A priori Estimation of Reproducibility Odds Informs the Sizing of Omic Data Cohorts

JOBIM 2025 08>11 JUIL Bordeaux

M. Colange[1], A. Nordor[1], and A. Behdenna[1].

[1]Epigene Labs, Paris, France

Introduction

- **Fragmented Data Hinders Progress:** Omic studies are constrained by fragment or poorly integrated datasets, weakening generalizability and reproducibility.
- **Data Integration is Under-Resourced:** Data integration is typically approached on a "best effort" basis, with little guidance on how much effort or resources are truly needed.
- **Costs Remain Invisible:** The scientific and opportunity costs of limited integration are widely acknowledged but rarely quantified in a rigorous and systematic way.

Contribution

- **Quantifying What's at Stake:** We introduce mathematical formulas that link cohort size and statistical power, making the cost of limited integration explicit.
- **Practical Tools for Study Design:** These ready-to-use formulas apply broadly across data types and can inform both new and secondary analyses.
- **Enabling Strategic Commitment:** By revealing the price of underpowered studies, we aim to shift data integration from an ad-hoc task to a justified, well-resourced priority.

Quantifying the gap between observations and the hidden signal

Law of large numbers

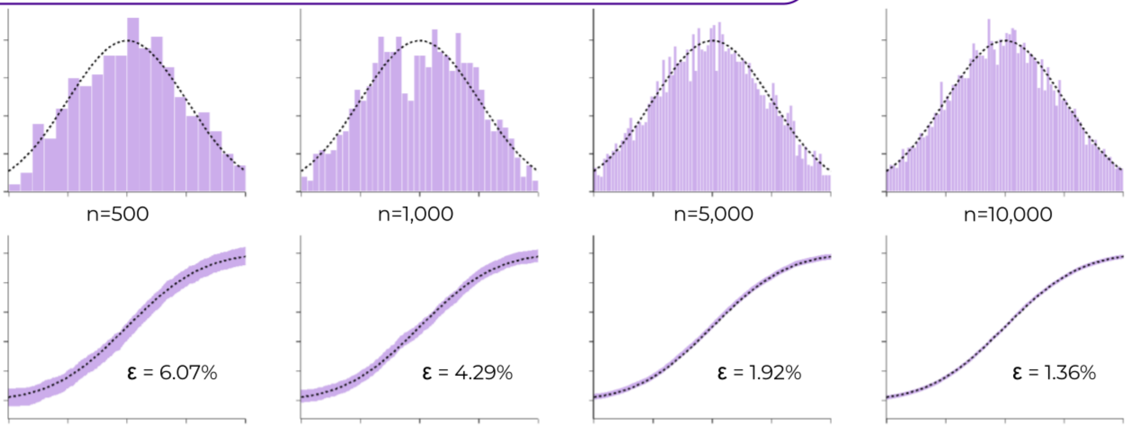
The distribution of observations converges to the true distribution.

"the more samples, the better"

Dvoretzky-Kiefer-Wolfowitz bound

Gives a rate of convergence for the law of large numbers.

"n samples needed for a 95% CI on the signal distribution narrower than ε."



Linking CI width, confidence level and cohort size

confidence level on the approximation of the signal distribution

number of observations

$$1 - \alpha \leq 2ke^{-2n\varepsilon^2}$$

number of dimensions

CI width

k=5,000 features

How many samples to have 95% CI narrower than 5%? α = 95%, ε = 5%	**N = 2,441**
Width of the 95% CI with 2000 observations? α = 95%, n = 2,000	ε **= 5,5%**
What CI are narrower than 5% with 2,000 observations? ε = 5%, n = 2,000	α **= 54,6%**

Application to cancer datasets from GEO

Microarray

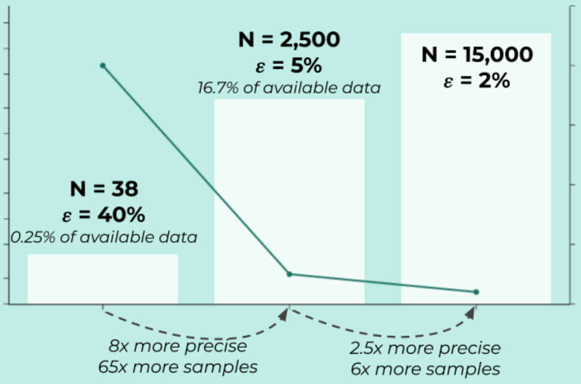
> $ 250K
for *de novo* data generation

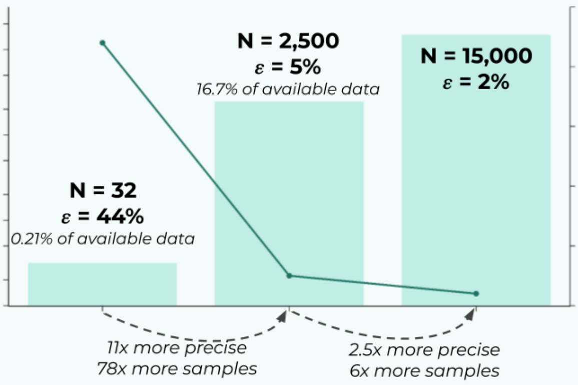
Target cohort size ~ 2,500 samples
for 95% CI narrower than 5%

RNA-Seq

> $ 500K
for *de novo* data generation

k=5,000 features





- **Data- and model-agnostic:** our formulas only depend on the number of *features* modeled. This guarantees their wide applicability.
- **Fast estimates:** our formulas can be used to inform project feasibility, before experimental design is finalized and before data is collected.
- **Untapped potential:** without data integration capabilities, public data remains under-utilized. Our study shows that only 15% of GEO data would improve study precision by a factor 8 to 11.
- **Addressing biases in available data:** whatever the cohort size, representativity is key. Data integration allows tailor-made cohorts, built to alleviate biases.
- **Challenges in data heterogeneity:** evolving disease classifications, non-standardized nomenclatures, improving sequencing technologies, changing gene name references...
- **Solutions exist:** batch effect correction, gene name harmonization, AI-powered clinical metadata cleaning...

Would you rather invest in under-powered *de novo* data, or in data integration capabilities?

CHECK OUR ONLINE CALCULATOR

Contact Akpéli Nordor, PharmD, PhD (akpeli@epigenelabs.com).

The authors have no conflict of interest to declare.

CHECK US OUT!