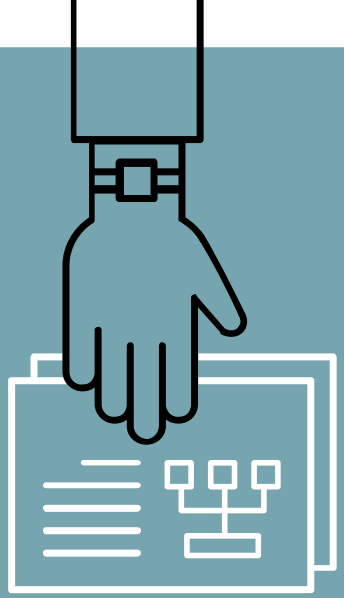
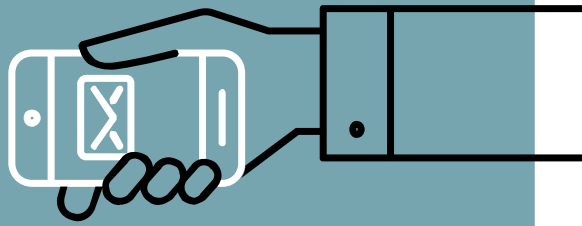
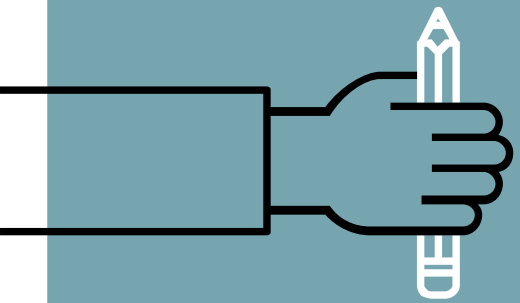
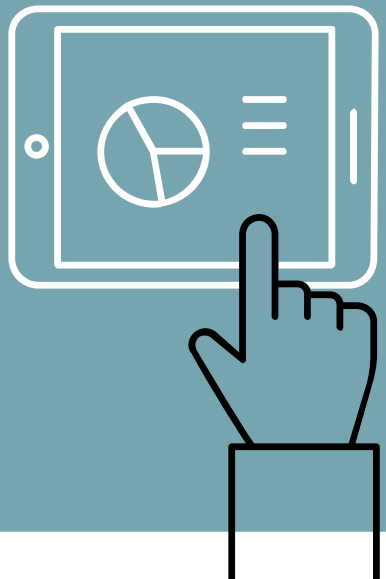




BST 270
Reproducible Data Science

Winter 2022
Session 5



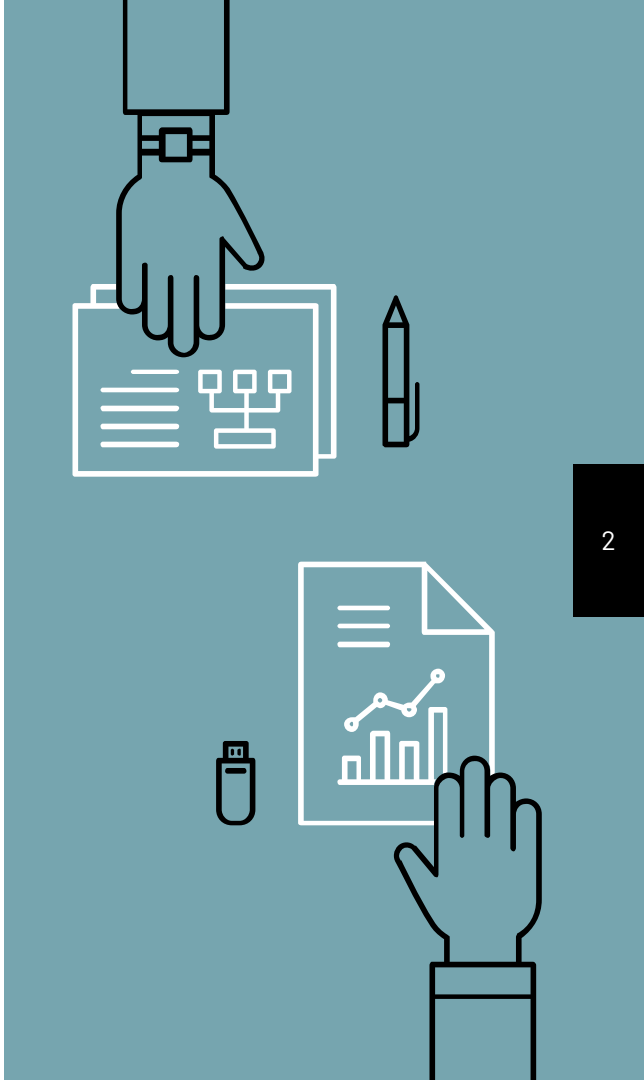


Module 5 Comments

I have learned Coefficient of Determination in the past, but I had never interpreted the scaling by the variance as comparing this estimator to the mean of the responses. That's a pretty intuitive way of looking at validation metrics, which had not previously occurred to me.

While I have only used CV in my research, it is interesting to learn from the modules that CV can be over-optimistic about the models and it is always better to include multiple datasets (from different studies with different populations and thus different underlying distribution) and use both CV and CSV. This again underscores the importance of having data available on public data repositories, not only for reproducibility but also for providing researchers with various datasets to work with.

Performing simulations is a great way to evaluate possible algorithms to train prediction models, and to evaluate model operating characteristics.



Module 5 Comments

Validating the performance of a ML model on other studies is hard because the data distribution will be different across different studies, and that will make a highly predictive model fail in other studies. People in computer science refer to this as covariate shift problem (i.e. X has different distributions).

It will be even harder for multi-study machine learning when the distribution of Y given X are different, and Prof. Giovanni Parmigiani's group has some wonderful papers to deal with this problem. For example, they train study-specific models on different studies, and then ensemble the predictions. Here is one of the papers: https://pubmed.ncbi.nlm.nih.gov/29531060/

Module 5 Comments

Another good approach for assessing the performance of a model used for predicting a binary outcome is to make calibration plots. The idea is bin your predicted values (e.g. from 0 to 5 percent, 5 to 10, … , 95 to 100), take the average predicted probability in each bin, and plot them against the corresponding average value of the outcome. For a well-calibrated model, we'd expect the plotted values to fall roughly on a 45-degree line — ideally, if we took all of the instances that our model assigned a probability of (approximately) X%, the corresponding true outcome would occur roughly X% of the time.

FiveThirtyEight keeps track of their prior forecasts and has a nice write-up on calibration plots with some examples from their own models: https://projects.fivethirtyeight.com/checking-our-work/

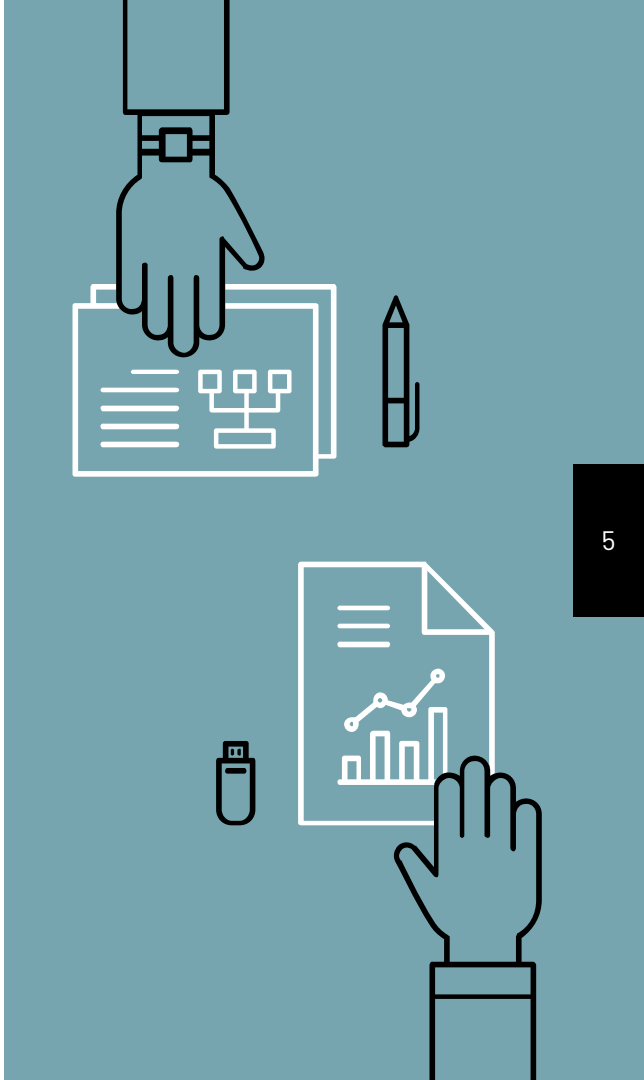
Module 5 Comments

This recent paper gives a discussion of cross validation that some members of the class might find interesting: https://arxiv.org/pdf/2104.00673.pdf

I'm only familiar with meta-analysis in the context of GWAS, so I was interested to see its application to studying model validation statistics in the Clustering video.

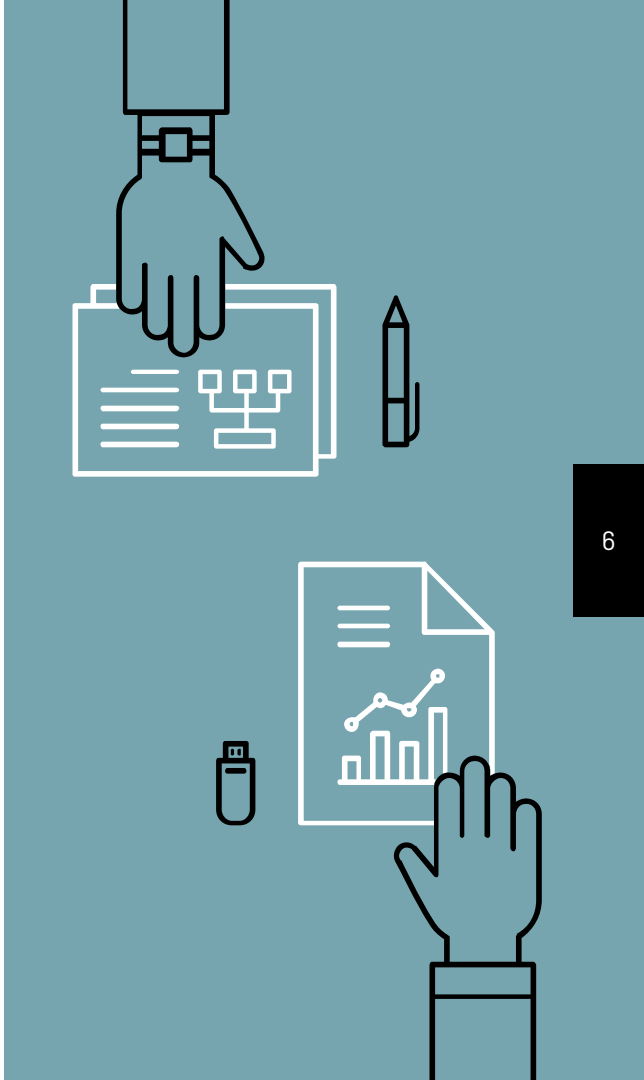
Professor Trippa introduces an interesting idea of using summary level statistics from different studies to form a federated / pooled estimate of interest. This is an emerging field in both statistics and machine learning, especially in the biomedical research. This technique is especially useful when we have limited sample sizes from individual studies, pooling the estimates together can possibly increase the power of the test and get more efficient estimates. This is also very useful when we have restrictions in pooling (raw) medical data for confidentiality concern.

Many of the methods and techniques introduced are standard and simple, but they are widely used in almost every data science research. I remembered how I was amazed by how intuitive and simple these methods (i.e. bootstrap, CV, decision tree, etc.) are when I first saw them. I hope one day I can understand the rigorous statistical proof of these methods.



Module 5 Comments

I’ve been told that the ROC curve becomes less useful on data with imbalanced classes. This notebook goes over how the precision recall curve can be better in these cases, because “False Positive Rate (False Positives / Total Real Negatives) does not drop drastically when the Total Real Negatives is huge” https://www.kaggle.com/lct14558/imbalanced-data-why-you-should-not-use-roc-curve/notebook



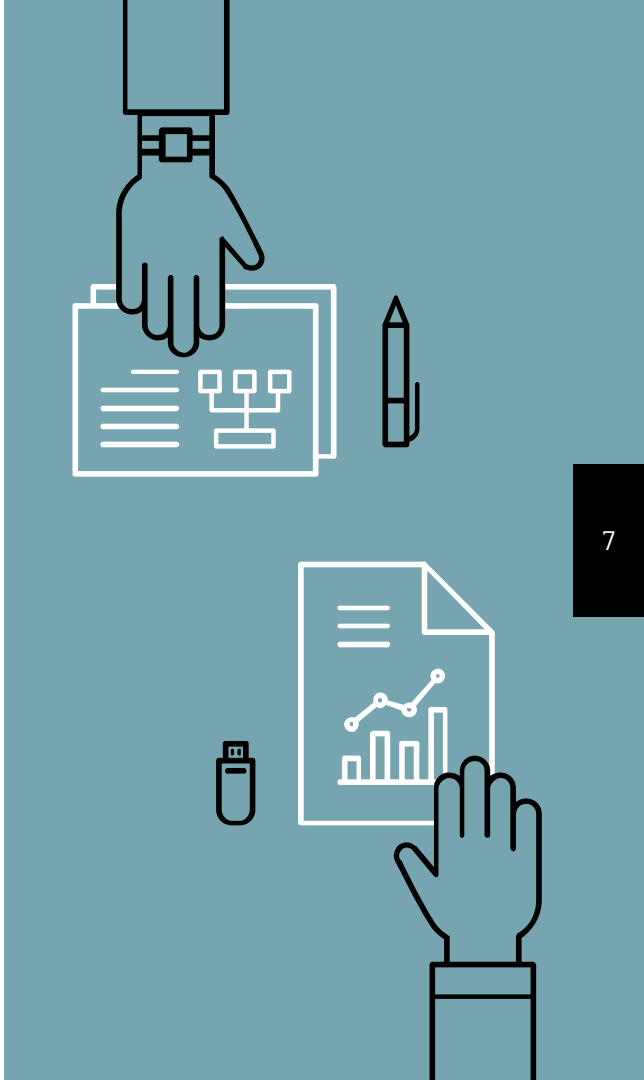
Module 5 Comments

Aren't these validation metrics more under the umbrella of replicability than reproducibility? I say this because they're trying to make the case that you have a good statistical model, whose results will agree with the results of other good models... right?

I understand the importance of all of the topics covered in this module, but how do they tie to reproducibility? For instance, using validation metrics to assess if a prediction model is reliable and using cross validation to select an algorithm are both clearly important in ensuring your research is valid, but it seemed like this section was a bit disconnected from the previous module in terms of staying on the topic of reproducibility.

These videos were a nice review of some statistical methods and Lorenzo's research area, but I don't really see how this section relates to reproducible research. Perhaps the connection is made later on?

- I agree and to be honest I'm not sure this module really fits in. The main purpose of the module is exactly what you pointed out and it is really about making sure your methods/results are correct/valid before sharing with other researchers.



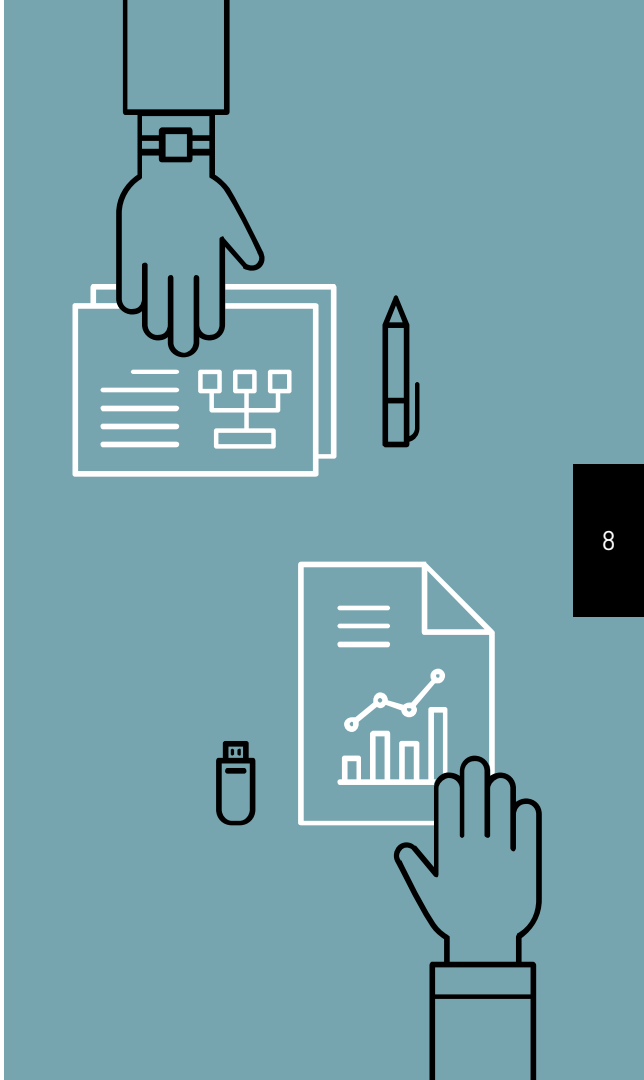
Module 5 Discussion

When submitting papers for review, how common is it for reviewers to run through code for the simulations? How much is reproducibility enforced when using simulated data as compared to real datasets that might need more standards for privacy/security?

- Not very common. I'm not sure – I would think for actual datasets it would be more stringent.

When we are using AUC as the validation metric, we often make the assumption that specificity and sensitivity have the same importance, which in many cases is not true. Can we find another asymmetric validation metric to focus more on either one side, if possible?

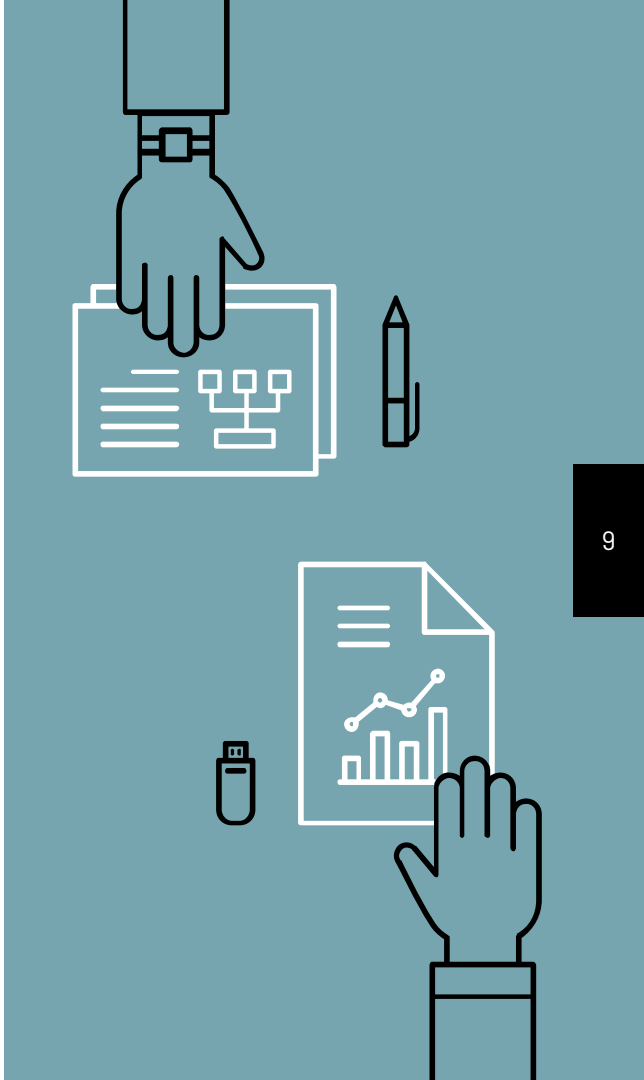
- The AUROC plot actually plots different pairs of sensitivity and specificity values and can be used to value one metric over the other.



Module 5 Discussion

When do we use Bootstrap instead of cross validation? What is the difference between them?

- Bootstrap resamples with replacement and usually produces “new” data sets with the same number of cases as the original data set. Due to the drawing with replacement, a bootstrapped data set may contain multiple instances of the same original cases, and may completely omit other original cases.
- Cross validation resamples without replacement and produces datasets that are smaller than the original. These data sets are produced in a systematic way so that after a pre-specified number k of surrogate data sets, each of the n original cases has been left out exactly once.
- Cross *validation*’s primary purpose is measuring the performance of a model. Bootstrapping is primarily used to establish empirical distribution functions for a widespread range of statistics (e.g. the variation of the mean to the variation of models in bagged ensemble models).



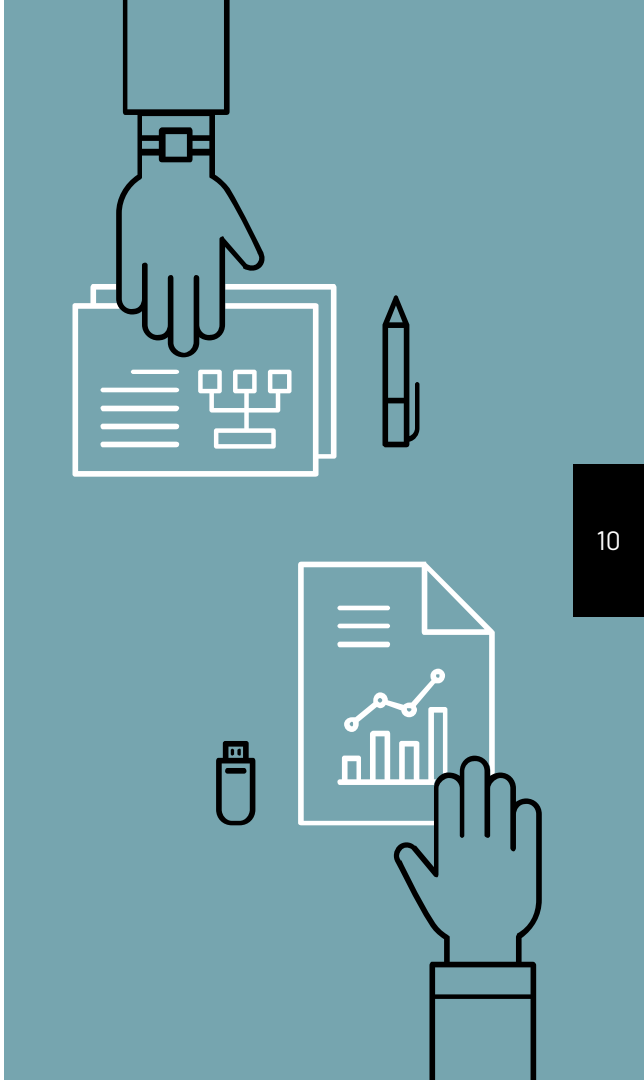
Module 5 Discussion

The construction of the C-index seems like it came out of thin air. Is there any place we can see more about the Kaplan Meier indices Lorenzo had mentioned in the videos?
Could you explain the estimator for censoring cases?

- I haven't been able to find anything besides the notes that go along with that video. I'm sure one of Lorenzo's papers has it but I need to look.

One new thing I learned was the basics of the concordance measure in survival analyses. I suppose this can be interpreted like an AUC measure that accounts for censoring?

- Exactly

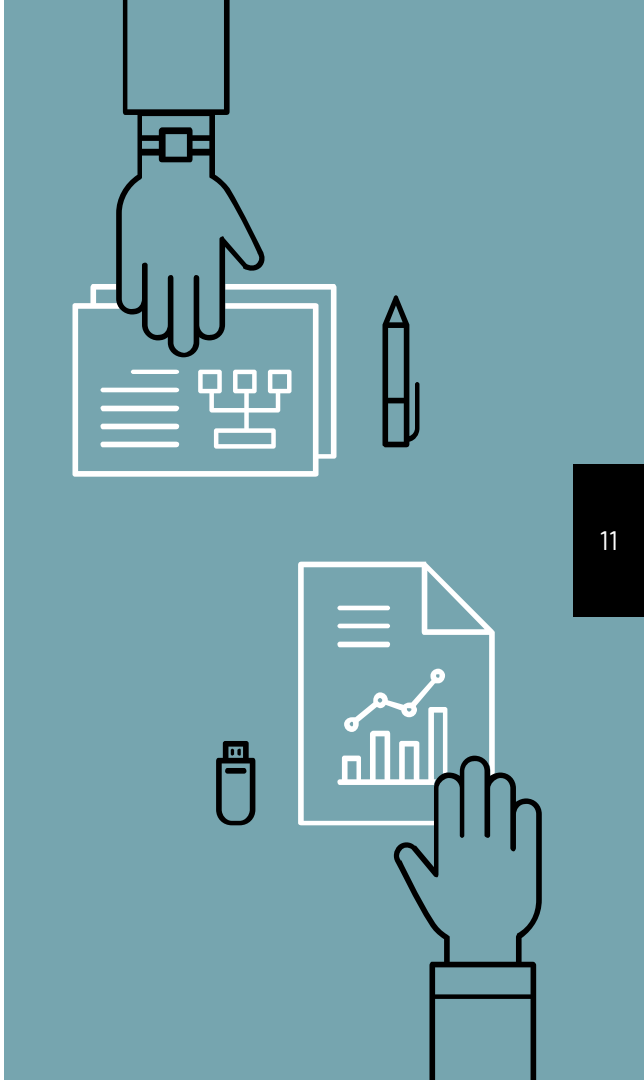


Module 5 Discussion

I'd never heard of the Brier score before only AUC so I'm wondering if disciplines or situations tend to prefer one over the other or is really a case-by-case basis?

I was quite surprised that I hadn't heard of Brier score before this. It was interesting to learn about this score. Is it common to report Brier score in research for a classification problem? Mostly, scores like recall, precision, F1-score, AUC, AUPR are used and taught.

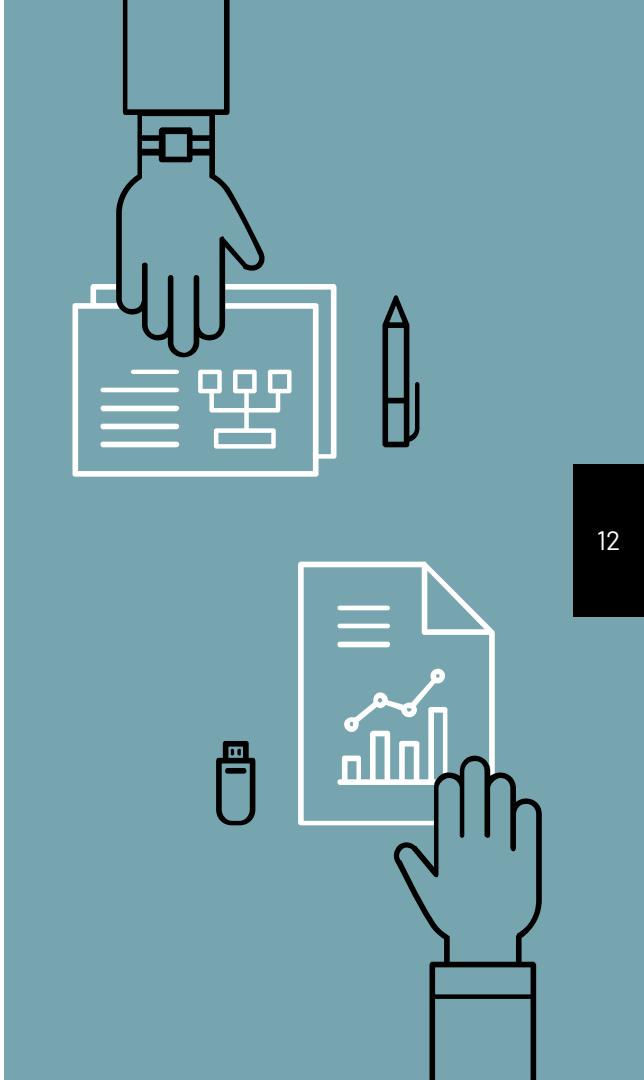
- I think it depends on the discipline - the Brier score is typically used by people who do survival analysis.



Module 5 Discussion

Are these measures (AUC, R^2, etc.) ever actually meaningful on their own or just when comparing several models? If the latter, how would you use these metrics to assess whether a first attempt at a problem were any good?

- They can be good on your own. For example, if your model has an AUC value of 0.5, the model is basically performing as well as a coin flip in terms of classification, basically meaning your model is useless.

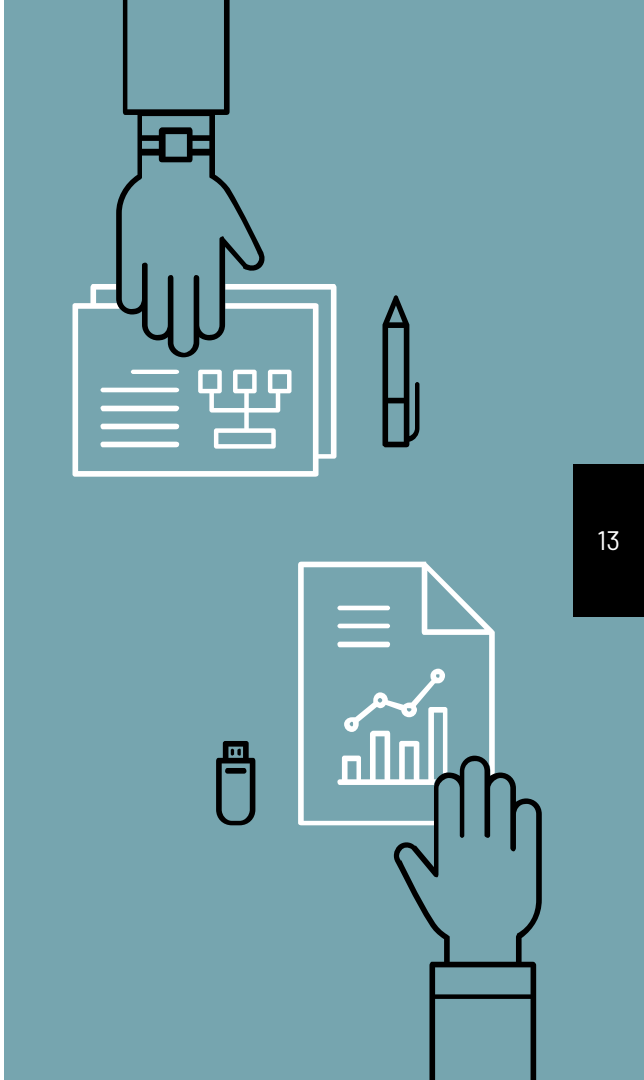


Module 5 Discussion

How much and what kind of due diligence is necessary to ensure that a training dataset is appropriate for the test dataset?

Something that came up in the modules and in several of the summer project talks is how to simulate data that is similar enough to the data behind real-world applications of the methodology. This seems like it could be a way more subtle issue than it would appear at first, and I'm wondering if there are best practices for simulated data generation.

- It is definitely more subtle and not at all simple. In my experience, simulating datasets with the same summary statistics is usually a good start, but I'm sure there are many more details to think about for different types of data.



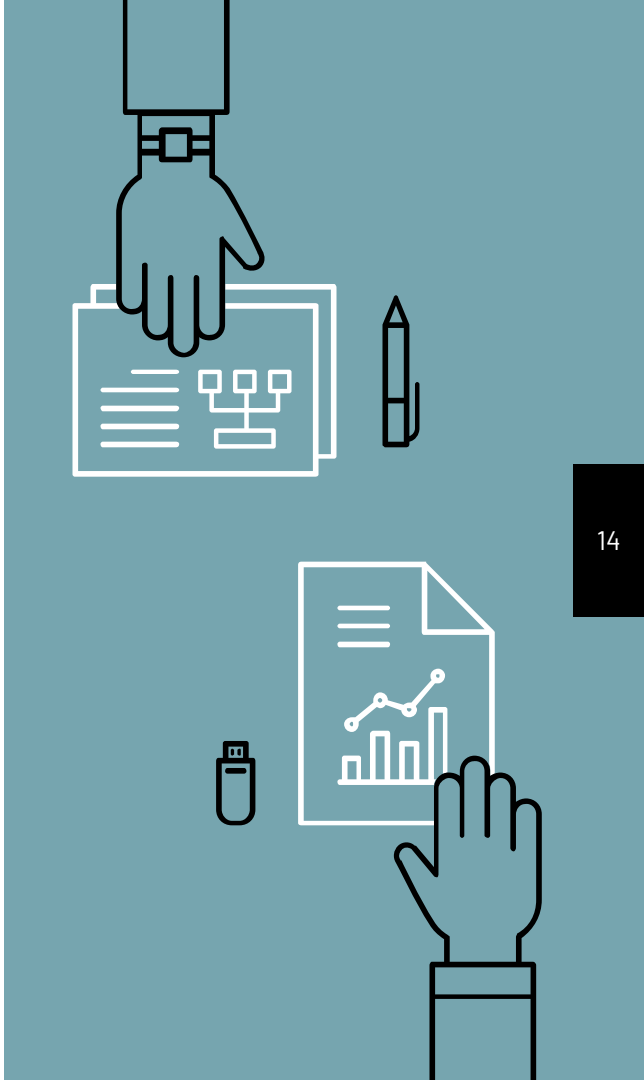
Module 5 Discussion

In video 6.4.1, while explaining the AUC, Lorenzo introduces a value called tau. Which he says dichotomize the values M(X1) to M(X100). What is the purpose of this value and how to determine its optimum value?

- Tau is the threshold you choose in order to classify something as a 0 or 1. The output of your model (for binary classification) is the P(Y = 1), and is a predicted probability ranging from 0-1. If your predicted probability is 0.6, and tau is 0.5, you would classify the 0.6 as a 1 (everything above tau is rounded up, everything below is rounded down). Different values of tau create different values of sensitivity and specificity. The optimum value is what you decide it is - usually, researchers are more concerned with having either a high sensitivity or a high specificity, depending on their classification task.

Have there ever been instances in your own work where bootstrapping seemed appealing but was computationally infeasible/expensive?

- Yes. Many metrics in network science are computationally expensive because of the size of the network. I've had to split up my simulations as different jobs on the cluster.



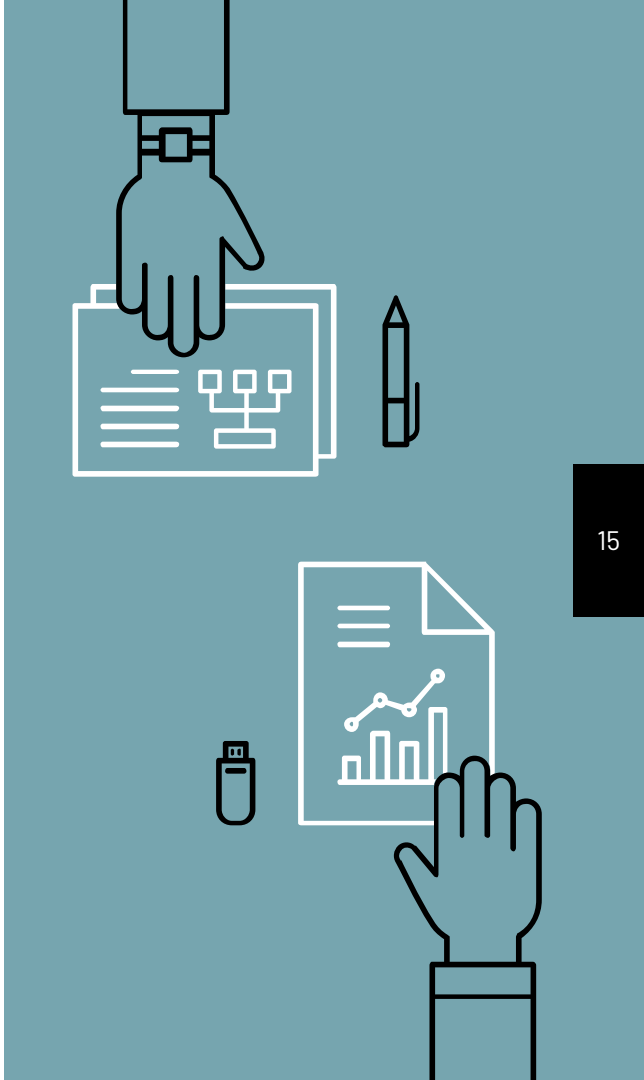
Module 5 Discussion

Does using Cross Study Validation (CSV) pose additional issues for reproducibility since the data comes from various sources?

- It could.

Does using CSV also pose difficulties if one of the underlying data sets used was ripe with bias or was fabricated?

- Absolutely.



Module 5 Discussion

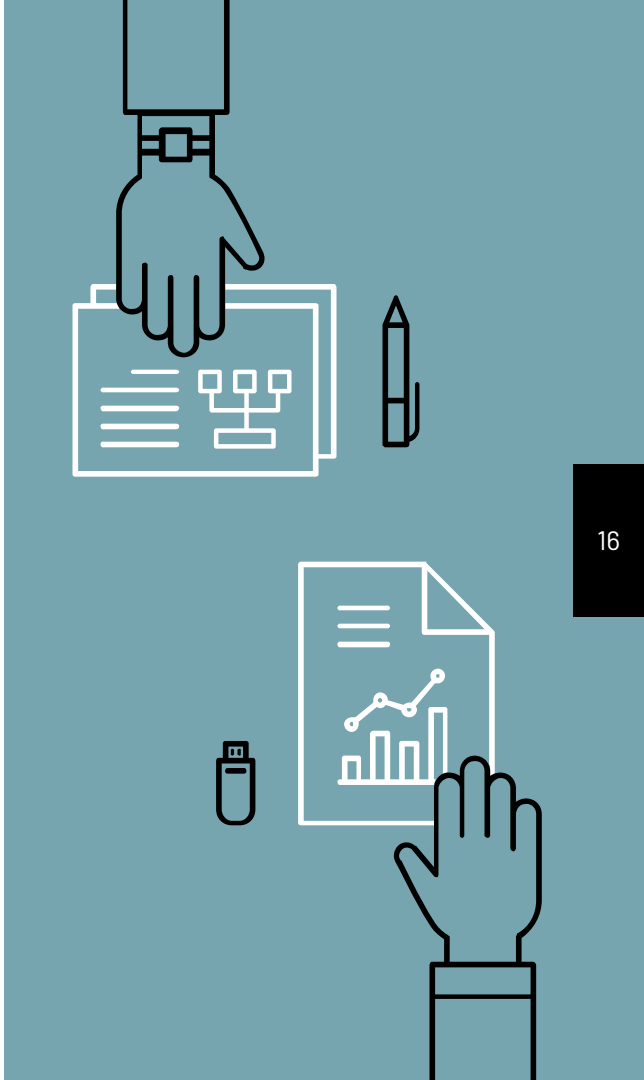
Likely an involved answer, but how does estimation change in the RE models for treatment effect?

What simulation procedure was used in the Simulations video?

- From the paper: "We define the model through a resampling procedure that we apply to the eight breast cancer datasets in Table 1. The resampling scheme is a combination of parametric and nonparametric bootstrap." Link to the paper

Given that cross validation is mostly useful for prediction, are there forms of cross validation that are used mainly in inferential problems?

- Not that I know of.

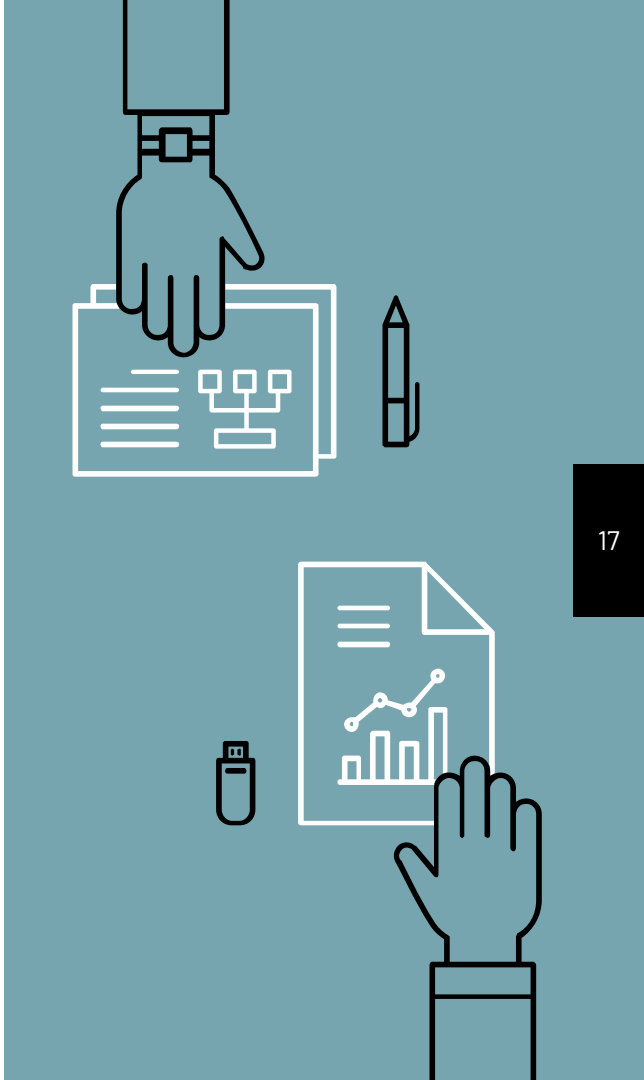


Module 5 Discussion

In the introduction video, how do we determine which dataset is of good quality and should be added for low noise while others are of bad quality and to be added of high level of noise?

Is there any metrics that we can use to reflect how closely our empirical estimation of ROC actually is to the true ROC given our model M? Or perhaps are any of these metrics overlapping with the metrics that we use to evaluate the fit of a logistic regression model?

- I need to look into this more, but you could simulate data that is similar to the data you have and calculate the true ROC, and then use your model on the simulated data to come up with the estimated ROC.

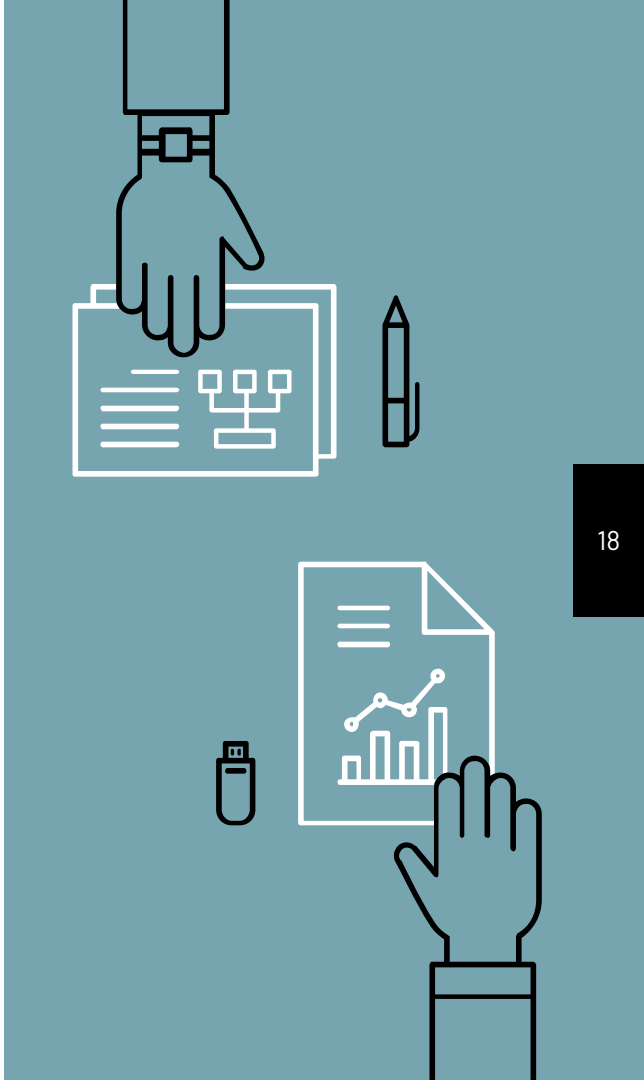


Module 5 Discussion

What are some strategies to develop cross validation methods that don't leak information (when you are pooling many studies together). How do you go about thinking about your Cross Validation pipeline?

- Here is one strategy, but it's only for when using one dataset. I'll need to look into this more when pooling studies. Although I think for cross-study-validation you aren't pooling the datasets.

I know recently R has introduced tidymodels, effectively a tidyverse for model fitting. Within this framework, they consider data pre-processing (thinks like missing value imputation, PCA, etc.) to be part of the CV process. Do you agree with this philosphy? When may or may not it be practical?



Module 5 Discussion

How do people choose which evaluation metrics to use/report in their studies/papers?

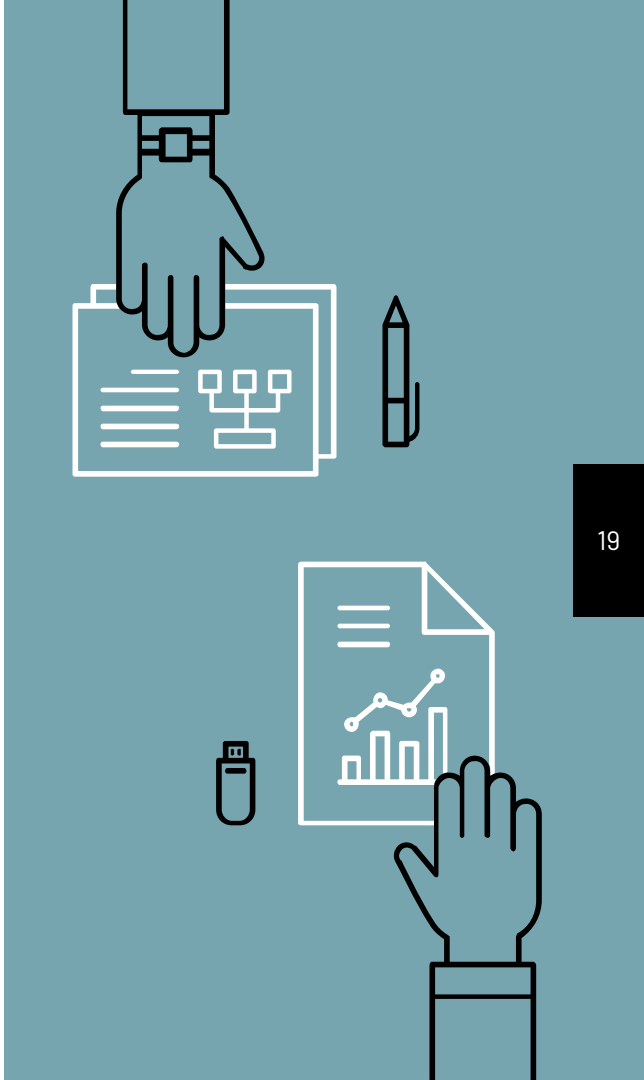
- Usually, it's whichever ones make their model look the best. Some researchers, including myself, tend to report as many as I can to give the reader a clearer picture of what my model is good at and where it struggles.

Brier score is useful when the primary goal is binary prediction, and we want our estimated probabilities to be correctly near 0 or 1. What if, instead, we want a realistic probability. That is, what if we're more concerned about the actual predicted probabilities rather than the classification accuracy?

- So, predicted probabilities are generally for estimating how sure we are that something should be classified as a 1, or some other category. Can you give me an example?

Are simulations really reliable? How to make it close to the real data?

- They can be, depends on the data and the method. There are ways to make it close to the data but this is an active area of research across Biostatistics subfields.

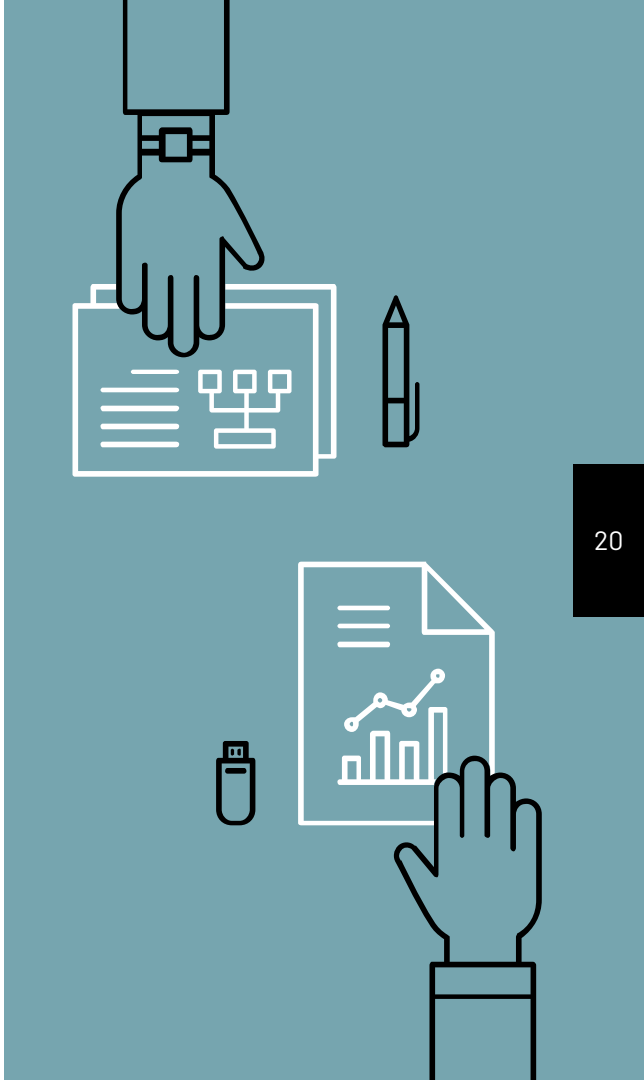


Module 5 Discussion

I've been curious about 'reproducibility' in terms of finding the same effects/associations in different studies (i.e., meta-analyses). Does that even fall under the umbrella of 'reproducibility'? What sorts of statistical techniques are there for meta-analyses that compare different studies and what conclusions might be able to make of any differences between effect sizes from the different studies?

When doing a cross validation, if you have multiple datasets you can combine, is it better to train a model on one dataset and test it on the other, or combine the two datasets and then split the combination into train/test sets?

- Great question.



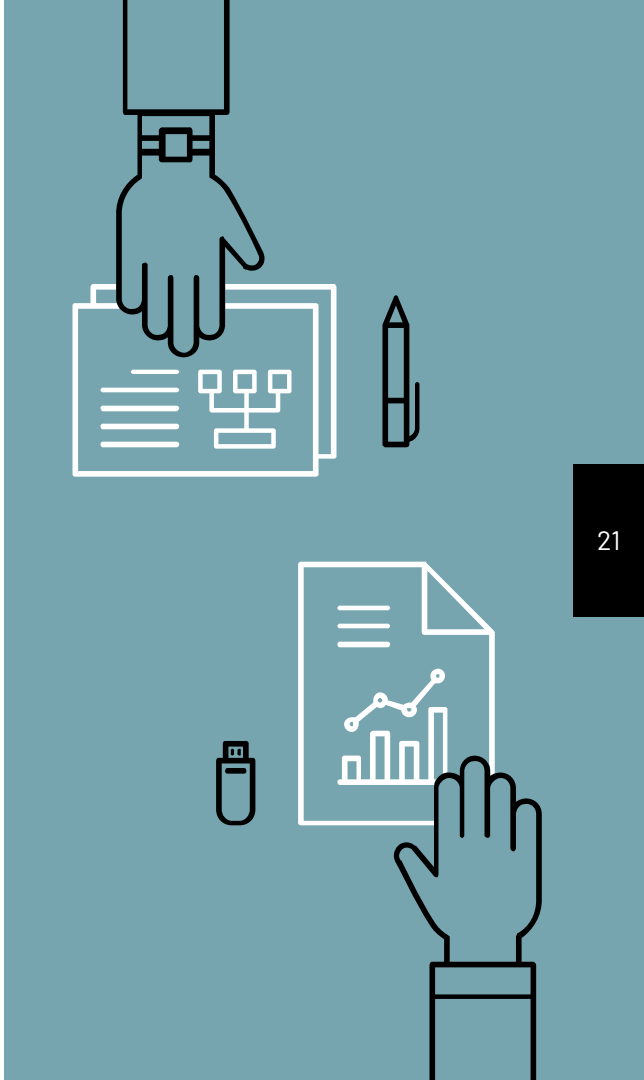
Module 5 Discussion

Do people still do cross validation if they have a replication dataset?

- If I understand the question correctly, yes. Cross validation is like having multiple validation sets to test the performance and generalizability of a model. A replication dataset is for replicating the results of the model on a completely different dataset.

In the examples they split the dataset into the number of rows contained in the dataset and have the training dataset made up of every observation except one, and the testing dataset be made up of one observation. What is the advantage to doing this over something like 5-fold cross validation (which will have more than one observation in the test dataset)?

- Leave one out cross validation (LOOCV) is a special case of k-fold cross-validation where k is equal to the size of data (n). LOOCV is ideal for small datasets (not many rows) because it allows for the largest amount of training data. However, it is prone to higher variance since the validation set is only 1 observation. If you have a larger number of rows, I would suggest using k-fold cross-validation to avoid this variance issue.



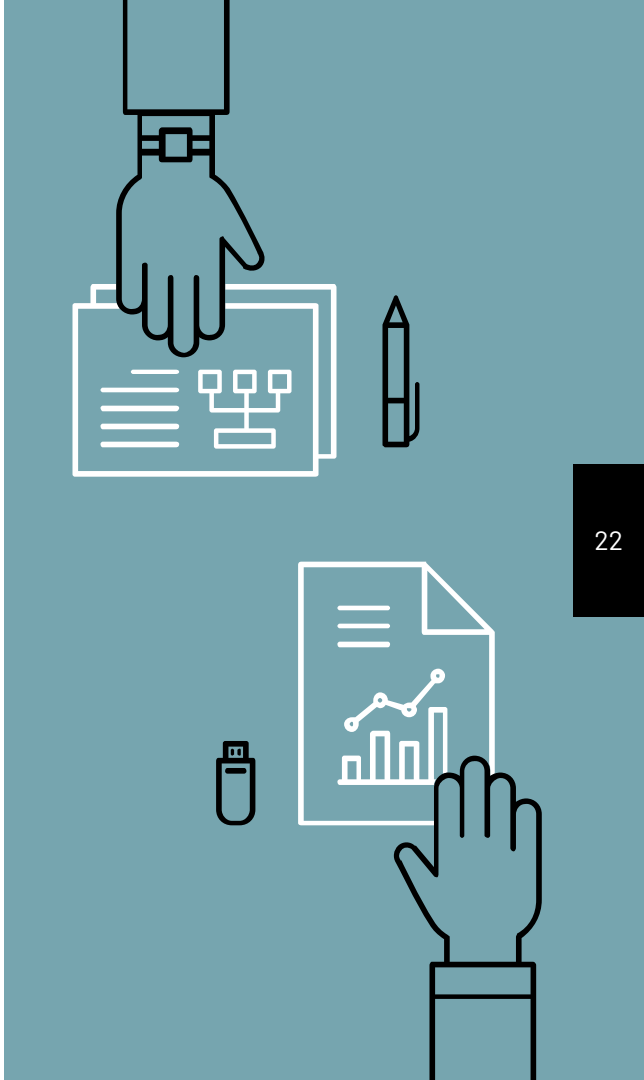
Module 5 Discussion

Although it similar to residual errors, how popular is the brier score in statistics? I haven't heard about it before today.

- In my experience it's more common in machine learning when checking the calibration of a model. I like this post and I think it explains it really well.

The Brier Score and AUC are used for binary outcome variables, but what validation metrics are used for categorical outcome variables that have 3 or more levels?
Are Brier score and ROC only applicable to binary outcomes?

- You can still use the Brier score for categorical outcome variables. The only difference is using 1 for the correct class (category) and 0 if it's anything else (all other categories) - so one category versus all of the others rather than one versus another.



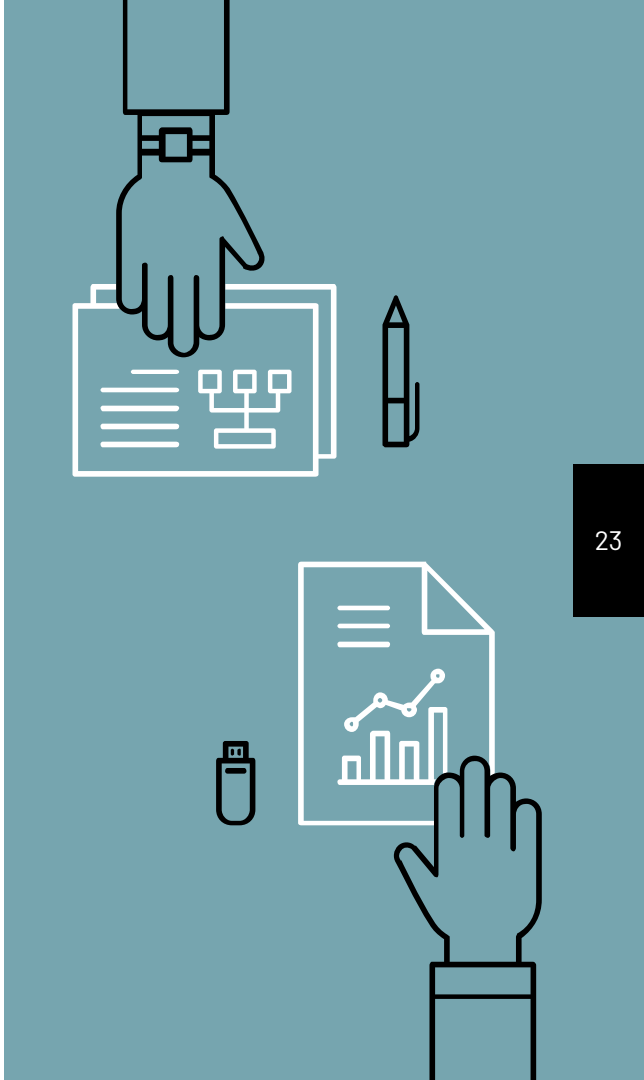
Module 5 Discussion

If you're testing/developing some new method, how should you balance between simulated data resembling a real dataset you're working with vs. satisfying the assumptions you're making about method (even if the dataset might not satisfy them)?

- I've seen researchers create multiple simulated datasets. One to show how their method works when all assumptions are met, and one or more that resemble empirical data or data that doesn't meet all of the assumptions. That way you can also see how your method breaks down when the assumptions aren't met.

Can we go over the intuition behind the estimate of concordance measure?

- Concordance, or the C-statistic, is interpreted as the probability that a randomly selected subject who experienced the outcome will have a higher predicted probability of having the outcome occur than a randomly selected subject who did not experience the outcome.
- I like this resource



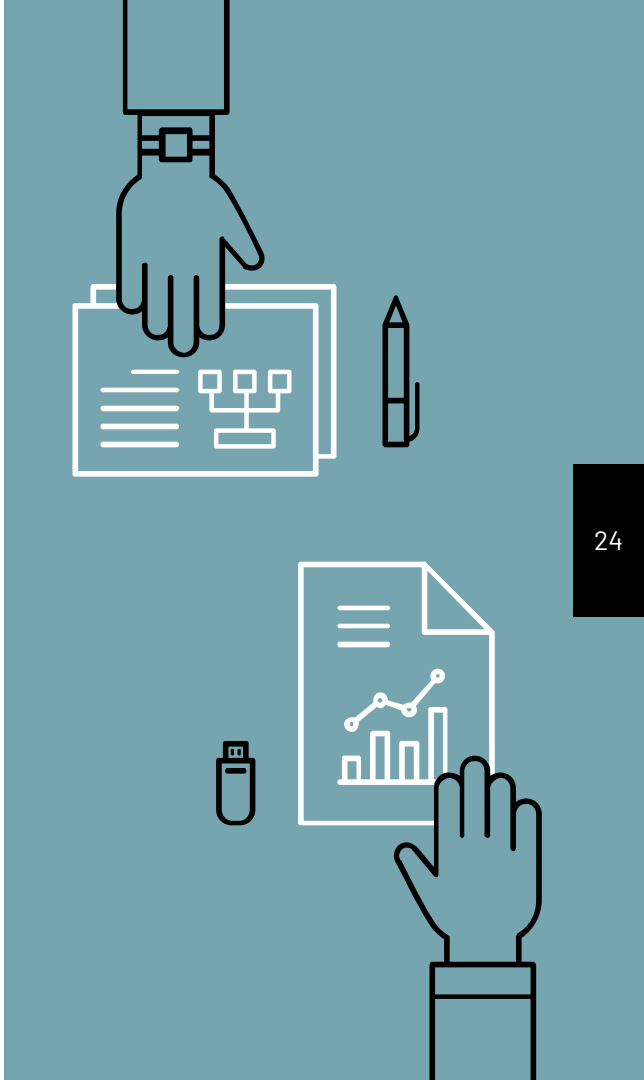
Module 5 Discussion

Is the only difference between cross validation and cross-study validation the number of datasets used?

- Basically. Cross validation uses one dataset and splits it into some number *k* groups for validation. Cross-study validation uses multiple independent datasets for validation (similar to replication).

Do you have strategies for communicating stats measures (AUC, p-values, regression outputs) to non-stats people who are probably more interested in the scientific result rather than the math behind it? How much of the math do you explain?

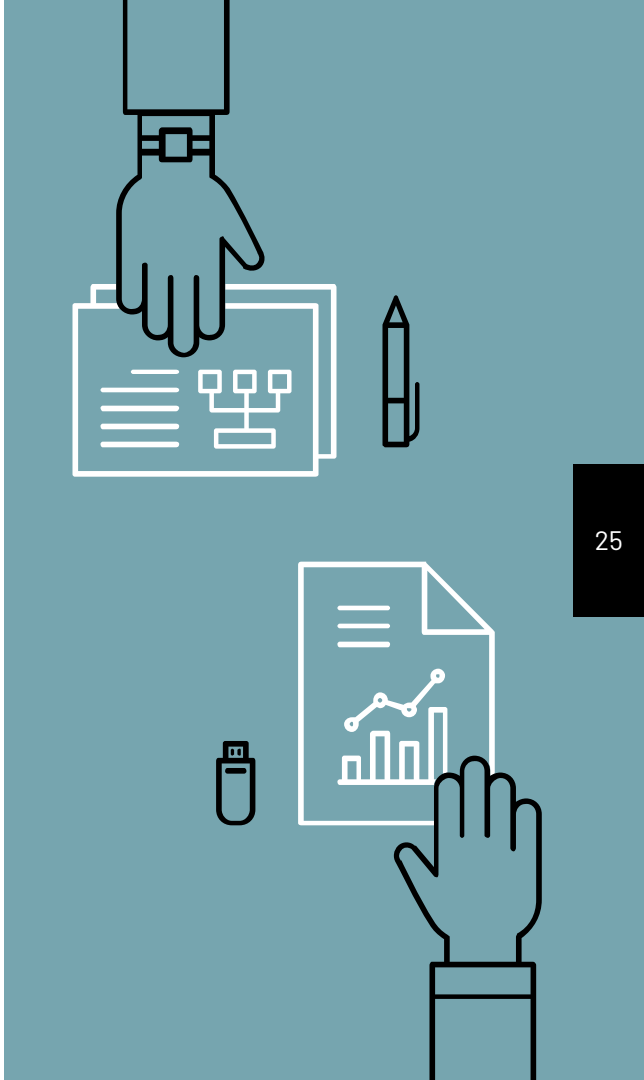
- This is such an important skill! And it takes years to master. If I am explaining this to clinicians or a very non-technical audience I don't mention/explain any of the math and just say what the outcome/result means in the context of the problem. For example: "The AUC for our model is .9, meaning our model is better at predicting cardiovascular disease than the standard model." (In this example the other AUC was .6)
- If the audience is a bit more technical, I might go into more detail but still summarize results without using a ton of jargon.

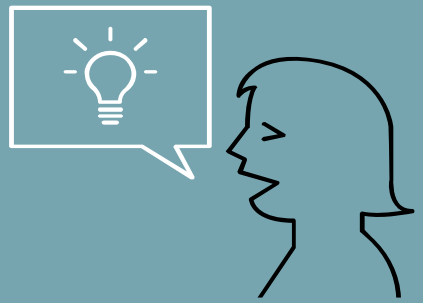


Module 5 Discussion

What exactly are the pros and cons of using the AUC vs. Brier score for binary outcomes?

- AUC is more common so people will usually know what you're talking about, and the Brier score isn't as well known. AUC is also a little easier to interpret/understand, but the Brier score gives more information about model performance. I personally like to report both when applicable.





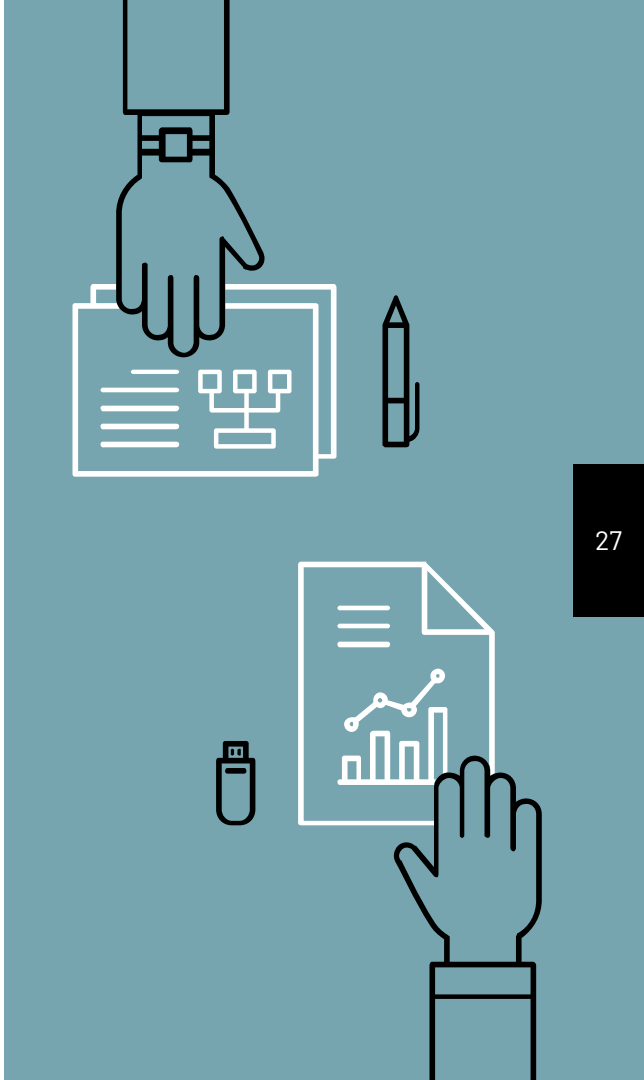
In-Class Project

We will attempt to reproduce the results from and critique the reproducibility of:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.

Specific Tasks:

- Create a data dictionary
- Wrangle data
- Recreate Figure 1
- Recreate Tables 1-5
- Critique reproducibility

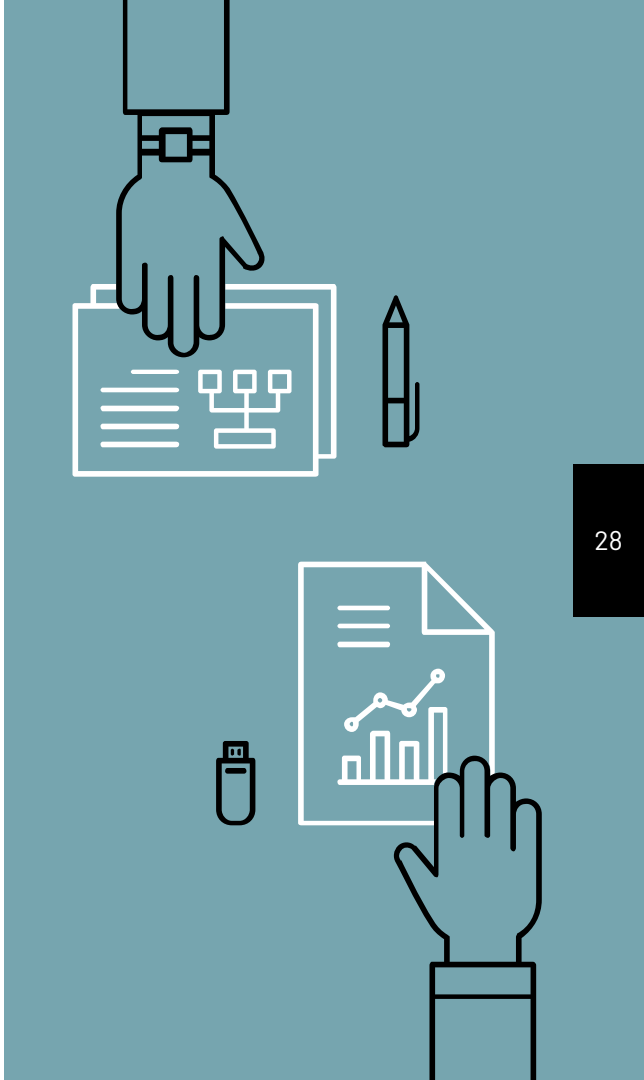


We will attempt to reproduce the results from and critique the reproducibility of:

[1] Boehm, J. K., Williams, D. R., Rimm, E. B., Ryff, C., \& Kubzansky, L. D. (2013). Relation between Optimism and Lipids in Midlife. The American Journal of Cardiology, 111(10), 1425-1431.

Specific Tasks:

- Create a data dictionary
- Wrangle data
- Recreate Figure 1
- Recreate Tables 1-5
- Critique reproducibility



Homework

- Watch Module 6 part 1 videos
  - 6.1.1 - 6.3.10

- Submit Module 6 part 1 discussion points