SISCER Module 3

Part IV:

A. Combining Biomarkers and Developing Risk Models;

B. Setting Target performance for Early Phase Biomarker Research

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

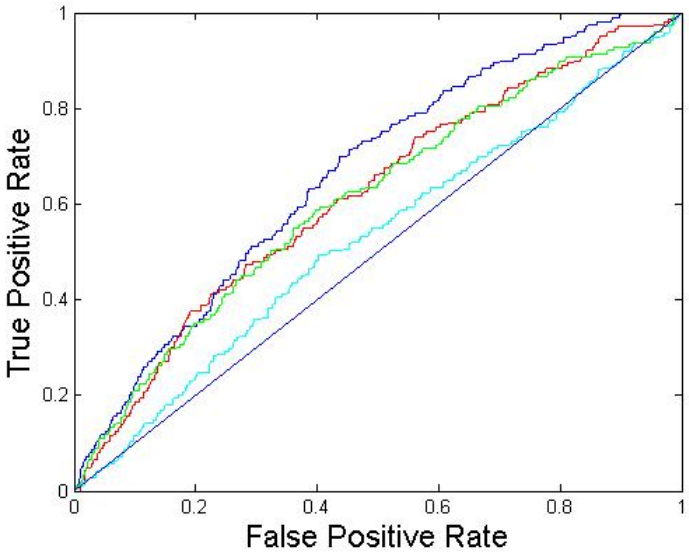
Caveat

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

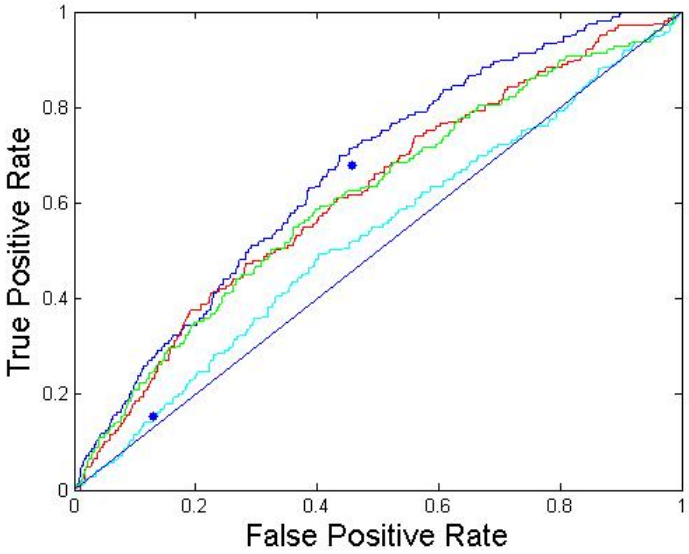
A shared experience

- Investigators interested in predicting 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…

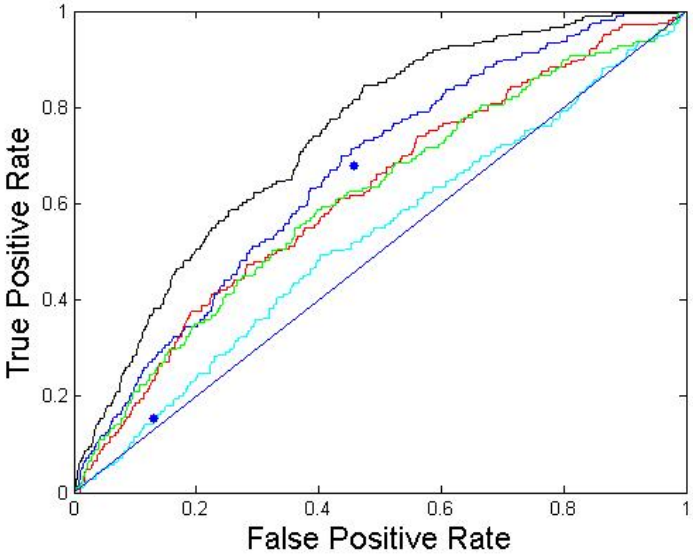


Framingham risk factors individually…



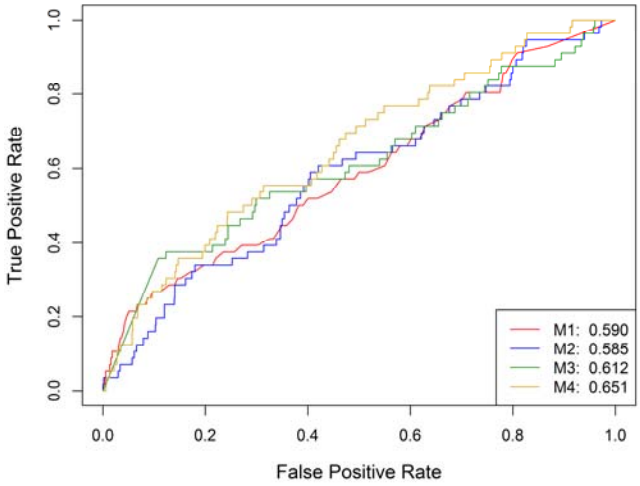
405

Framingham risk factors in combination

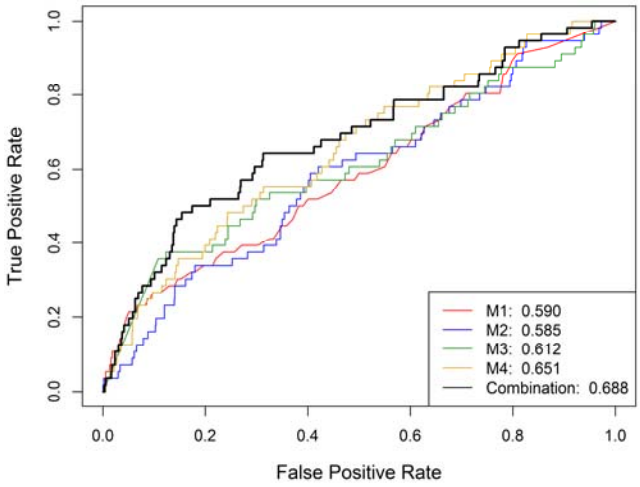


406

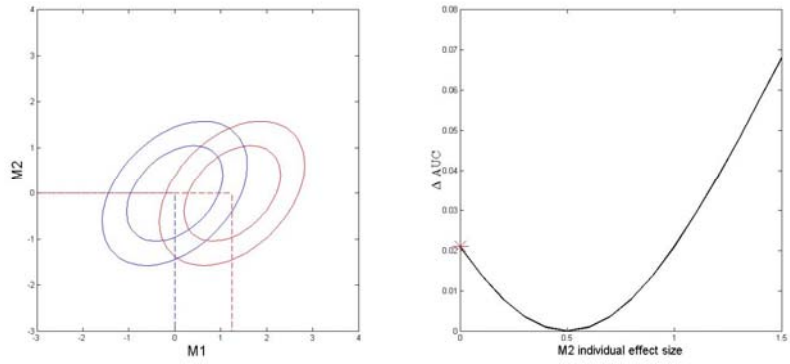
AKI biomarkers individually…

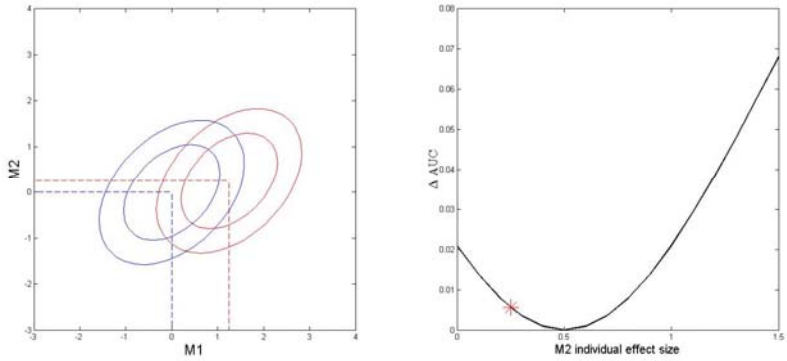


AKI biomarkers in combination

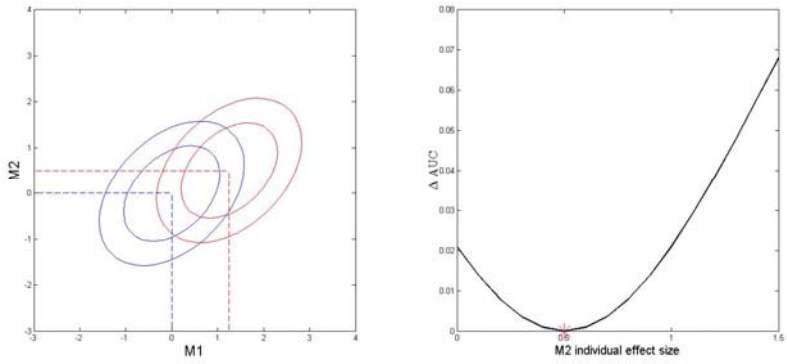


- The previous examples used linear combinations to combine predictors
- Is the problem that we don't know the right way to combine markers? Should we use something more sophisticated than logistic regression?
- Let's return to our BiNormal Model

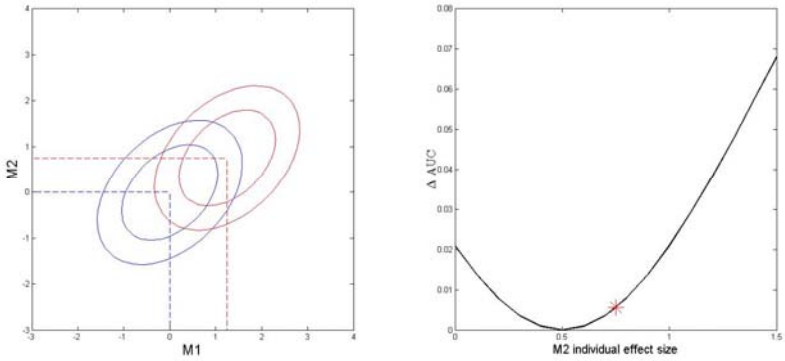



M2
M1
Δ AUC
M2 individual effect size

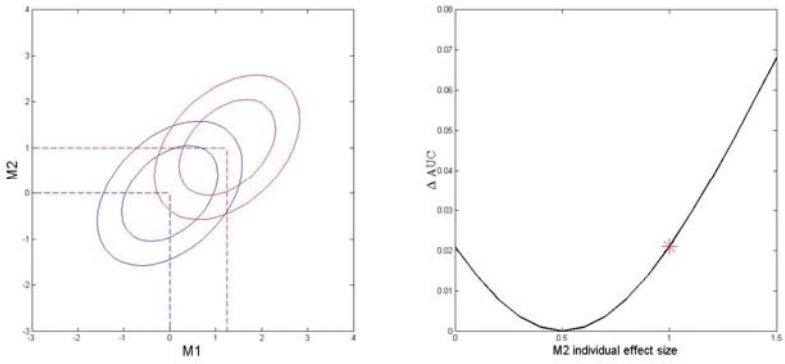
411


M2
M1
Δ AUC
M2 individual effect size

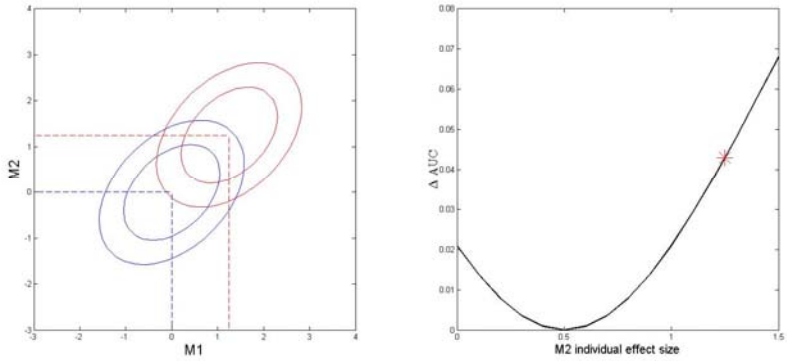
412


M2
M1
Δ AUC
M2 individual effect size

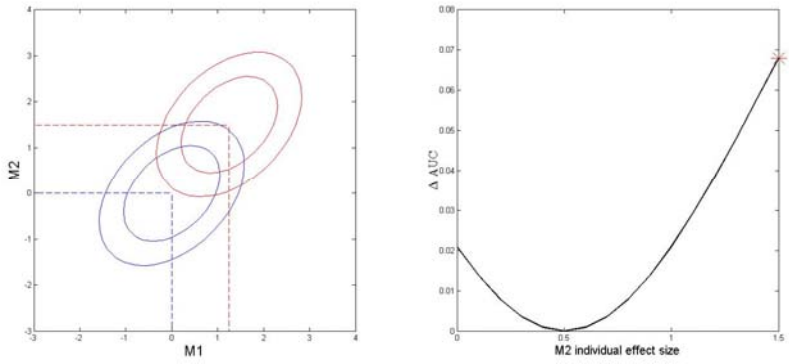
413


M2
M1
Δ AUC
M2 individual effect size

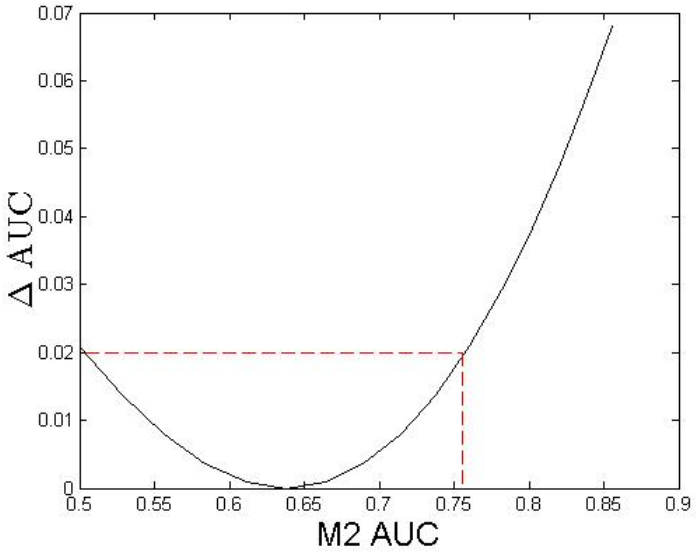
414


M2
M1
Δ AUC
M2 individual effect size

415


M2
M1
Δ AUC
M2 individual effect size

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 as good 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

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 bias results, no evidence of a systematic benefit for machhine 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 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"
  - Fairly extensive set of methods exist to correct for this bias when evaluating a risk model
    - bootstrapping
    - cross-validation
    - Harrell, *Regression Modeling Strategies* text and `rms` R package: "optimism-corrected AUC" etc
      - R demo

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 option

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?

One Method for 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)

One Method for 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 general "optimal" way to build a prediction 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 to do better (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 for part A

- 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

Early-Phase Biomarker Research

- Early phase biomarker research project
  - "We seek a biomarker with 80% sensitivity and 90% specificity"…. What makes a reasonable goal?
  - We can borrow principles from risk model assessment to inform and set performance targets

Early-Phase Biomarker Research

- If the marker is used to direct a clinical decision about an intervention, the context can help set performance standards
- Example
  - Seek a biomarker to select women for mammography
  - Let B be the benefit of mammography to a women with undiagnosed breast cancer
  - Let C be the cost/harms of mammography to a women without breast cancer
  - Let ρ be the prevalence of undiagnosed breast cancer in the target population
  - The total benefit derives from positive tests in cases: $\rho \cdot TPR \cdot B$
  - The total harms derives from positive tests in controls: $(1-\rho) \cdot FPR \cdot C$

Early-Phase Biomarker Research

- For the marker to have net positive value:

$$\rho \cdot TPR \cdot B > (1-\rho) \cdot FPR \cdot C$$

i.e., $\frac{TPR}{FPR} > \frac{1-\rho}{\rho}\frac{C}{B} = \frac{1-\rho}{\rho} r$, where r is the Cost/Benefit ratio $\frac{C}{B}$.

Specifying or soliciting $\mathrm{r} = \frac{C}{B}$ is difficult

Intuitive Measures of the Cost/Benefit Ratio

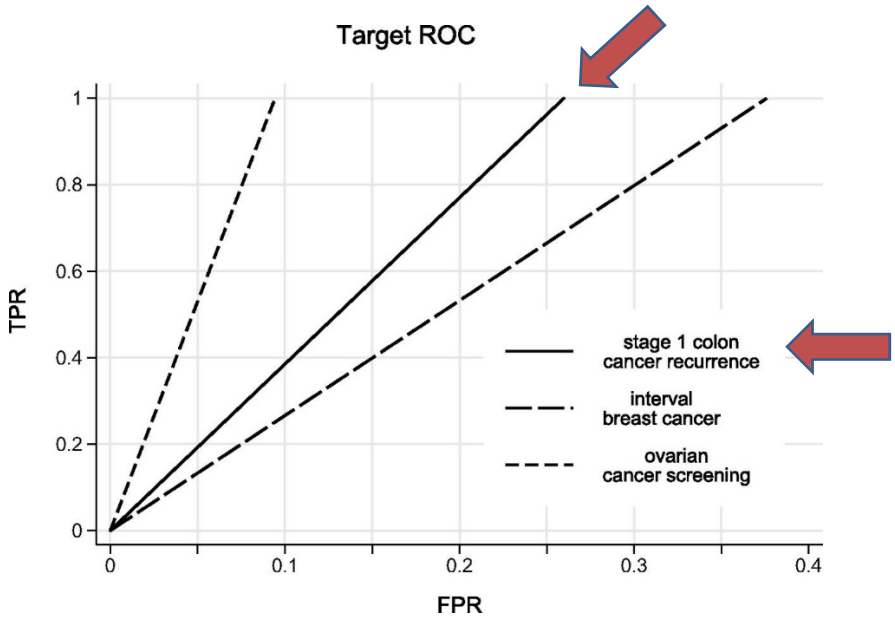
- One can articulate r=C/B in terms of the maximum number of controls N_{max} we are willing to work up in order to receive the benefit of working up one case.
- The cost of working up N_{max} controls is $N_{max} \cdot C$.
- From the definition of N_{max}: $N_{max} \cdot C = B$.
- So $\mathrm{r} = \frac{C}{B} = \frac{1}{N_{max}}$

Intuitive Measures of the Cost/Benefit Ratio

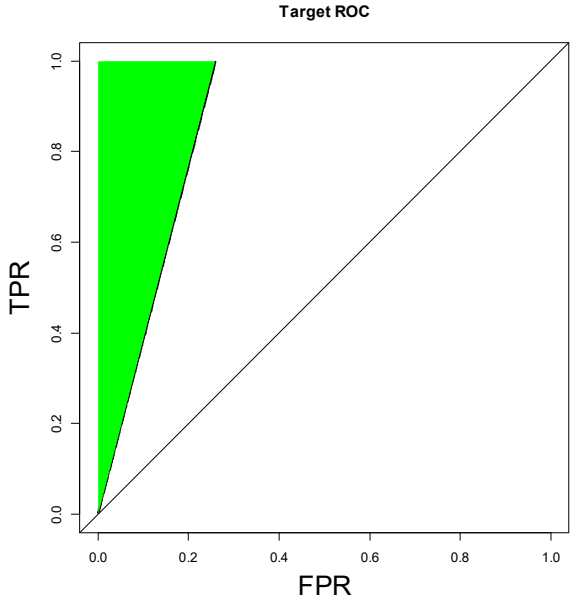
- What is the minimum level of risk R at which work-up is warranted?
- E.g., a woman might feel a mammogram is warranted if her risk or having breast cancer is at least 5/1000 but not if it is less.
- $\frac{C}{B} = \frac{R}{1-R}$

Example: Chemotherapy for Stage 1 Colon Cancer

- Consider biomarker for risk of recurrence within the first year after surgery for stage 1 color cancer patients.
- Stage 1 patients are not normally offered chemotherapy, which would reduce risk of recurrence.
- The 1-year recurrence rate for stage 1 patients is 10% (ρ).
- Stage 3 colon cancer patients are routinely offered chemotherapy; without it, their risk of recurrence is 30%.
- Therefore, R≤30%. If we take R=30%, then $r = \frac{0.3}{1-0.3} = 0.43$.
- Thus $\frac{TPR}{FPR} \geq \frac{1-0.1}{0.1} \times 0.43 = 3.85$.



A marker with a single (FPR, TPR) above the target could have clinical utility.
Since TPR cannot exceed 1, markers with FPR>1/3.85=26% cannot have clinical utility.

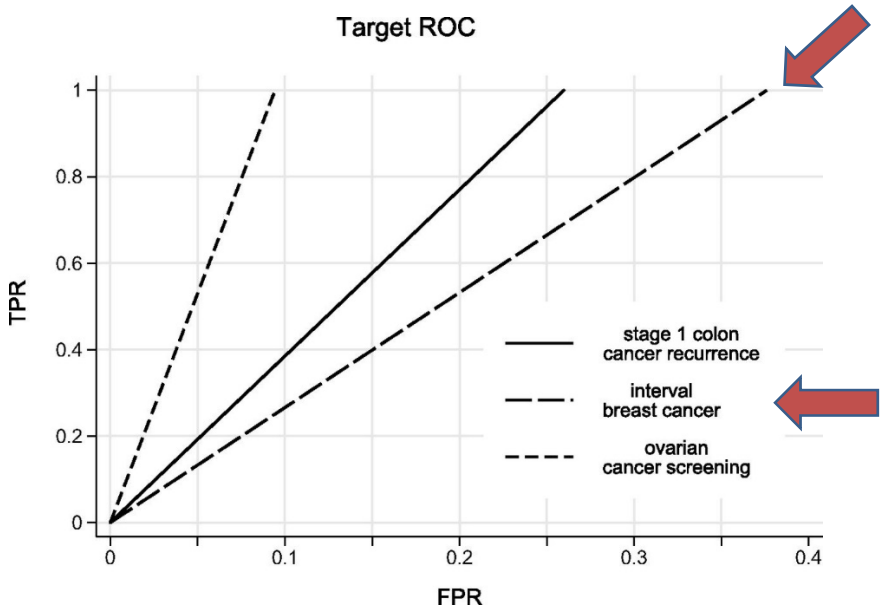


Example: Interval Breast Cancer Screening

- Women 50-74 are recommended for screening mammography every two years.
- Suppose we seek a biomarker to identify women for additional screening (mammograms) 8 and 16 months after a negative mammogram.
- During this interval, the expected incidence of breast cancer is 0.15%.
- A panel decides that the health care system should support 500 additional mammograms (250 women getting 2 "extra" mammograms) to catch 1 woman with interval cancer.
  - N_{max}=250 → $r = \frac{1}{250} \rightarrow \frac{TPR}{FPR} \geq \frac{1-0.0015}{0.0015} \times \frac{1}{250} = 2.66.$

Example: Interval Breast Cancer Screening

- N_{max}=250 → $r = \frac{1}{250} \rightarrow \frac{TPR}{FPR} \geq \frac{1-0.0015}{0.0015} \times \frac{1}{250} = 2.66.$
- If we limit FPR at 5%, then the TPR must exceed 2.66 · 0.05 = 13% for the biomarker to be useful

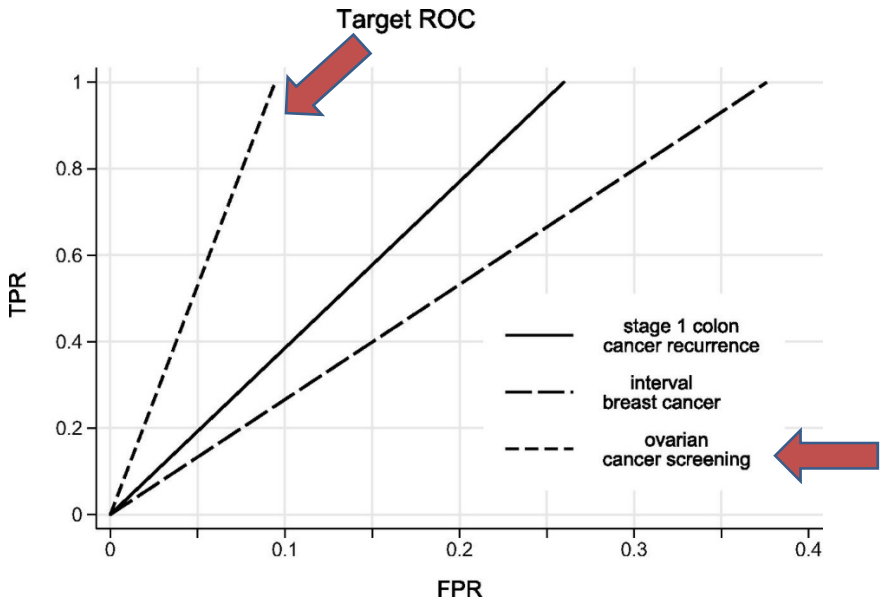


Example: Ovarian Cancer Screening

- Incidence of ovarian cancer in women 50-64 is 25 in 100,000
- We seek a biomarker for annual screening; biomarker positive women will receive surgery for definitive diagnosis.
- We require 1 discovery of ovarian cancer for every 10 surgeries. That is, we tolerate 9 unnecessary surgeries to find one cancer.
- $N_{max}=9 \rightarrow r = \frac{1}{9} \rightarrow \frac{TPR}{FPR} \geq \frac{1-0.00025}{0.00025} \times \frac{1}{9} = 444.$

Example: Ovarian Cancer Screening

- More realistically, marker positive women would receive transvaginal ultrasound to decide on surgery. If TVS is also positive, then surgery.
- If marker results and TVS results are independent (big assumption), then the TPR for the combined test is the TPR for ultrasound (0.755) times the TPR for the marker; the FPR for the combined test is the FPR for ultrasound (0.018) times the FPR for the marker.
- $\frac{0.755 \times TPR}{0.018 \times FPR} \geq 444 \rightarrow \frac{TPR}{FPR} \geq 10.6.$
- A biomarker that detects 80% of cancers must have an FPR ≤ 0.075.



Reference for part B

Clinical Chemistry 62:5
737–742 (2016)

Cancer Diagnostics

Early-Phase Studies of Biomarkers: What Target Sensitivity and Specificity Values Might Confer Clinical Utility?

Margaret S. Pepe,[1*] Holly Janes,[2] Christopher I. Li,[3] Patrick M. Bossuyt,[4] Ziding Feng,[5] and Jørgen Hilden[6]