powerROC: An Interactive Web Tool for Sample Size Calculation in Assessing Models' Discriminative Abilities

François Grolleau

HealthRex Lab group meeting
August 28, 2024

In medicine, validating existing prediction models may be more valuable than developing new ones!

- Abundance of existing models

  eg >400 models for COPD, >300 for cardiovascular disease, >600 for covid-19†

- Lack of validation

  Despite many models being developed, only a small fraction are externally validated

- What matters is model's generalization

  External validation is crucial for demonstrating that a model can be applied beyond the population used in its development

→ **For validating prediction models, sample size calculation is critical!**

- Effective resource allocation
- Minimize type II errors (conclude nothing when we should conclude something)

†Collins et al. BMJ 2024; 384: e074819.

How to calculate the sample size for evaluating prediction models?

Talk outline:

- Review of AUROC (Area Under the Receiver Operating Characteristic curve)
- Determining the necessary sample size to evaluate a single prediction model
- Establishing the sample size needed to compare two prediction models

Objectives:

- Describe CAN estimators and show how to calculate confidence intervals for them
- Explain the concept of statistical power and how it pertains to AUROC
- Illustrate the principles of both nonparametric and parametric Monte-Carlo simulations

How to calculate the sample size for evaluating prediction models?

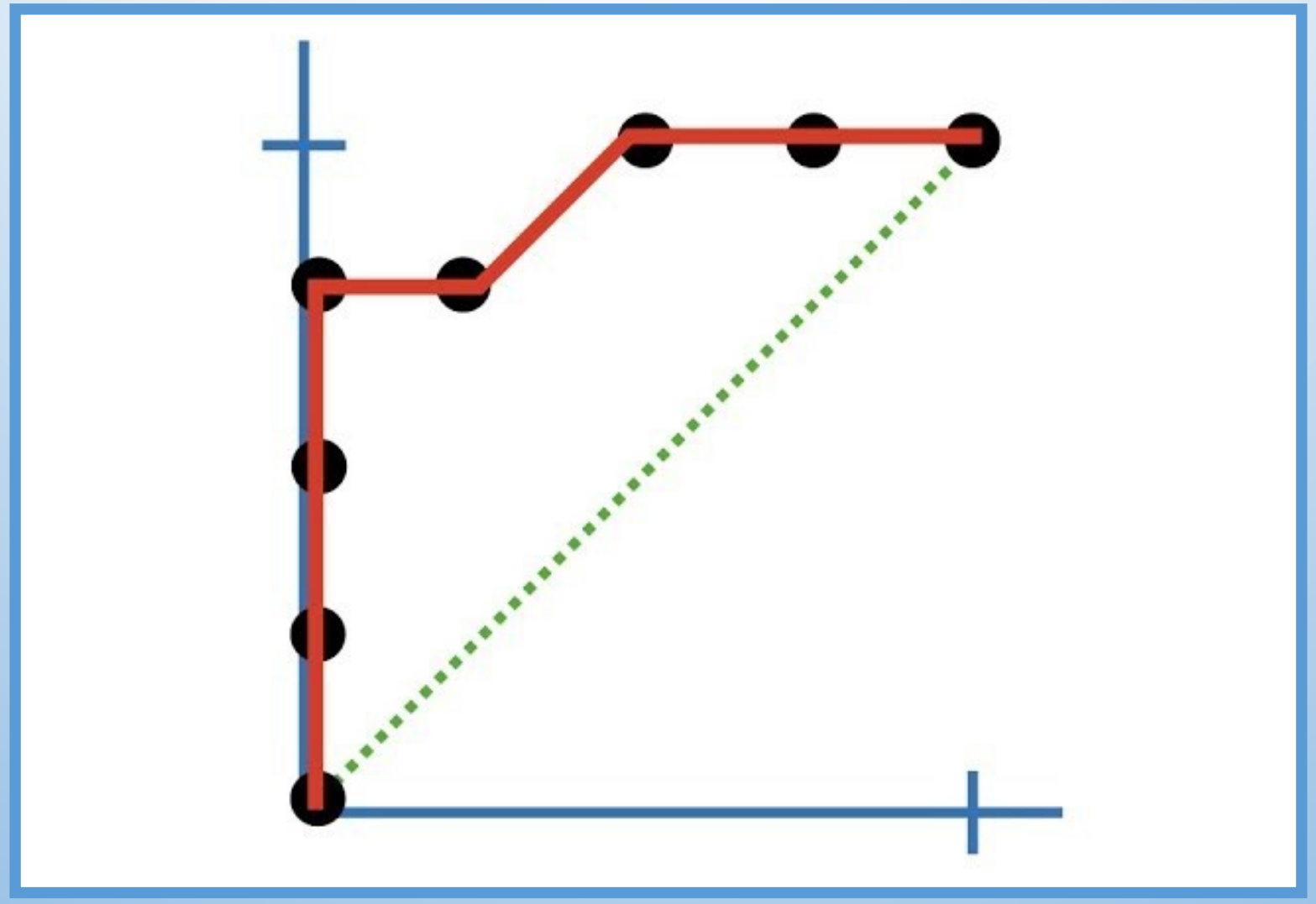
Talk outline:

- **Review of AUROC (Area Under the Receiver Operating Characteristic curve)**
- Determining the necessary sample size to evaluate a single prediction model
- Establishing the sample size needed to compare two prediction models

Objectives:

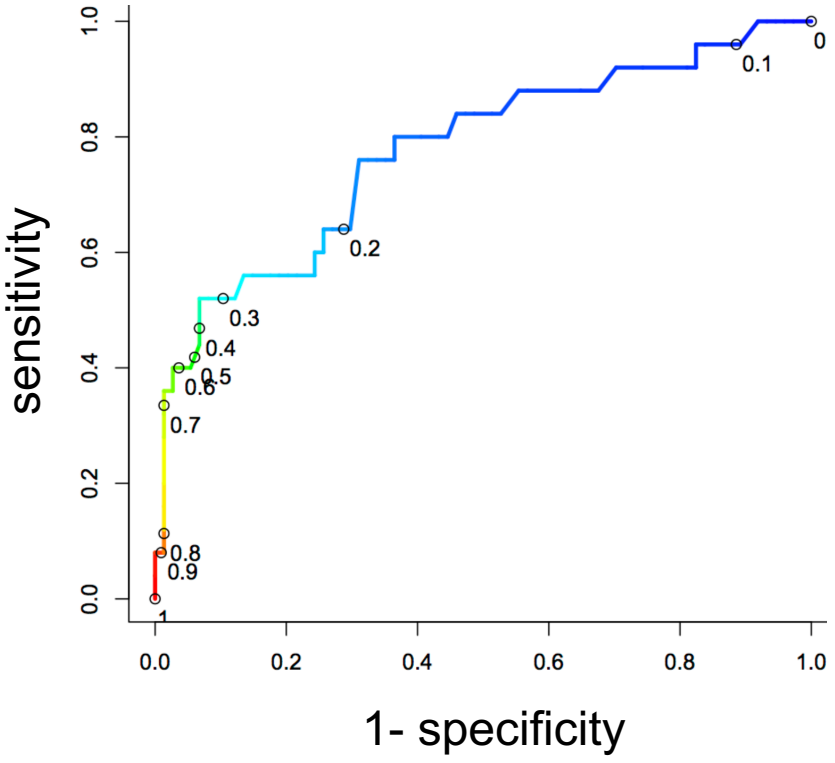
- Describe CAN estimators and show how to calculate confidence intervals for them
- Explain the concept of statistical power and how it pertains to AUROC
- Illustrate the principles of both nonparametric and parametric Monte-Carlo simulations

Receiver Operating Characteristic Curve



Receiver Operating Characteristic Curve

Row #	Label	Prediction
1	1	0.8
...	...	...
$n+1$	0	0.9
...	...	...
$n+m$	0	0.4



- For each threshold, create a 2X2 table
  and calculate the corresponding sensitivity & specificity
- Plot sensitivity and specificity for all thresholds

Aera under the ROC curve: definition

AUROC measures the discriminative ability of a classifier
i.e. the ability to distinguish between cases and controls

- AUROC = 0.5

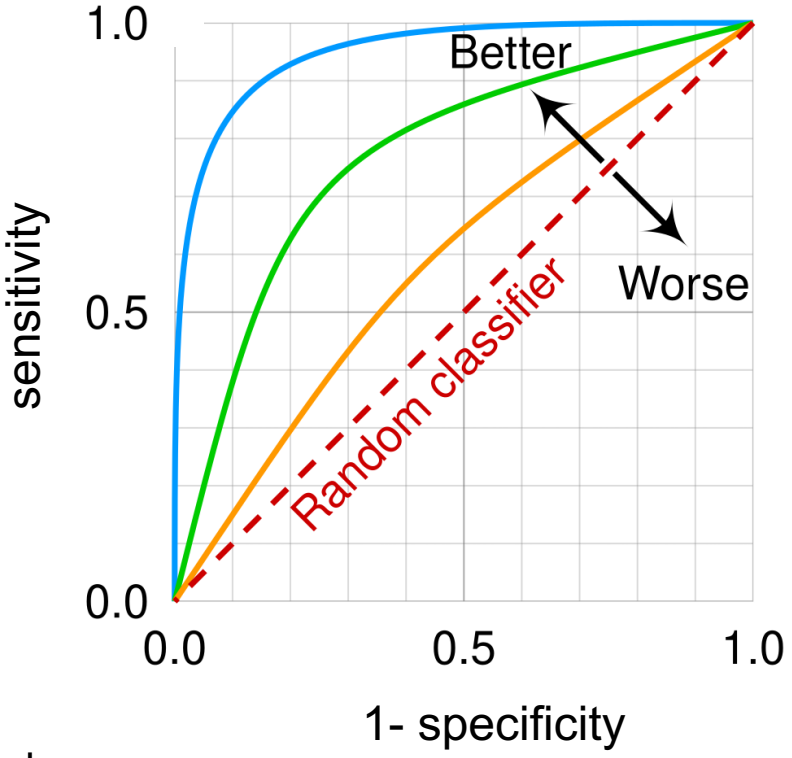
the model doesn't do better than the flip of a (possibly flawed) coin

- AUROC = 1

the model always predicts higher values for cases than for controls

- AUROC $\in]0.5, 1[$

more often than not the model predicts higher values for cases than for controls



Calculating AUROC: the standard view

- Calculate the AUROC by…

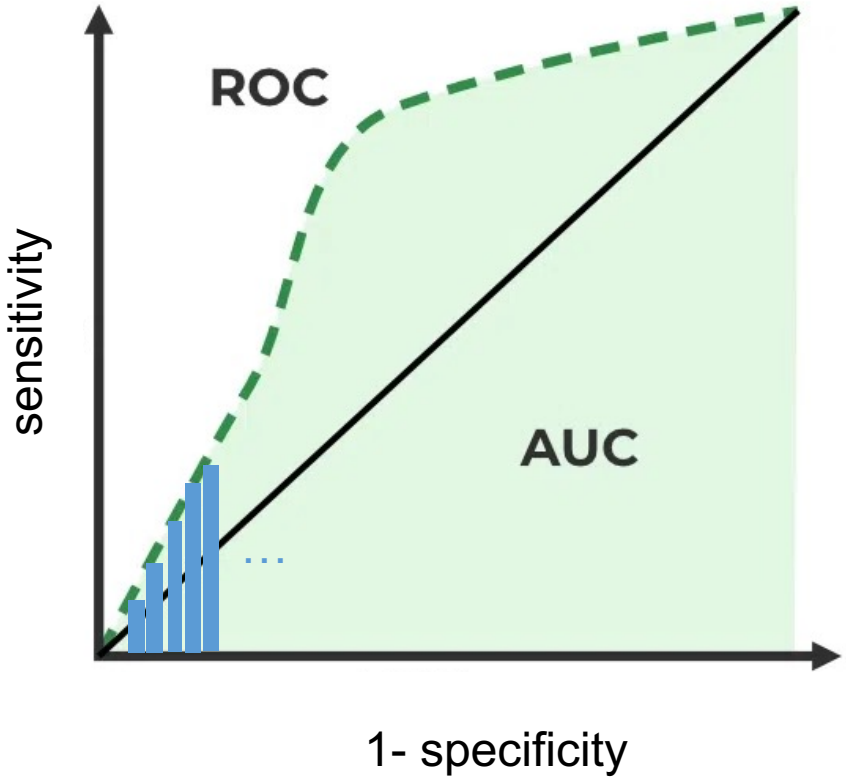
… measuring the area under the ROC curve!

ie summing up the area of little rectangles for numerical integration

- Nothing wrong with this approach

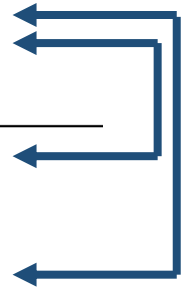
But it's hard to do statistical inference and sample size calculation

if we stick with this view of AUROC



Calculating AUROC by looking at all possible pairs (case & control)

Row #	Label	Prediction
1	1	0.8
...	...	...
$n+1$	0	0.9
...	...	...
$n+m$	0	0.4



1. Create a dataset of all possible pairs

2. Compare predictions in each pair

Pair #	Prediction (case)	Prediction (control)	Prediction case > control?
1	0.8	0.9	0
2	0.8	0.4	1
$n \times m$	...	...	0 or 1

3. $\widehat{AUROC}$=Mean of this column†

†This is a mathematical fact that we haven't proven here. (see Krzanowski & Hand 2009, p. 27. for a proof)

Calculating AUROC by looking at all possible pairs (case & control)

$$\widehat{\mathrm{AUROC}} = \frac{1}{nm} \sum_i \sum_j \mathbb{1}\{\text{prediction on case } i > \text{prediction on control } j\}$$

- $\widehat{\mathrm{AUROC}}$ is an estimator

It takes data as input and outputs a number → that's why we put a hat on AUROC

- $\widehat{\mathrm{AUROC}}$ is a random variable

If you sample a different dataset of the same size, AUROC takes a different value

- $\widehat{\mathrm{AUROC}}$ is a U-statistic

The equation above has a particular form, it "looks at all possible pairs"

→ The theoretical properties of U-statistics have been described by Hoeffding and others
We can directly apply powerful results from this literature!

The payoff of viewing AUROC as a U-statistic

$\widehat{\text{AUROC}}$ is a consistent asymptotically normal (i.e. "nice") estimator:

Centered on the true value of AUROC

Approximately follows

Normal distribution

$$\widehat{\text{AUROC}} \dot{\sim} \mathcal{N}\left(\text{AUROC}, SE^2\right)$$

Variance of the estimator (called standard error squared) is known, and it shrinks to zero as the sample size increases (see below)

and Hoeffding gives us a formula for its standard error:

$$SE(C) \approx \sqrt{\frac{C(1-C)\left(1+\left(\frac{N}{2}-1\right)\left(\frac{1-C}{2-C}\right)+\frac{\left(\frac{N}{2}-1\right)C}{1+C}\right)}{N^2\phi(1-\phi)}}$$

See Figure 3 in Riley et al. BMJ 2024.

The payoff of viewing AUROC as a U-statistic

$\widehat{\text{AUROC}}$ is a consistent asymptotically normal (i.e. "nice") estimator:

Centered on the true value of AUROC

Approximately follows

Normal distribution

$$\widehat{\text{AUROC}} \dot{\sim} \mathcal{N}\left(\text{AUROC}, SE^2\right)$$

Variance of the estimator (called standard error squared) is known, and it shrinks to zero as the sample size increases (see below)

and Hoeffding gives us a formula for its standard error

We can use this machinery to calculate asymptotic (i.e. non bootstrap) confidence intervals

$$\hat{\text{CI}}_{95\%}(\text{AUROC}) = \widehat{\text{AUROC}} \pm 1.96 \times \hat{SE}$$

How to calculate the sample size for evaluating prediction models?

Talk outline:

- Review of AUROC (Area Under the Receiver Operating Characteristic curve)
- **Determining the necessary sample size to evaluate a single prediction model**
- Establishing the sample size needed to compare two prediction models

Objectives:

- Describe CAN estimators and show how to calculate confidence intervals for them
- Explain the concept of statistical power and how it pertains to AUROC
- Illustrate the principles of both nonparametric and parametric Monte-Carlo simulations

The payoff of viewing AUROC as a U-statistic

$\widehat{\text{AUROC}}$ is a consistent asymptotically normal (i.e. "nice") estimator:

Centered on the true value of AUROC

Approximately follows

Normal distribution

$$\widehat{\text{AUROC}} \dot{\sim} \mathcal{N}\left(\text{AUROC}, SE^2\right)$$

Variance of the estimator (called standard error squared) is known, and it shrinks to zero as the sample size increases (see below)

and Hoeffding gives us a formula for its standard error

We can reverse engineer this machinery to calculate the sample size needed for a level of precision

i.e. use the standard error formula to iteratively increase the sample size until the target standard error is reached

https://fcgrolleau.github.io/powerROC/

Questions?

How to calculate the sample size for evaluating prediction models?

Talk outline:

- Review of AUROC (Area Under the Receiver Operating Characteristic curve)
- Determining the necessary sample size to evaluate a single prediction model
- **Establishing the sample size needed to compare two prediction models**

Objectives:

- Describe CAN estimators and show how to calculate confidence intervals for them
- Explain the concept of statistical power and how it pertains to AUROC
- Illustrate the principles of both nonparametric and parametric Monte-Carlo simulations

Comparing two prediction models

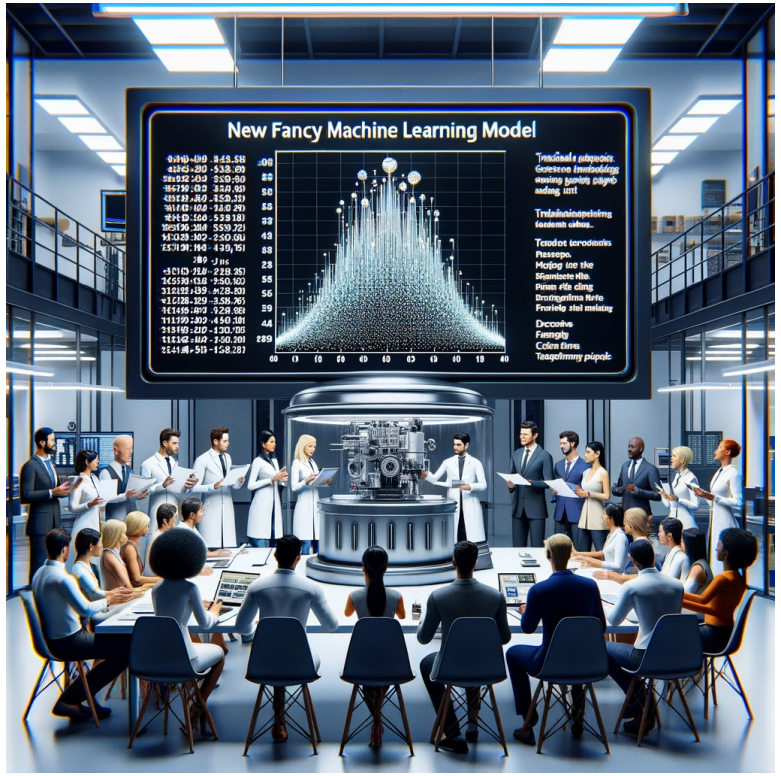
- Imagine the following:

- A new fancy ML model comes out for prognostic in cardiovascular disease
- Authors claim it outperforms all other models

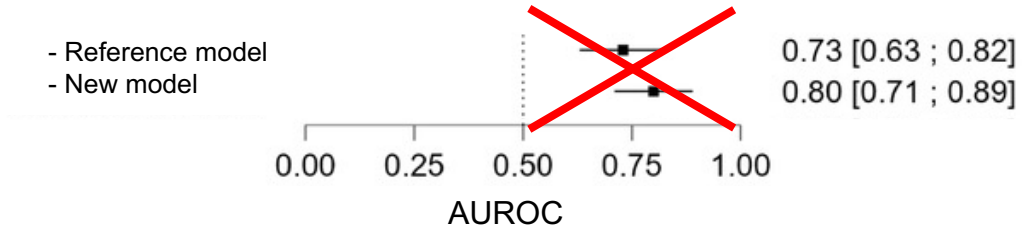
→ How would we know if it's better than the reference model used in practice?

Answer: conduct an external validation study compare AUROCs:

new vs reference model (e.g. Framingham Risk Score)



- How <u>not</u> to compare AUROCs



→ When both models are applied to the same patients, it's not OK to conclude a difference by checking if 95%CIs overlap!

Comparing two prediction models

Goal: Conduct an external validation study compare AUROCs:

new vs reference model (e.g., Framingham Risk Score)

Question 1: How to compare AUROCs for models applied to the same patients?

1. Calculate $\hat{\mathrm{CI}}_{95\%}(\mathrm{AUROC}_{\mathrm{new}} - \mathrm{AUROC}_{\mathrm{ref}})$ via the bootstrap and check if the resulting CI contains zero. If not, conclude a difference. Valid but slow…

or

2. Compare $\mathrm{AUROC}_{\mathrm{new}}$ and $\mathrm{AUROC}_{\mathrm{ref}}$ by calculating a p-value from DeLong test for paired data†

Valid, fast, relies on the theory of U-statistics. Implementations available in R and Python.

†DeLong. Biometrics. 1988.

Comparing two prediction models

Goal: Conduct an external validation study compare AUROCs:

new vs reference model (e.g., Framingham Risk Score)

Question 2: How many samples are needed for the external validation study?

→ Specify the statistical power we wish to attain. The idea is this:

1. Assuming there is a real difference $\mathrm{AUROC}_{\mathrm{new}} \neq \mathrm{AUROC}_{\mathrm{ref}}$

2. and you could repeat the external validation study infinitely many times,

3. "how often" would you like to see $\mathrm{DeLong}\ p < 0.05$?

(i.e., rightly conclude a difference)

$$\mathrm{Power} = \mathbb{P}(\mathrm{DeLong}\ p < 0.05 | H_1)$$

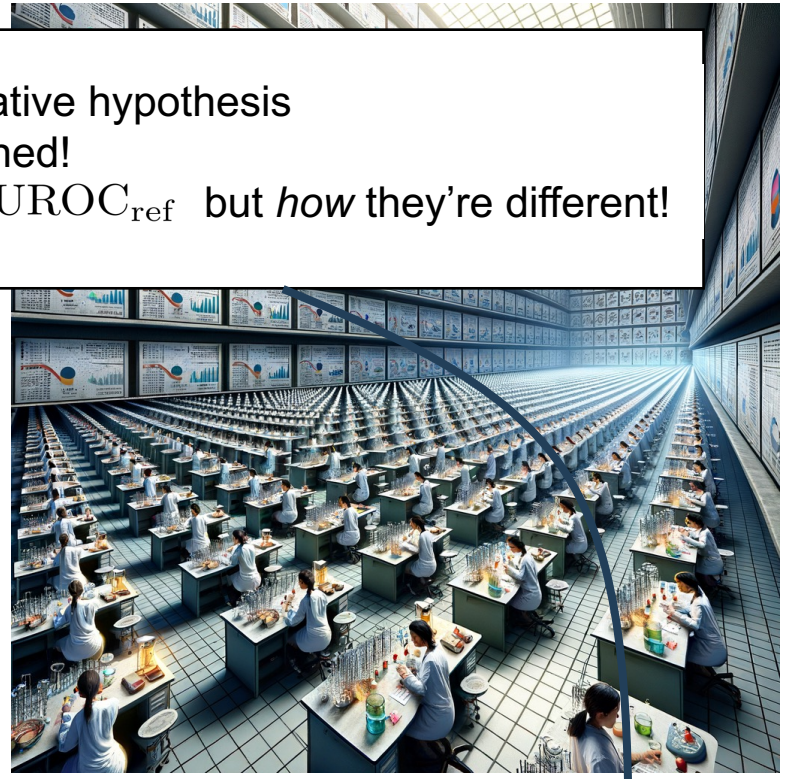
Typically, 80% power is chosen

Comparing two prediction models

Goal: Conduct an external validation study com

new vs reference model (e.g., F

⚠ For power analyses, the alternative hypothesis needs to be very precisely defined!
i.e., not just $\mathrm{AUROC}_{\mathrm{new}} \neq \mathrm{AUROC}_{\mathrm{ref}}$ but *how* they're different!



Question 2: How many samples are needed for the external validation study?

→ Specify the statistical power we wish to attain. The idea is this:

1. Assuming there is a real difference $\mathrm{AUROC}_{\mathrm{new}} \neq \mathrm{AUROC}_{\mathrm{ref}}$

2. and you could repeat the external validation study infinitely many times,

3. "how often" would you like to see $\mathrm{DeLong}\ p < 0.05$?

(i.e., rightly conclude a difference)

$$\mathrm{Power} = \mathbb{P}(\mathrm{DeLong}\ p < 0.05 | H_1)$$

Typically, 80% power is chosen

Calculate sample size for comparing two prediction models: Using a pilot validation study

Question 2: How many samples are needed for the external validation study?

→ Specify the statistical power we wish to attain: typically, 80%

→ Specify how $\mathrm{AUROC}_{\mathrm{new}}$ and $\mathrm{AUROC}_{\mathrm{ref}}$ are different by providing a pilot test set

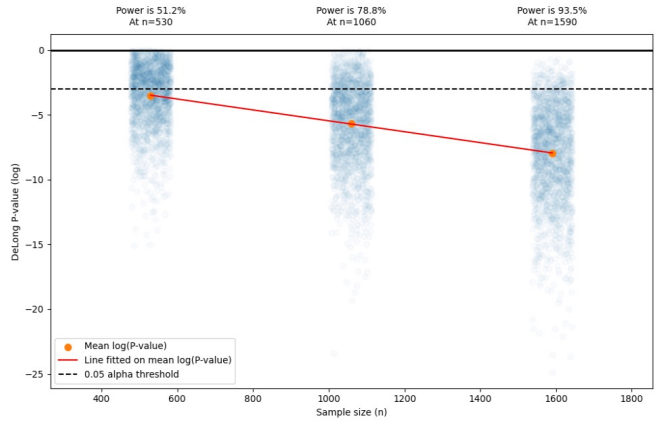
→ From this test set, resample (with replacement) to create B datasets of each of size N

→ For each dataset $b = 1, 2, \ldots, B$ calculate a Delong p-value $p_1, p_2, \ldots, p_B$

→ Estimate power at sample size N as $\mathbb{P}(\text{DeLong } p < 0.05 | H_1) \approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{\text{DeLong } p_b < 0.05\}$

Pilot test set used:

Event	Prediction from model A	Prediction from model B
0	0.694	0.121
1	0.526	0.825
1	0.687	0.584
0	0.136	0.059
0	0.315	0.127
0	0.137	0.052
1	0.689	0.294
0	0.248	0.649
0	0.228	0.033
0	0.292	0.040



This is an example of nonparametric Monte Carlo simulation
It's nonparametric because we're sampling from empirical data

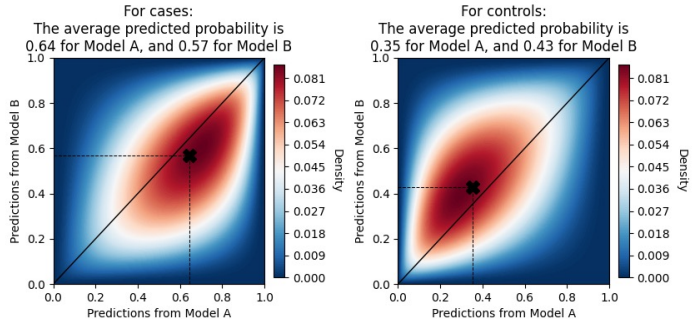
https://fcgrolleau.github.io/powerROC/

Calculate sample size for comparing two prediction models: Without a pilot validation study

Question 2: How many samples are needed for the external validation study?

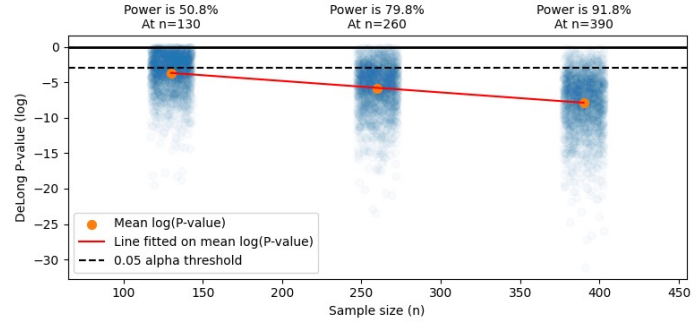
→ Specify the statistical power we wish to attain: typically, 80%

→ Specify how $\mathrm{AUROC}_{\mathrm{new}}$ and $\mathrm{AUROC}_{\mathrm{ref}}$ are different by specifying parameters for the data-generating process



→ Sample from this data-generating process to create B datasets of each of size N

→ For each dataset $b = 1, 2, \ldots, B$ calculate a Delong p-value $p_1, p_2, \ldots, p_B$



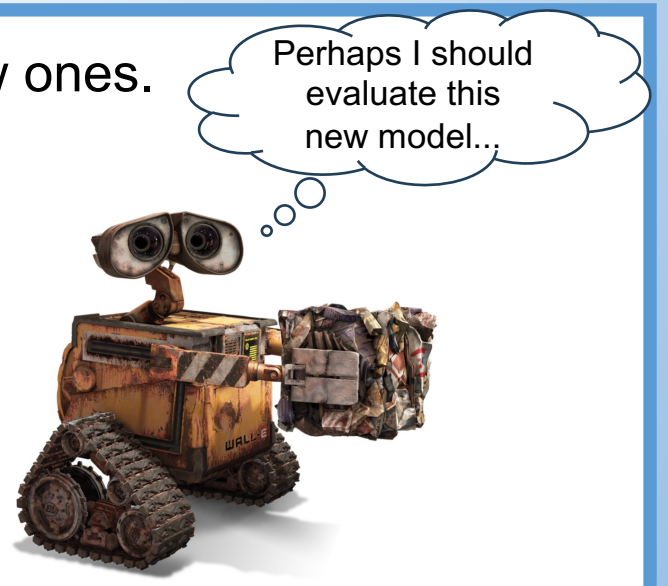
→ Estimate power at sample size N as $\mathbb{P}(\text{DeLong } p < 0.05 | H_1) \approx \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\{\text{DeLong } p_b < 0.05\}$

This is an example of parametric Monte Carlo simulation
It's parametric because we're sampling from a specified probability distribution

https://fcgrolleau.github.io/powerROC/

Conclusion

- In healthcare, validating prediction models often outweighs creating new ones.



- We need external validation studies to:
  - Precisely evaluate the performance of new models
  - Compare the performance of new models vs the reference models
- International guidelines† mandate reporting the process of determining the sample in these studies.
- Research is needed to make the theory and practice of sample size calculation more accessible.

Thank you for your attention!

†TRIPOD+AI statement. BMJ 2024.