SISCER 2023 Module 5: Evaluation of Biomarkers and Risk Models

Part IV: Combining Biomarkers and Developing Risk Models

July 13-14, 2023

8:30am-Noon PT / 11:30am-3pm ET

Kathleen Kerr, PhD
Professor of Biostatistics
SISCER Director
University of Washington

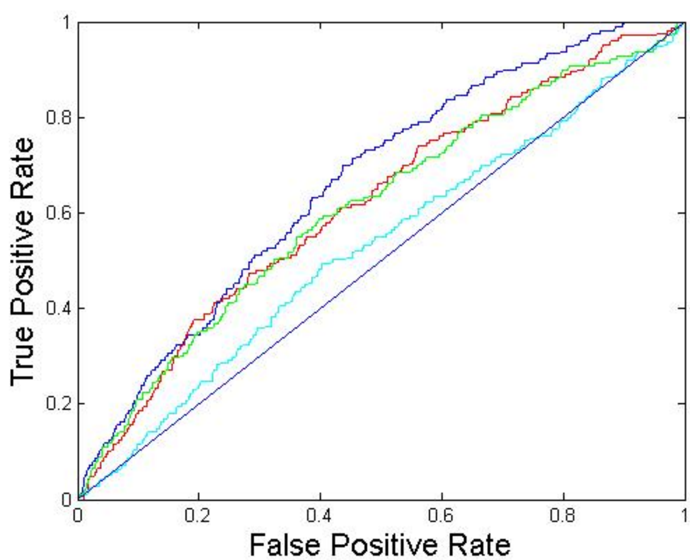
Caveat

- This set of material provides guidance, but does not provide a recipe for developing risk models.

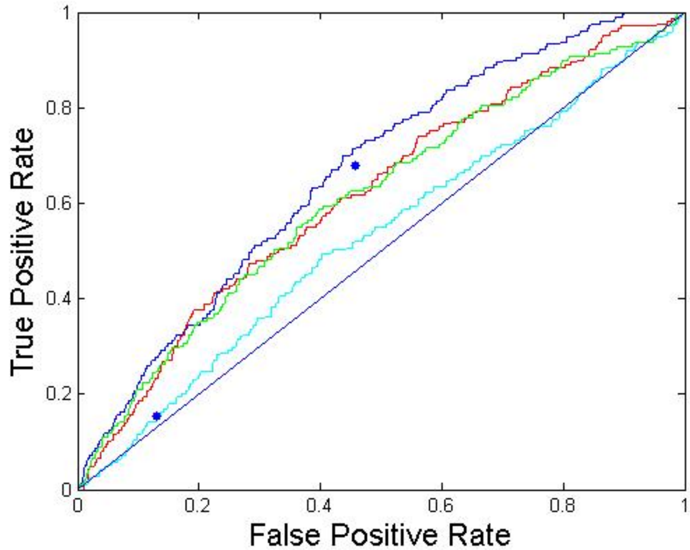
A shared experience

- Investigators interested in predicting an outcome D have a collection of modestly predictive biomarkers
- They combine the markers together with logistic regression. This results in…
- … a modestly predictive combination

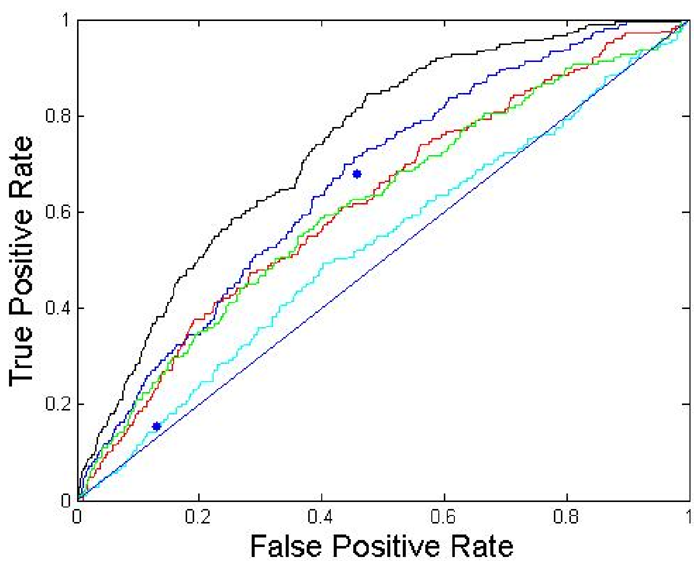
Framingham risk factors individually…



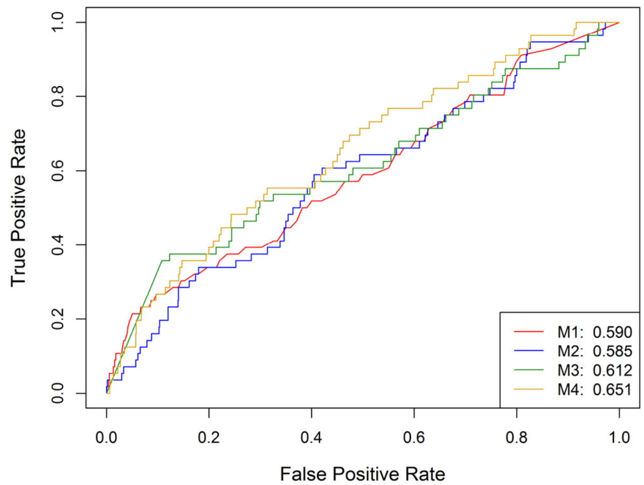
Framingham risk factors individually…



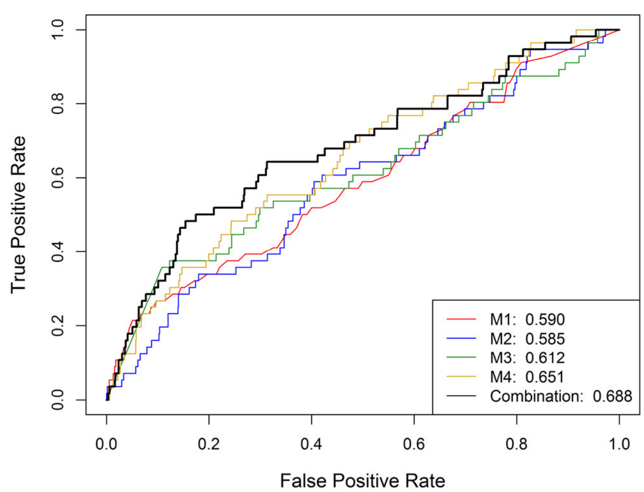
Framingham risk factors in combination



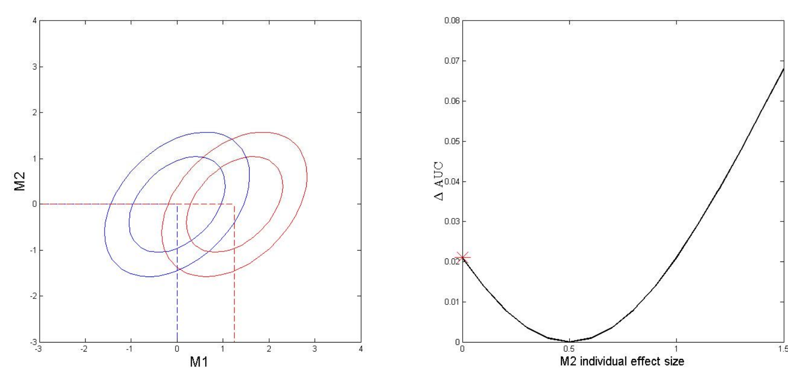
AKI biomarkers individually…

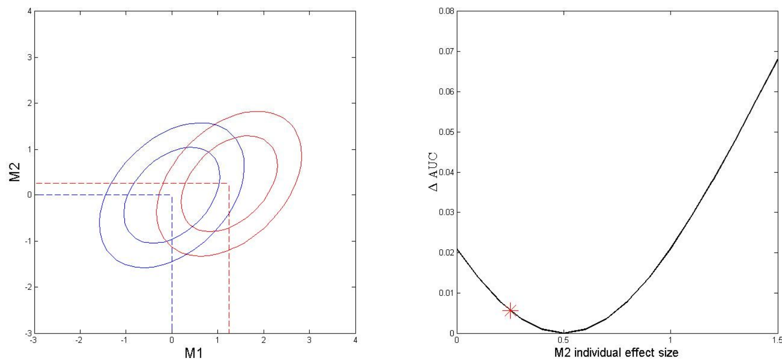


AKI biomarkers in combination

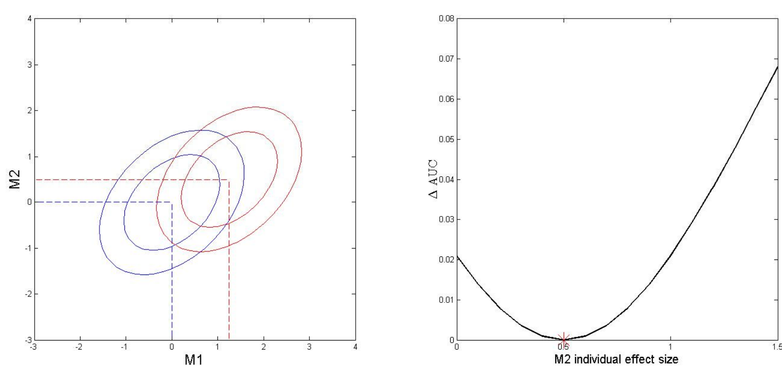


- The previous examples used linear combinations to combine predictors
- Is the problem that we don’t know the right way to combine markers? Should we use something more sophisticated than logistic regression?
- Let’s return to the BiNormal Model

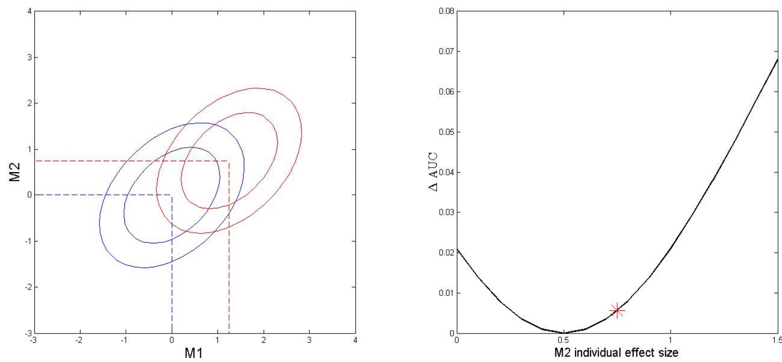


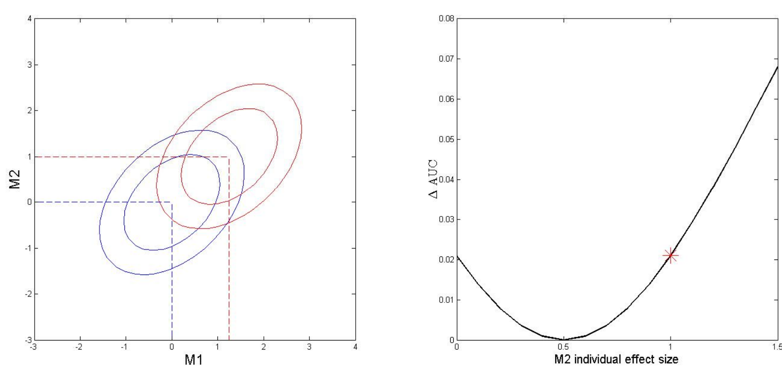

M2
M1
Δ AUC
M2 individual effect size

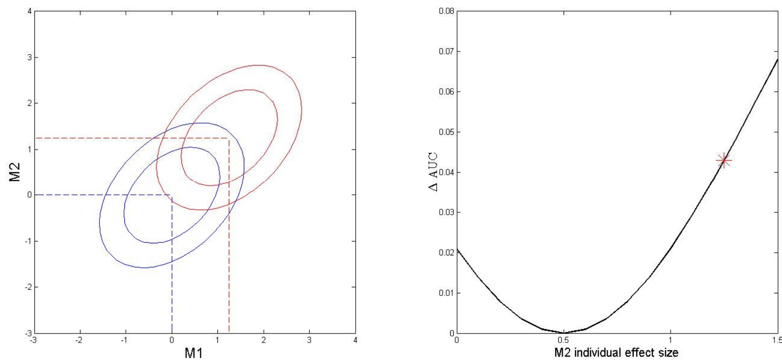
411

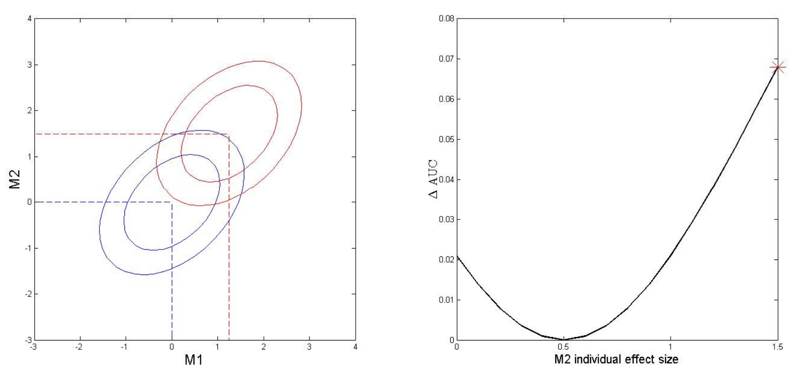

M2
M1
Δ AUC
M2 individual effect size

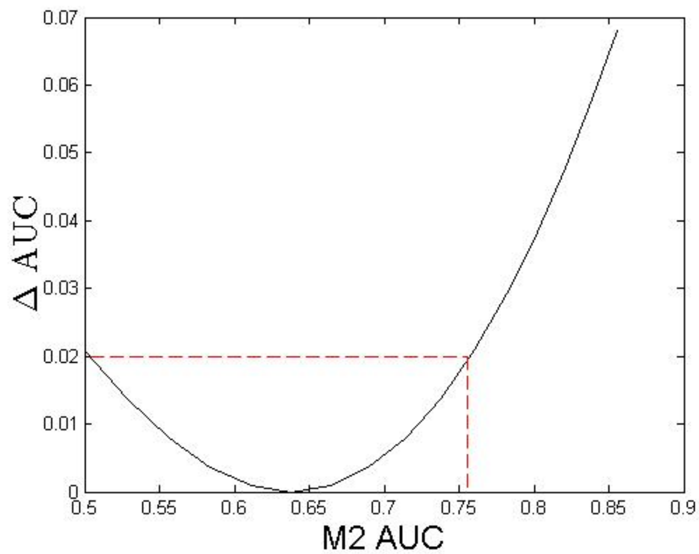
412


M1
M2
Δ AUC
M2 individual effect size


M1
M2
Δ AUC
M2 individual effect size


M2
M1
Δ AUC
M2 individual effect size


M2
M1
Δ AUC
M2 individual effect size



Lessons from the example:

- A marker with no predictive capacity by itself can have positive incremental value.
- A marker with prognostic capacity by itself can have 0 incremental value.
- Incremental value is **not** a monotone function of a biomarker's individual predictive capacity.
- To get large incremental value, we may need new biomarkers that are *as good as or better than* existing markers.

Observations about the example:

- In the example, the true risk scores are known theoretically and exactly
  - risk(D | M1)
  - risk(D | M2)
  - risk(D | M1, M2)
- In particular, we are not *estimating* risk P(D | M1, M2).
- Conclusion: “better methods for combining biomarkers” is not what is lacking in this example

New example

- X1 has mean 0 in controls and mean 1 in cases. SD_{X1}=1 in both.
- X2 has mean 0 in controls and mean 2 in cases, SD_{X2}=1 in both.
- We consider the optimal combination of X1 and X2 for discriminating cases and controls When will we have highest AUC for the combination?

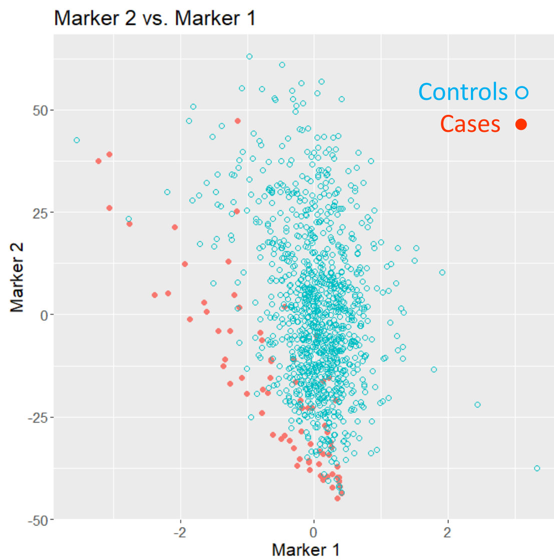
$corr_{cntl}$(X1, X2)=$corr_{case}$(X1, X2)= 0, 0.3, 0.6, or 0.9

(X1, X2) has mean (0,0) in controls and mean (1,2) in cases. Conditional correlation between X1 and X2 is 0, 0.3, 0.6, or 0.9.

(X1, X2) has mean (0,0) in controls and mean (1,2) in cases. Conditional correlation between X1 and X2 is 0, 0.3, 0.6, or 0.9.

Recent real data example

AUC$_{M2}$ 0.75



AUC$_{M1,M2}$: 0.94 after optimism-correction

Figure credit: David Hu

AUC$_{M1}$: 0.71

Lessons from Machine Learning

- Lim et al (2000) compared 33 classification algorithms on 32 datasets
  - 22 algorithms to build decision trees
  - 9 statistical algorithms
  - 2 neural network algorithms
- The best performing algorithm “was not statistically different” from 20 other algorithms.
- Logistic regression came in second

Lessons from Machine Learning

- Christodoulou et al (2019) reviewed published papers that reported both logistic regression and a machine learning technique to develop a predictive model
- For studies using best practices to avoid biased results, no evidence of a systematic benefit for machine learning or logistic regression
  - LR included penalized, “boosted”, and “bagged” versions
  - Evaluative metric: AUC

Lessons from Machine Learning

- There is no universally “optimal” way of combining biomarkers
  - For every method, there is probably some data structure for which it is optimal.

Lessons from Statistics and Machine Learning

- Different methods are optimal for different data structures, so should we try out lots of methods?
  - We should worry about "model selection bias"
  - If we try out lots of methods on our data and choose the best, we will have biased estimates of model performance without special methods
  - For modestly sized datasets in biomedicine, choose a sensible approach (or a few) and move on.

Reporting standards and guidelines for publishing risk models: TRIPOD and RiGoR

Annals of Internal Medicine RESEARCH AND REPORTING METHODS

Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration

Karel G.M. Moons, PhD; Douglas G. Altman, DSc; Johannes B. Reitsma, MD, PhD; John P.A. Ioannidis, MD, DSc; Petra Macaskill, PhD; Ewout W. Steyerberg, PhD; Andrew J. Vickers, PhD; David F. Ransohoff, MD; and Gary S. Collins, PhD

(TRIPOD co-published in 11 journals)

Kerr *et al. Biomarker Research* (2015) 3:2
DOI 10.1186/s40364-014-0027-7

BIOMARKER RESEARCH

REVIEW Open Access

RiGoR: reporting guidelines to address common sources of bias in risk model development

Kathleen F Kerr[1*], Allison Meisner[1], Heather Thiessen-Philbrook[2], Steven G Coca[3] and Chirag R Parikh[4]

TRIPOD

- Response to common problems with risk models presented in the literature
- In some areas, many risk models are being developed (diabetes, prostate cancer) – which should clinicians use?
- This problem is exacerbated by poor reporting.
  - The existence of existing models not acknowledged, new model not compared to existing models
  - Failure to provide information on the actual model (!)
- https://www.tripod-statement.org/

RiGoR

- Similar effort to TRIPOD (less prominent than TRIPOD). More emphasis on addressing sources of bias that can arise in risk model development
- Various terms are used to describe these biases
  - optimistic bias
  - overoptimistic bias
  - overfitting bias
  - selection bias
  - parameter uncertainty bias (Steyerberg)
  - model uncertainty bias (Steyerberg)
- Better to have terms that are descriptive and specific. RiGoR paper proposes “resubstitution bias” and “model-selection bias” for two sources of bias that commonly arise in risk model development

Resubstitution bias

- If the same data are used to fit a risk model and evaluate its performance, the evaluation will be biased in the “optimistic” direction
  - The process of evaluating a model on the dataset used to fit the model has been called “resubstitution”
- If we pre-specify the exact form of our model and use the data only to estimate model parameters, then only resubstitution bias is a concern. Methods to correct for resubstitution bias may assume this is the situation.
- There are methods to correct for resubstitution bias:
  - cross-validation.
    - Note that cross-validation does not actually assess the final, fitted model
  - bootstrapping
  - Harrell, *Regression Modeling Strategies* text and `rms` R package: “optimism-corrected AUC” etc. [R demo]

Model-selection bias

- Often we also use the data to help us choose our model
  - which variables to include in the model
  - transformations of those variables
  - form of the model (square terms, interaction terms)
- Even if we correct for resubstitution bias in our evaluation of the final model, we can still have model-selection bias

Model-selection bias

- Methods here are less-developed
- If using bootstrapping or cross-validation, a common practice is to incorporate model-selection into the procedure
  - not entirely clear how well this works
  - requires a completely algorithmic method of model-selection

Sample-splitting

- Randomly split the data into a training set and a test set (often 50-50, or 2/3-1/3)
  - all model development on the training set
  - when the final model is "locked down", evaluate its performance on the test set
  - addresses both resubstitution bias and model-selection bias
- Criticized for its statistical inefficiency
  - only using a fraction of the data to build/train your model
  - still, if you have lots of data this might be a good option
- Allows flexibility in developing the model as long as the test data are preserved for testing
    - No iteration allowed - next slide

Sample-splitting

- In order for sample-splitting to provide an unbiased assessment of model performance, you get “one look” at the test data
- Must “lock down” one or a few models to evaluate on the test data
- If you evaluate a model on the test data, then re-visit the training data to try to come up with a better model, you are no longer getting an unbiased assessment
  - the test data are informing model development, are no longer independent

Internal vs. External Validation

- All of the methods just discussed are methods of “internal” model validation
- “external” validation is a more challenging and more important hurdle: how does the model perform on a new sample of data from the appropriate clinical population?

Bootstrap Approach to Correcting for Resubstitution Bias

- "optimism-corrected estimate of model performance"
- Harrell text: "bias-corrected or overfitting-corrected estimate of predictive accuracy"
- (Illustrated in R Demo)

Bootstrap Approach to Correcting for Resubstitution Bias

1. Fit the (pre-specified) model (call it M) and calculate its performance on the same dataset.
   – "apparent performance" of M
2. Draw a bootstrap sample of size n. Re-fit the model to the bootstrap sample, get M*.
3. Evaluate M* on both the original dataset and the bootstrap dataset used to get M*. The difference between these is the estimate of optimism.
4. Repeat steps 2-3 many times. The average of the estimated optimisms across many bootstrap samples is the estimate of optimism. Subtract the estimated optimism from the apparent estimate of performance.

Summary

- There is no generally optimal way to build a prediction model or risk model
- Logistic regression has been observed to work well in lots of settings
  - need special methods for high-dimensional settings, not addressed here
- The variable that is most predictive on its own will not necessarily offer the most improvement to an existing risk model
- To improve upon an existing risk model we should not necessarily seek markers that are independent of existing markers

Summary

- Risk models are often poorly reported in the literature. Consult reporting standards (TRIPOD, RiGoR)
- Beware of optimistic biases in risk model development: resubstitution bias and model-selection bias
  - There are additional opportunities for biases to enter a study, e.g. selection of cases and controls

References

- Bansal and Pepe, When does combining markers improve classification performance and what are implications for practice? *Statistics in Medicine*, 2013.
- McIntosh and Pepe, Combining several screening tests: optimality of the risk score. *Biometrics*, 2002.
- Lim, Loh, and Shih, A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine Learning*, 2000.
- Chrisodoulou, Ma, Collins, Steyerberg, Verbakel, van Calster, A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction model. *J of Clinical Epidemiology* 2019
- Gary Collins et al, TRIPOD papers and website 2015
- Kerr et al, RiGoR, *Biomarker Research* 2015
- Harrell, Regression Modeling Strategies, Springer

Misconceptions about Biomarkers and Risk Models

- A large odds ratio implies that a biomarker is useful for prediction.
- A data analyst can identify the optimal threshold from an ROC curve.
- A data analyst can identify the optimal risk threshold from a Decision Curve.
- The best biomarker to improve a risk model is the one with strongest association with the outcome.
- To improve prediction, a new biomarker should be independent of existing predictors
- To assess whether to add new biomarker to a risk model, multiple stages of hypothesis testing are needed.
- We can often use biomarkers to identify which patients will benefit from treatment.

Misconceptions about Biomarkers and Risk Models

- A large odds ratio means a biomarker is useful for prediction.
- ROC curves are useful to identify the best biomarker cut-point.
- Decision curves are useful to identify the best risk threshold.
- To assess whether to add new biomarker to a risk model, multiple stages of hypothesis testing are needed.
- The best biomarker to improve a risk model is the one with strongest association with the outcome.
- To improve prediction, a new biomarker should be independent of existing predictors.
- We can often use biomarkers to identify which patients will benefit from treatment.