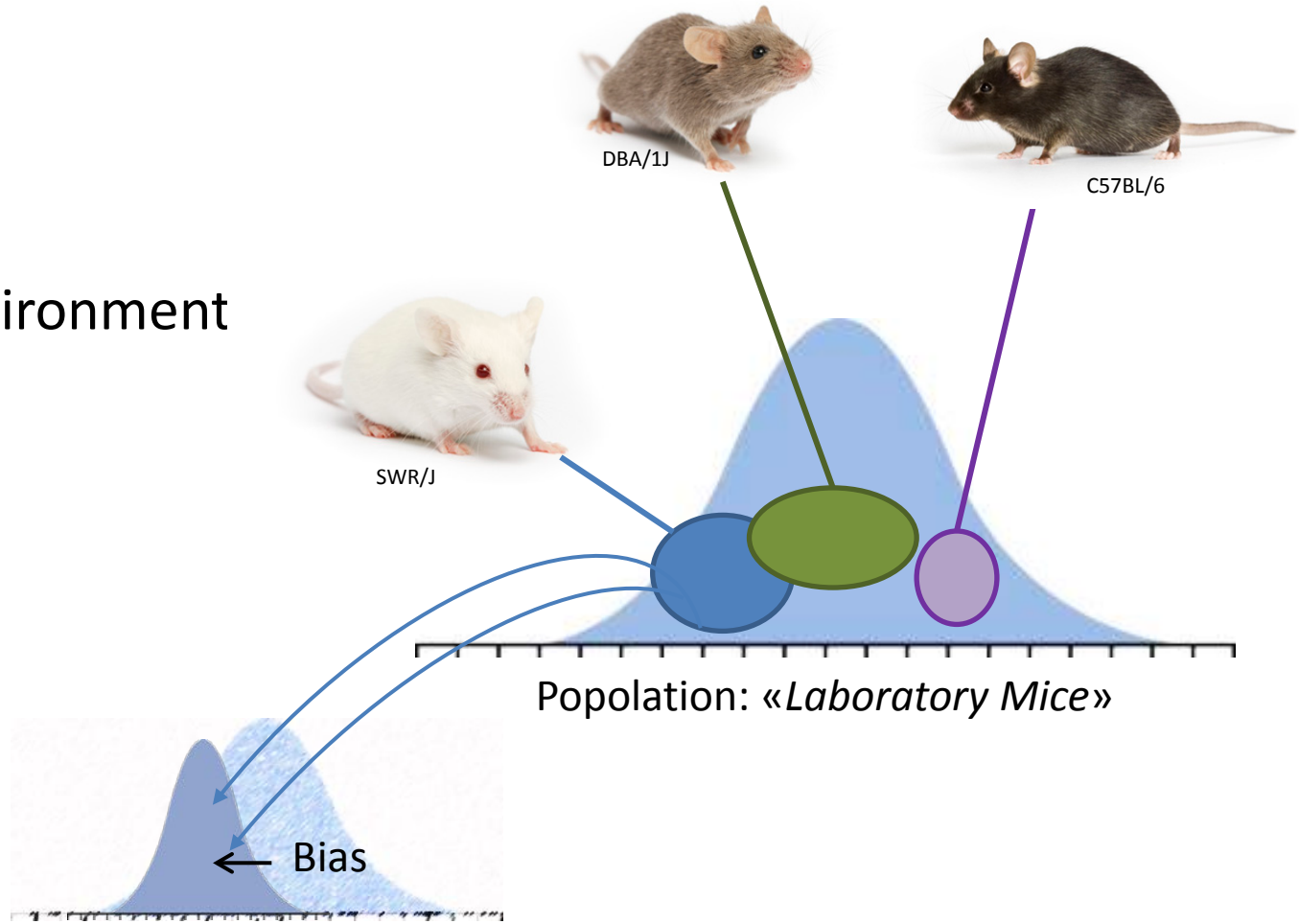


Bias: where does it come from?

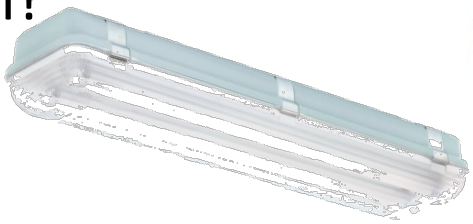
Genotype

Environment

Genotype × Environment



Bias: where does it come from?



Genotype

Environment

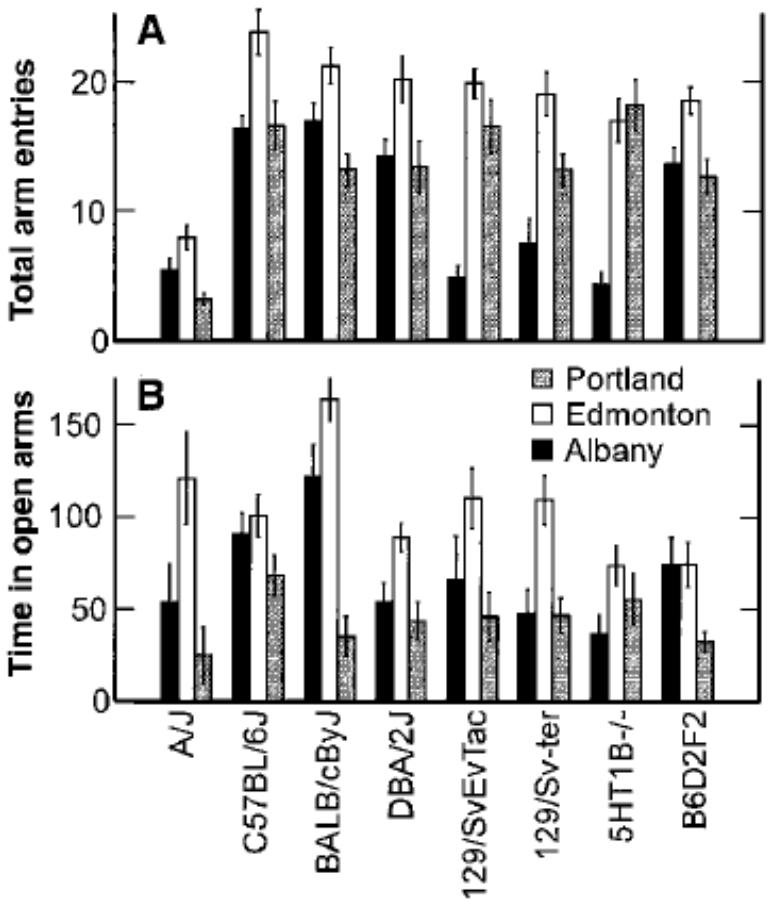
Genotype × Environment

Popolation: «*Laboratory Mice*»

G × E Interactions: Lab differences despite standardization

Genetics of Mouse Behavior: Interactions with Laboratory Environment

Crabbe et al. (1999): *Science*, **284**, 1670-1672.



Standardization:

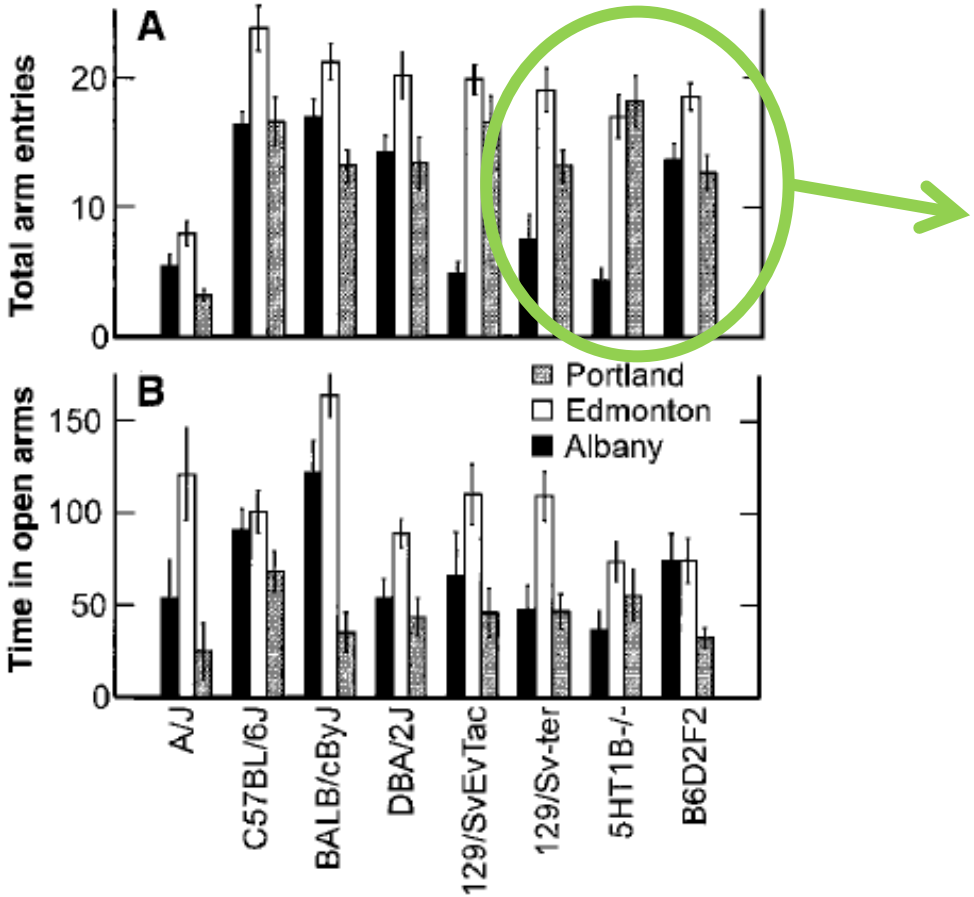
«*We went to extraordinary lengths to equate test apparatus, testing protocols, and all possible features of animal husbandry*»

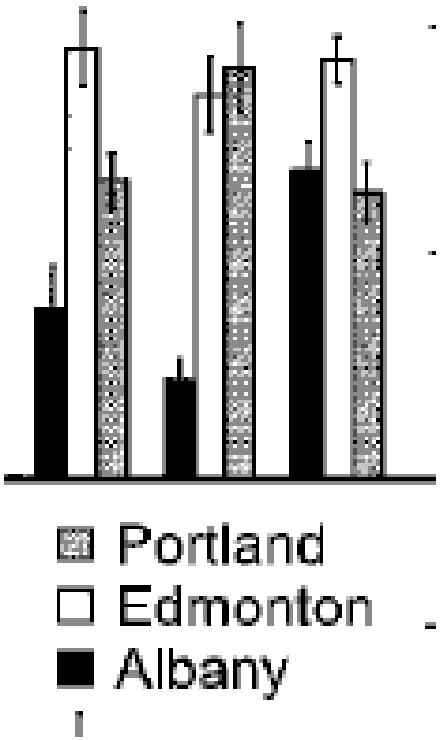
Variables explicitly equated across laboratories included: apparatus, exact testing protocols, age of shipped and laboratory-reared mice, method and time of marking before testing, food, bedding, stainless steel cage tops, four to five mice per cage, light/dark cycle, cage changing frequency and specific 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

Crabbe et al. (1999): *Science*, **284**, 1670-1672.

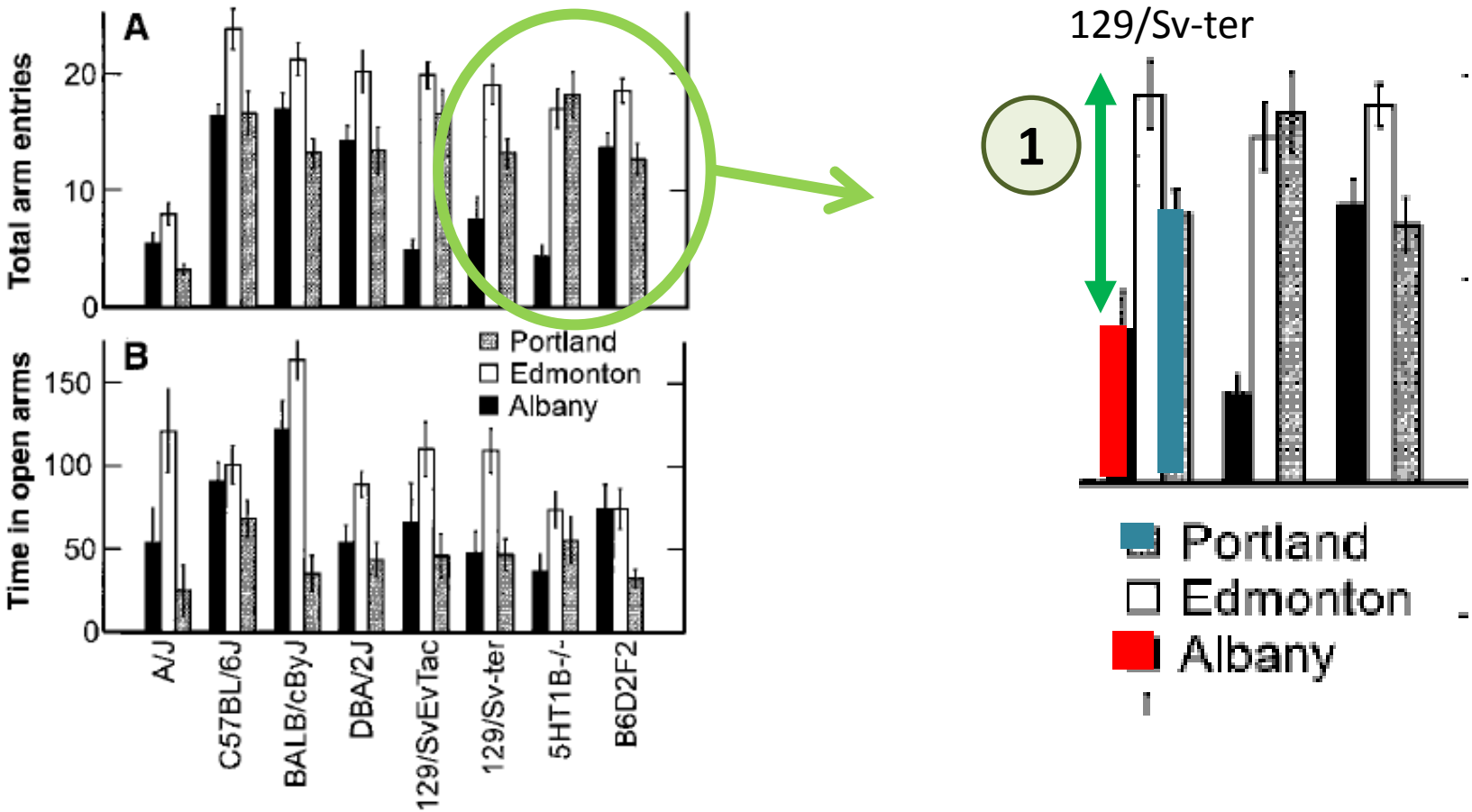




G × E Interactions: Lab differences despite standardization

Genetics of Mouse Behavior: Interactions with Laboratory Environment

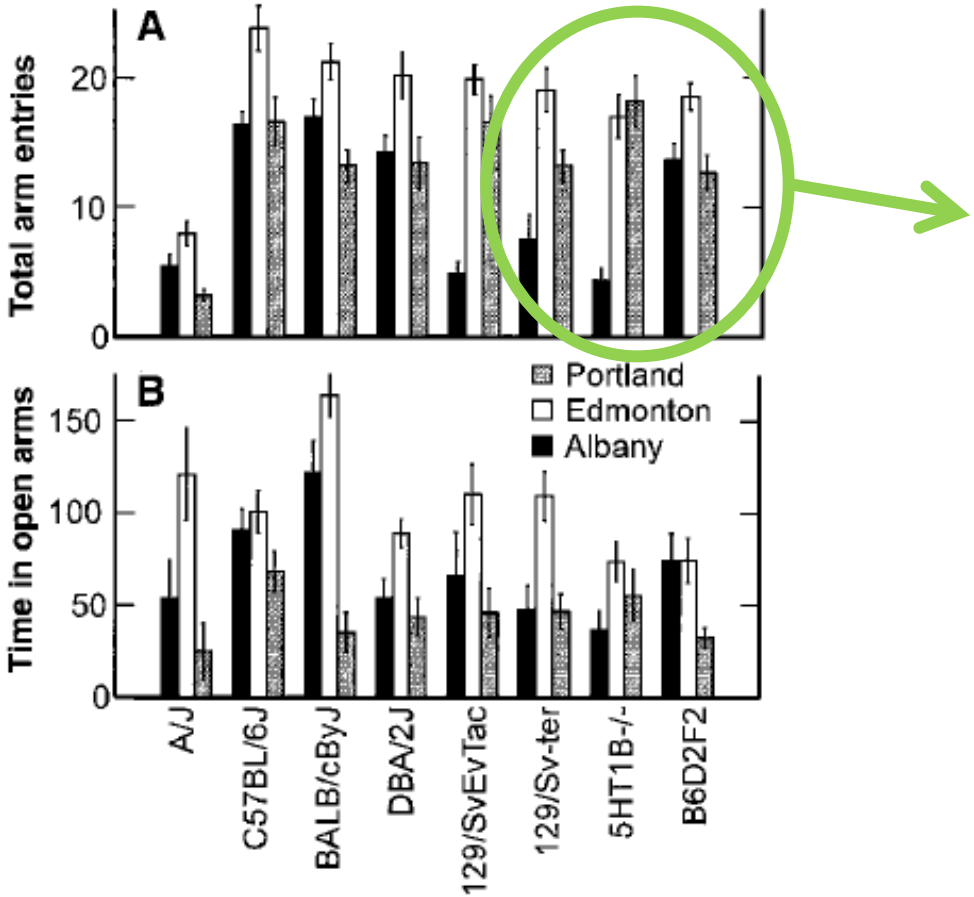
Crabbe et al. (1999): *Science*, **284**, 1670-1672.

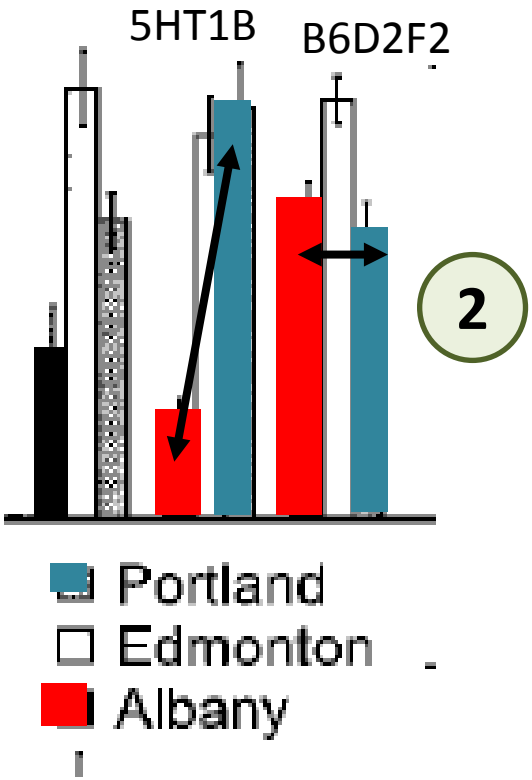


G × E Interactions: Lab differences despite standardization

Genetics of Mouse Behavior: Interactions with Laboratory Environment

Crabbe et al. (1999): *Science*, **284**, 1670-1672.

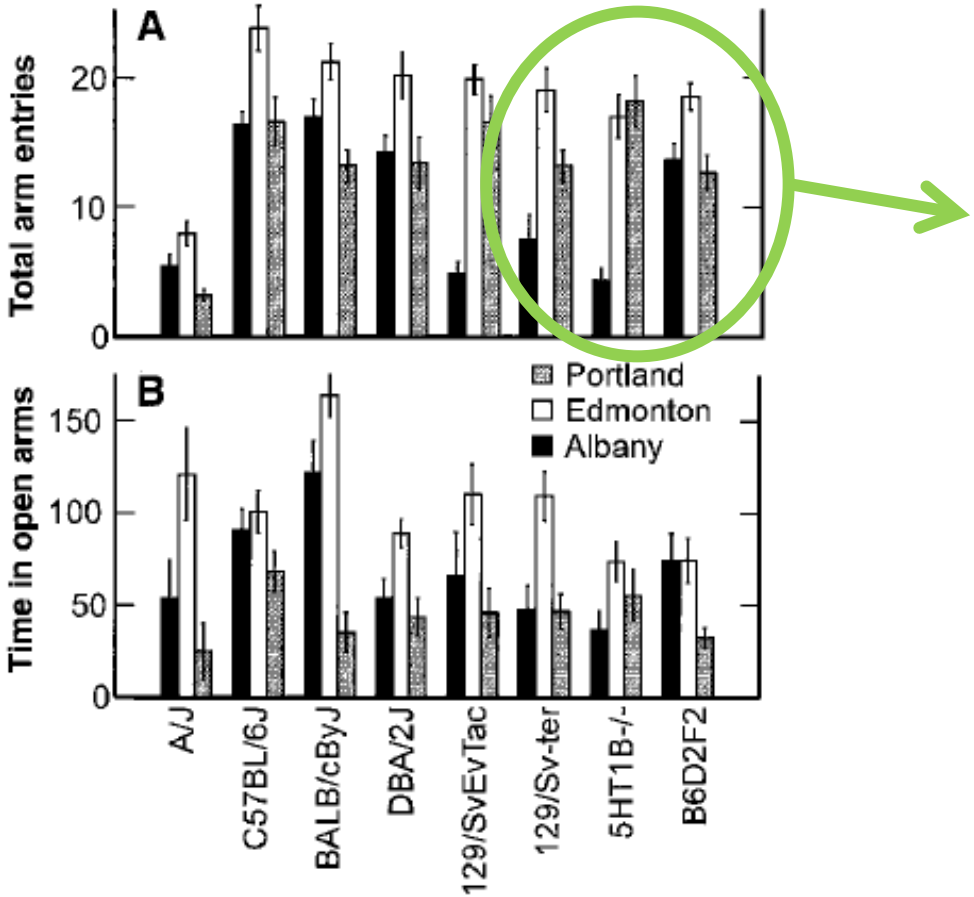


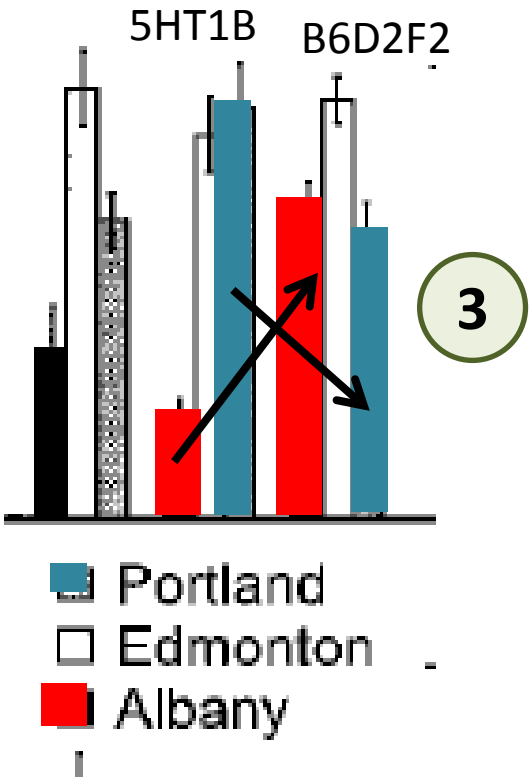


G × E Interactions: Lab differences despite standardization

Genetics of Mouse Behavior: Interactions with Laboratory Environment

Crabbe et al. (1999): *Science*, **284**, 1670-1672.



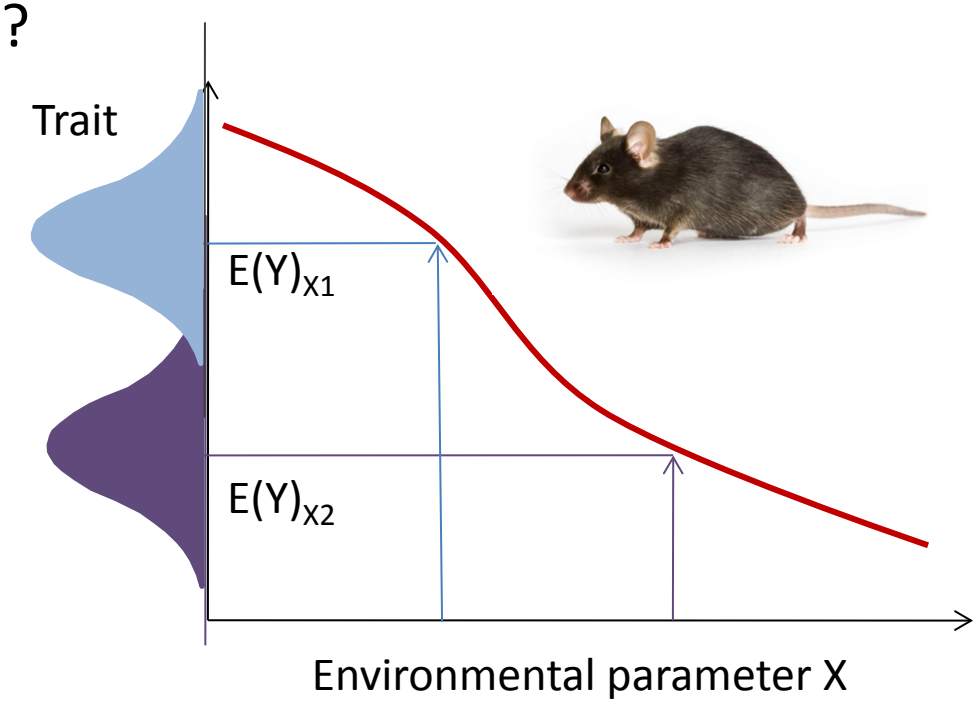


Bias: where does it come from?

Genotype

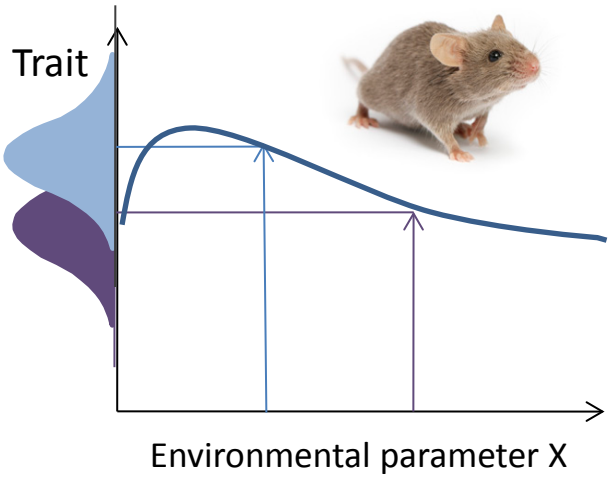
Environment

Genotype × Environment



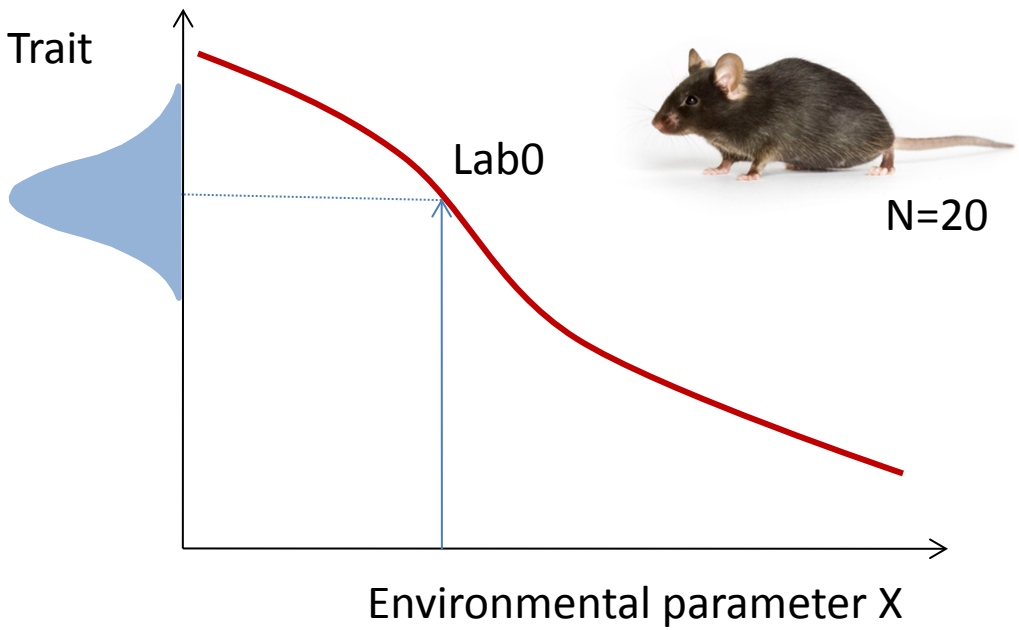
Reaction Norm:

Function describing the relationship between a specific environmental parameter and the expected value for a specific trait for one genotype.



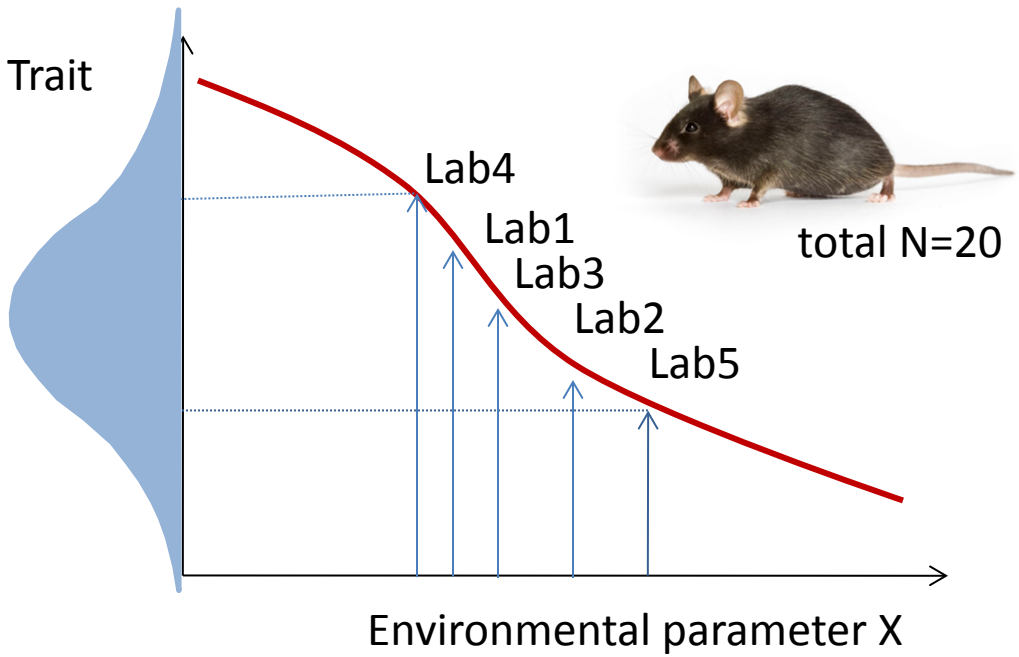
Multi-Laboratory Studies: Incorporating environmental variation

Norm of Reaction

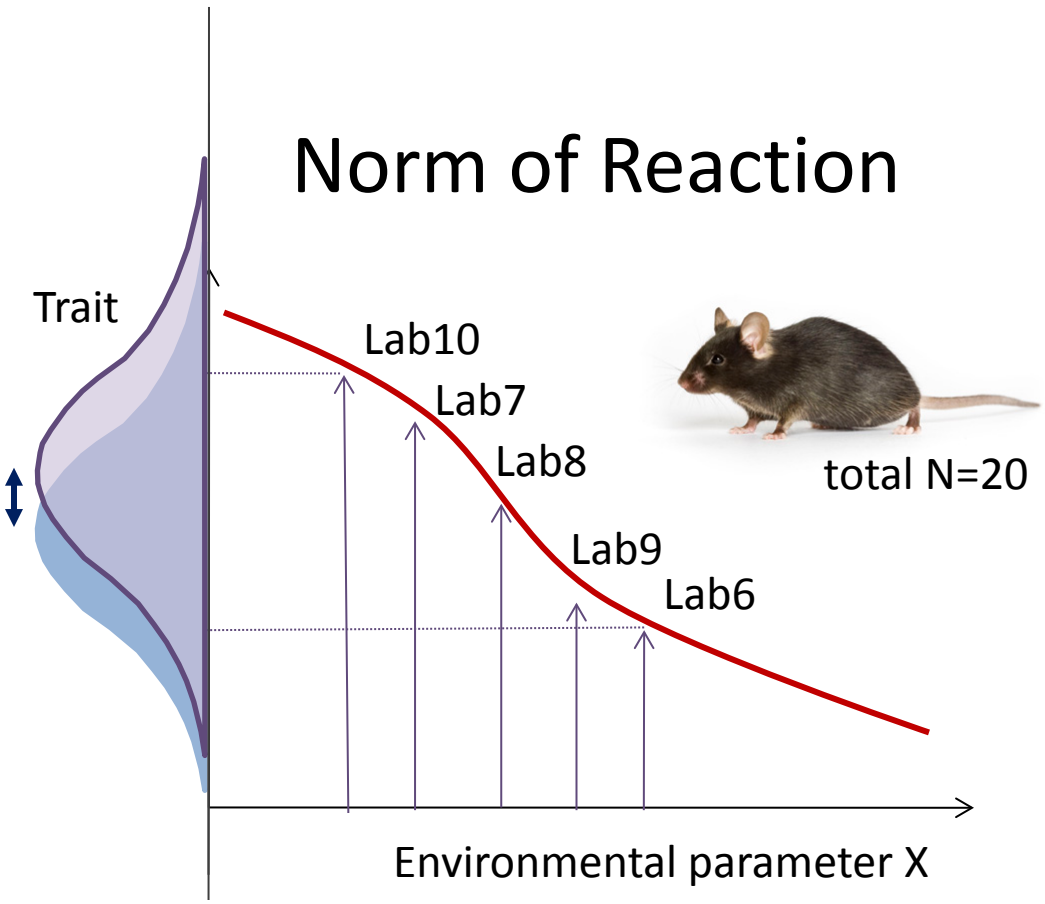


Multi-Laboratory Studies: Incorporating environmental variation

Norm of Reaction

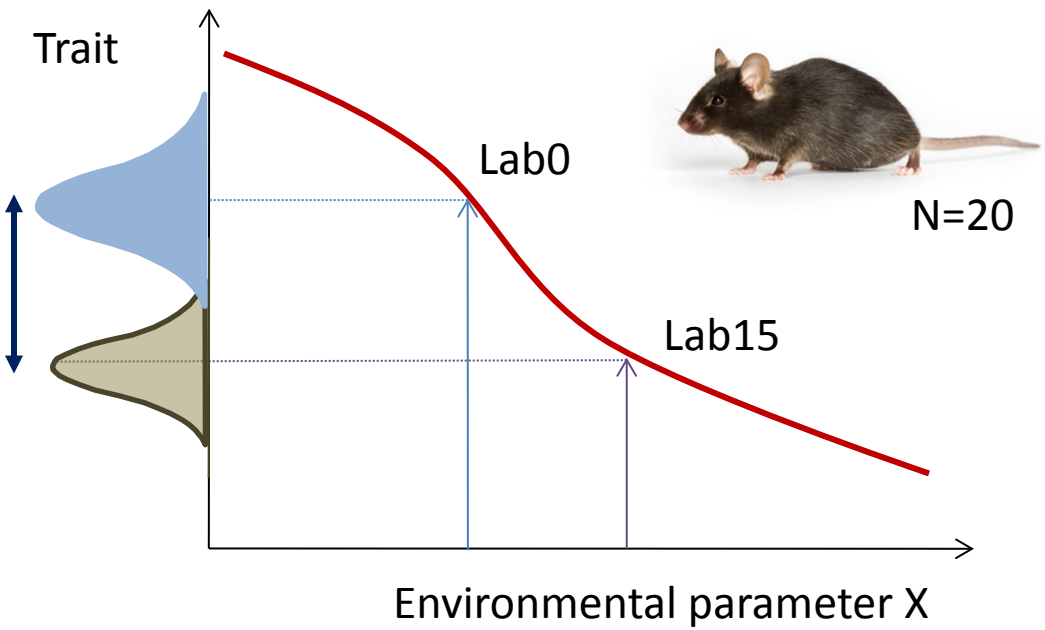


Multi-Laboratory Studies: Incorporating environmental variation



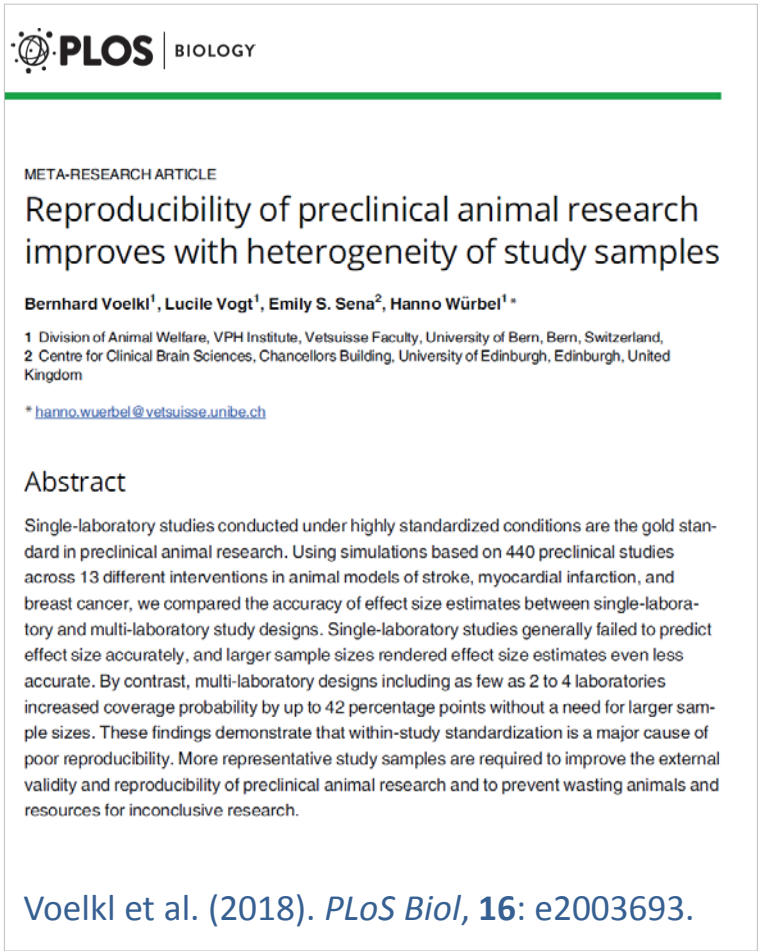
Multi-Laboratory Studies: Incorporating environmental variation

Norm of Reaction



Multi-Laboratory Studies: how much can we gain?

Simulated sampling from real-world data:



PLOS | BIOLOGY

META-RESEARCH ARTICLE

Reproducibility of preclinical animal research improves with heterogeneity of study samples

Bernhard Voelkl[1], Lucile Vogt[1], Emily S. Sena[2], Hanno Würbel[1]*

1 Division of Animal Welfare, VPH Institute, Vetsuisse Faculty, University of Bern, Bern, Switzerland,
2 Centre for Clinical Brain Sciences, Chancellors 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 440 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 42 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.

Voelkl et al. (2018). *PLoS Biol*, **16**: e2003693.

Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Studies

·C·A·M·A·R·A·D·E·S·

Set	Treatment	Measure	# Studies	Disease
1	tPA	Infarct Volume	**57**	Stroke
2	Trastuzumab	RX since Xmm3	**58**	Breast Cancer
3	FK506	Infarct Volume	**31**	Stroke
4	Rosiglitazone2	Infarct Volume	**21**	Stroke
5	IL1-RA	Infarct Volume	**36**	Stroke
6	Hypothermia	Infarct Volume	**98**	Stroke
7	Cardiosphere DC	EF (%)	**35**	Myocard Infarct
8	Tirilazad	Infarct Volume	**17**	Stroke
9	Estradiol	Infarct Volume	**24**	Stroke
10	Human MSC	Infarct Volume	**26**	Stroke
11	MK801	Infarct Volume	**30**	Stroke
12	TMZ	Volume	**26**	Glioma
13	Ckit CSC	EF (%)	**25**	Myocard Infarct
14	Rat BMSC	Infarct Volume	**25**	Stroke

Multi-Laboratory Studies: how much can we gain?

Simulated sampling from real-world data:

PLOS | BIOLOGY

META-RESEARCH ARTICLE

Reproducibility of preclinical animal research improves with heterogeneity of study samples

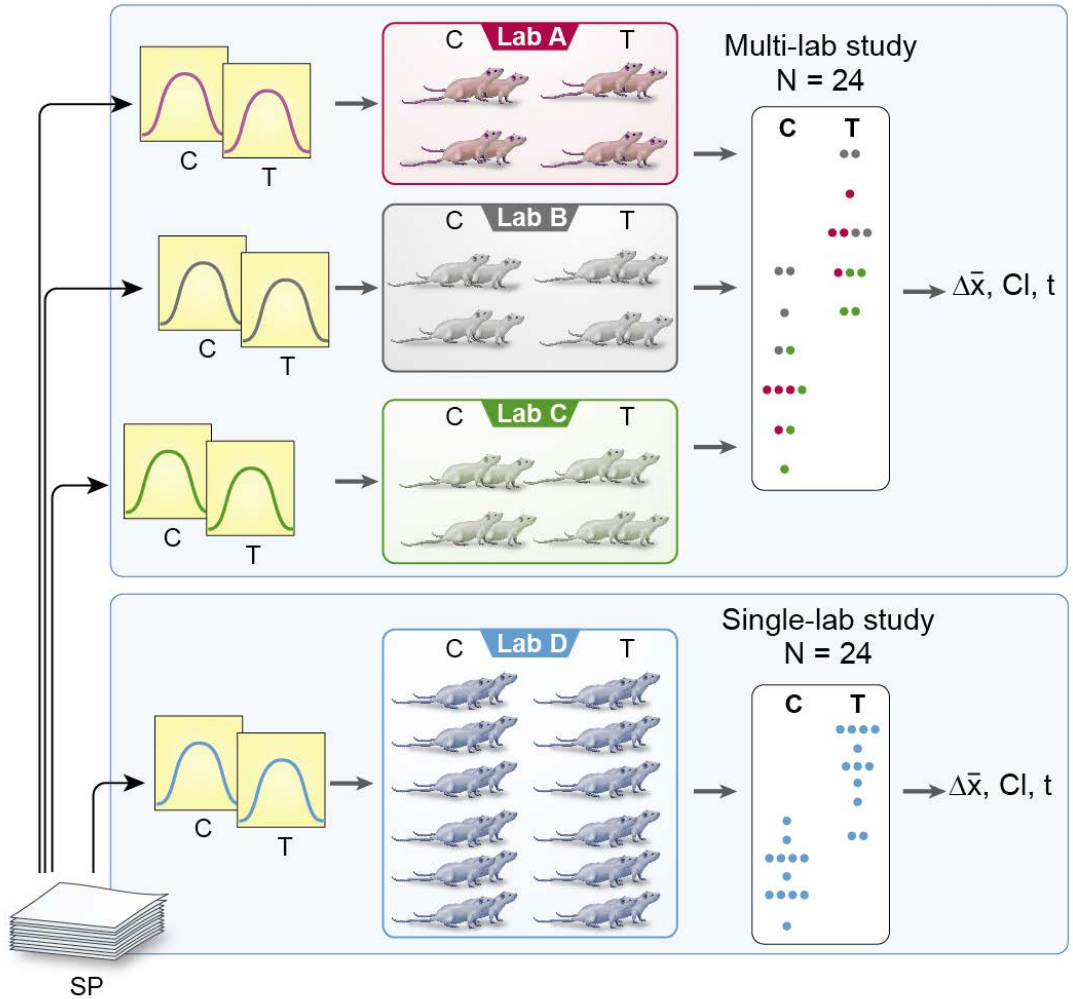
Bernhard Voelkl[1], Lucile Vogt[1], Emily S. Sena[2], Hanno Würbel[1]*

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

* hanno.wuerbel@vetsuisse.unibe.ch

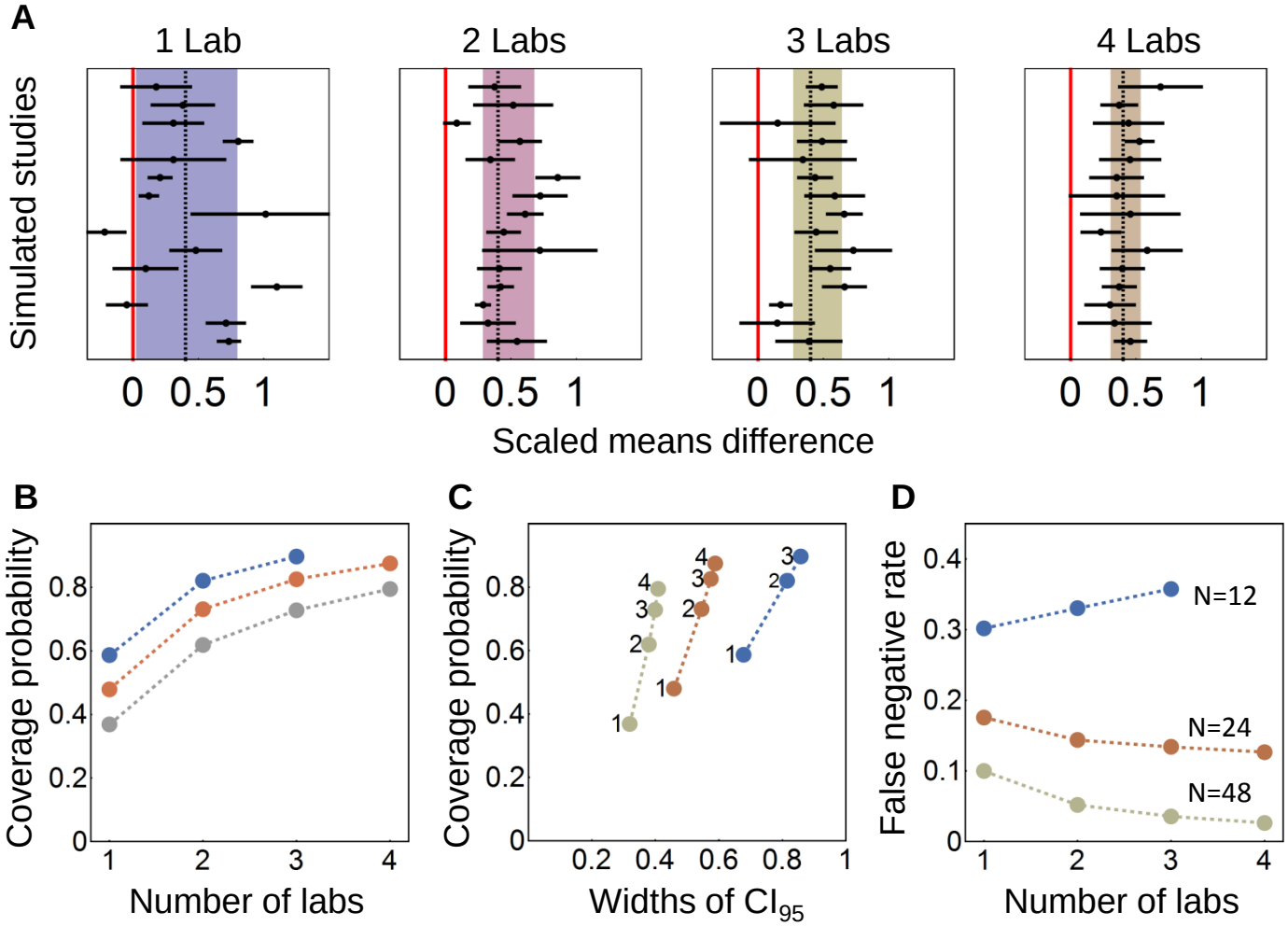
Abstract

Single-laboratory studies conducted under highly standardized conditions are the gold standard in preclinical animal research. Using simulations based on 440 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 42 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.



Voelkl et al. (2018). *PLoS Biol*, **16**: e2003693.

Multi-Laboratory Studies: how much can we gain?



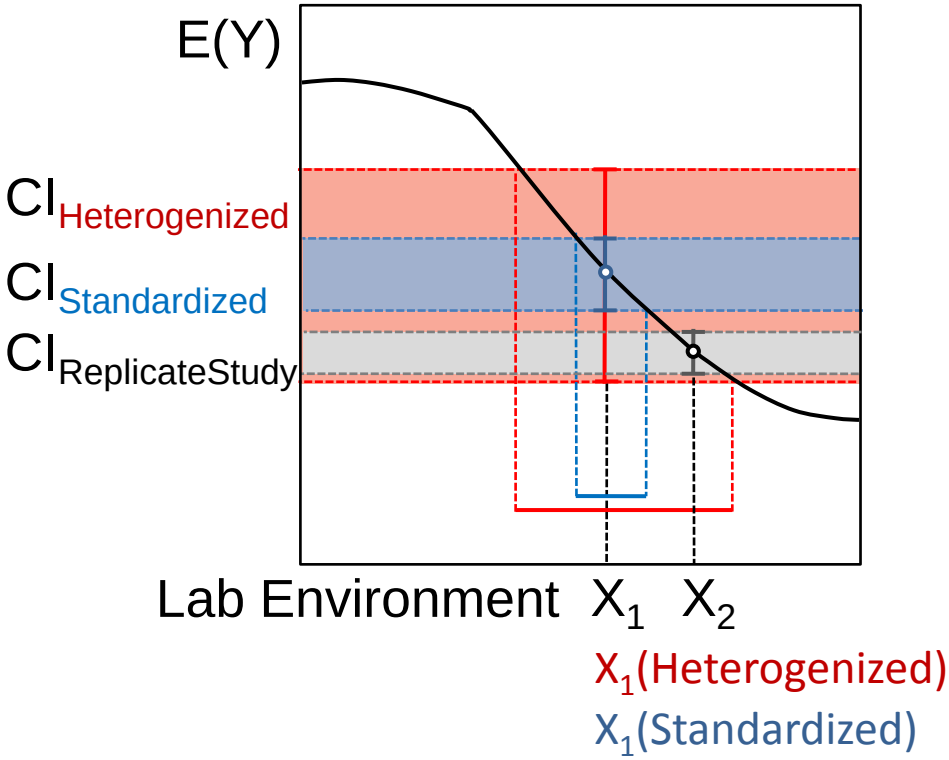
Result:

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.*»



Voelkl & Würbel (2016): *Trends Pharmacol. Sci.* **37**, 509-510.

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.

EQIPD imi innovative medicines initiative