SISCR Module 7

Part IV:
Notes on Combining Biomarkers and Developing Risk Models

Kathleen Kerr, Ph.D.
Associate Professor
Department of Biostatistics
University of Washington

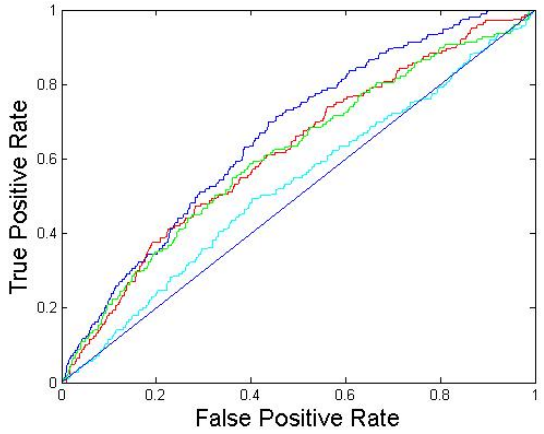
Caveat

- This set of material should provide you with some guidance, but will not provide you with a recipe.

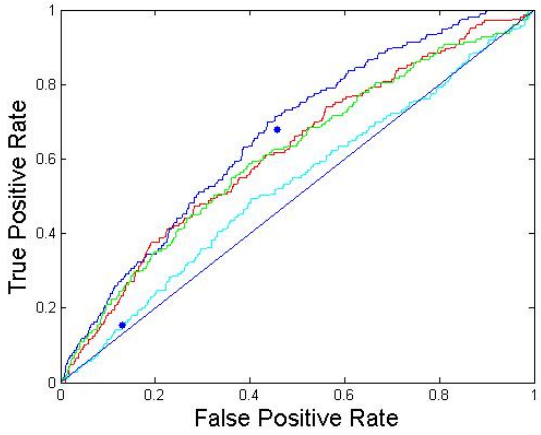
A shared experience

- Investigators interested in predicting a binary 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…

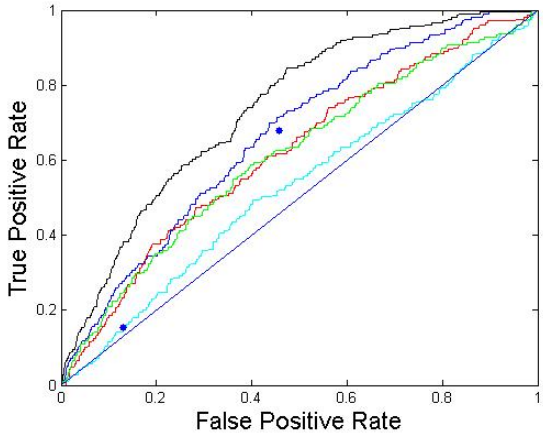


Framingham risk factors individually…



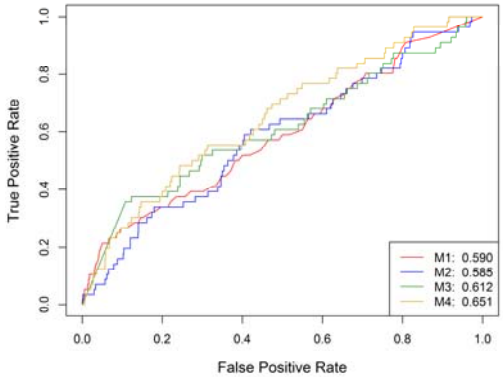
405

Framingham risk factors in combination



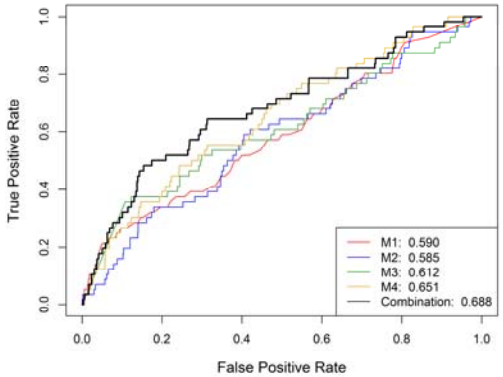
406

AKI biomarkers individually…



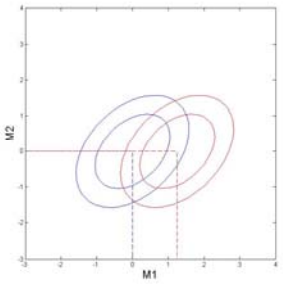
407

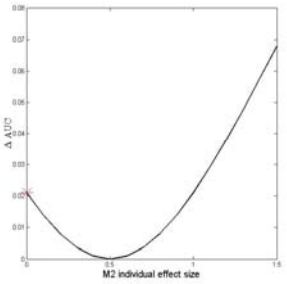
AKI biomarkers in combination

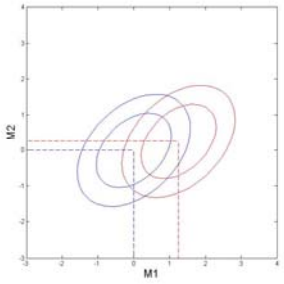


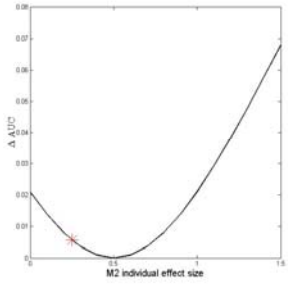
408

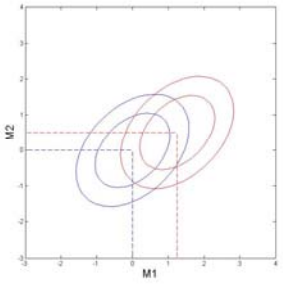
- The previous examples used linear combinations to combine predictors
- Is the problem that we don't know the right way to combine markers?
- Let's return to our BiNormal Model

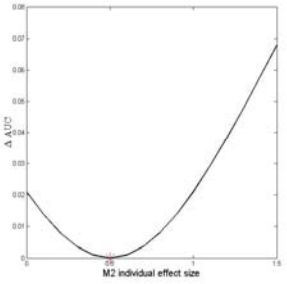


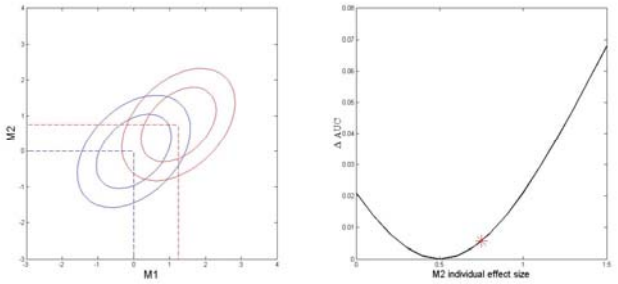




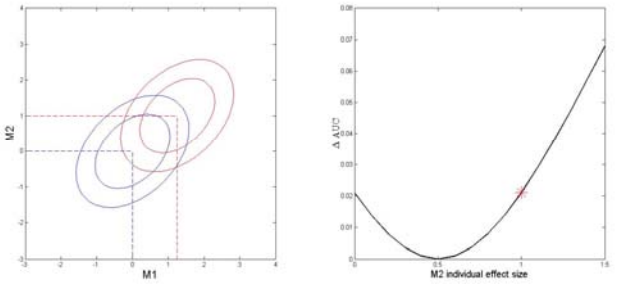




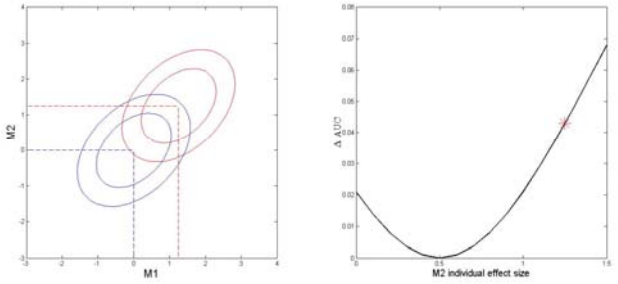




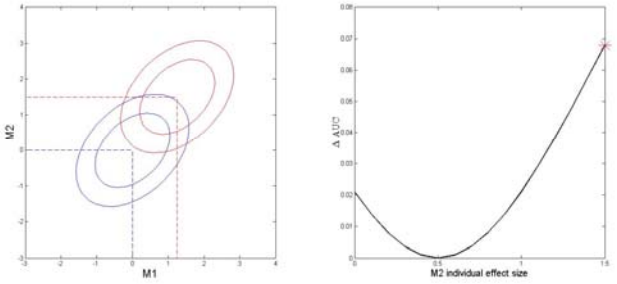
413



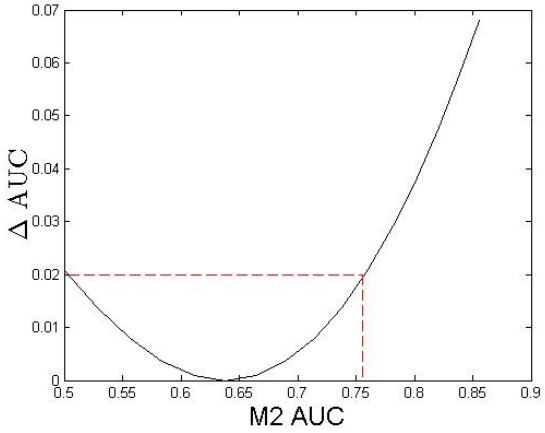
414



415



416



Lessons from the example:

- A marker with no predictive capacity by itself can have positive incremental value.
- Incremental value is **not** a monotone function of marginal predictive capacity.
- To get large incremental value, we may need new biomarkers that are nearly as strong as 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(D | M1, M2).
- Conclusion: "better methods for combining biomarkers" is not what is lacking in this example

Lessons from Machine Learning

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

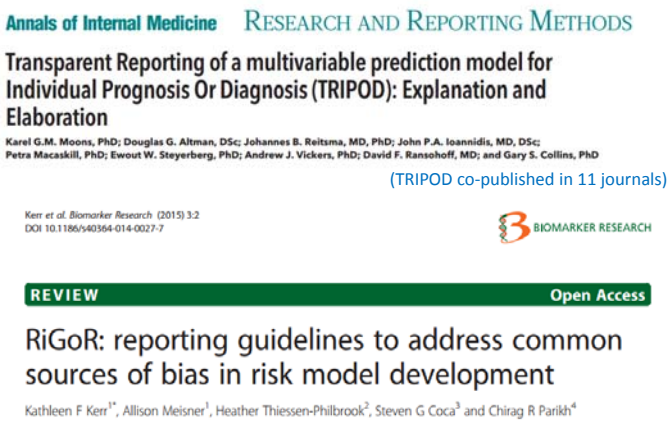
Lessons from Machine Learning

- There is no such thing as the "optimal" classifier
  - For every classifier, there is 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 something sensible and move on.

Recent efforts to provide reporting standards and guidelines for publications reporting new 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), making it challenging for clinicians to decide which one to 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
- Focus is addressing possible 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

- The RiGoR paper proposes the terms "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 applying a model to the dataset used to fit the model has been called "resubstitution", hence our name for this source of bias
  - Fairly extensive set of methods exist to correct for this bias when evaluating a risk model
    - bootstrapping
    - cross-validation
    - Harrell, *Regression Modeling Strategies* text: "optimism-corrected AUC" etc

Model-selection bias

- If we pre-specify the exact form for our prediction model, and use the data only to estimate model parameters, then only resubstitution bias is a concern
- More likely we used the data to help us choose our model
  - transformations of our variables
  - what variables to include in the model
  - 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
  - note that it doesn't actually assess the final, fitted model

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 the best way to go

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
  - Now the test data are informing model development

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?

Summary

- There is no general "optimal" way to build a prediction model
- Logistic regression has been observed to work well in lots of settings
  - NB: 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, recently-proposed standards to address this (TRIPOD, RiGoR)
- Beware of optimistic biases in risk model development: resubstitution bias and model-selection bias
  - There are plenty of other 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.
- Gary Collins et al, TRIPOD papers and website 2015
- Kerr et al, RiGoR, *Biomarker Research* 2015
- Harrell, Regression Modeling Strategies, Springer