Study Design

Mark Thornquist, Ph.D.

Co-PI, Data Management and Coordinating Center, EDRN

Fallacy vs. Truth

- Fallacy:
  Lung cancer is a disease of the lung

- Truth:
  Lung cancer is a *lot* of diseases of the lung, all given the same name

- We have no reason to think we could find a single marker that covers all non-malignant diseases of the lung: asbestosis, silicosis, pneumoconiosis, bronchitis, emphysema, infection, …
- Why should we think the situation is any different when it comes to malignant lung diseases?

Treatment vs. Early Detection

- Look at Herceptin:
  - Effective in ~10% of all breast cancers—this can make it sufficient to be FDA approved
- Contrast to a marker with similar performance
  - A marker that detects only 10% of a cancer will probably have great difficulty in being FDA approved
- A partial solution to the treatment problem is a solution; a partial solution to the screening problem is often still just a partial solution

Steps of Biomarker Research

- Step 1 Establish and progressively fill a library with markers each of which has *some* ability to differentiate cases from controls
- Step 2 Periodically, construct a panel from the markers in the library and test the panel
- Step 3 If the panel does not have the desired performance, wait until more markers are deposited in the library and try Step 2 again

Note

- The steps given on the previous slide are not the same as the phases of biomarker research described by Pepe *et al*.

Step 1: Fill Up a Library

- Markers must have *some* discriminating ability
- Measures of the value of a marker:
  - Sensitivity at very high specificity
  - Specificity at very high sensitivity
  - Area or partial area under the ROC curve
- Specificity to the specific cancer site of interest (rather than other cancer sites) is *not* important at this step
- Studies in this step will usually be Phase I studies in the Pepe phase scheme

Notes

- The library will grow over time with markers being added from all over
- Markers may not, and need not, all be measured in the same biological specimen
- The quality of the information about a marker is only as good as the study or studies it has undergone (quality of specimens, level of blinding, adjustment for multiple tests, …); more on this later

Role of an Organization Like Alliance of Glycobiologists in This Step

- Updating the library with the evolving information about markers is best done by a respected central organization (EDRN is doing this through its Biomarker Database)
- Discovery work in Step 1 is best and most efficiently accomplished through the use of consistent, high quality specimens such as a standard specimen reference set maintained by a central organization (EDRN has created such sets for many organ systems)

Step 2: Construct a Panel

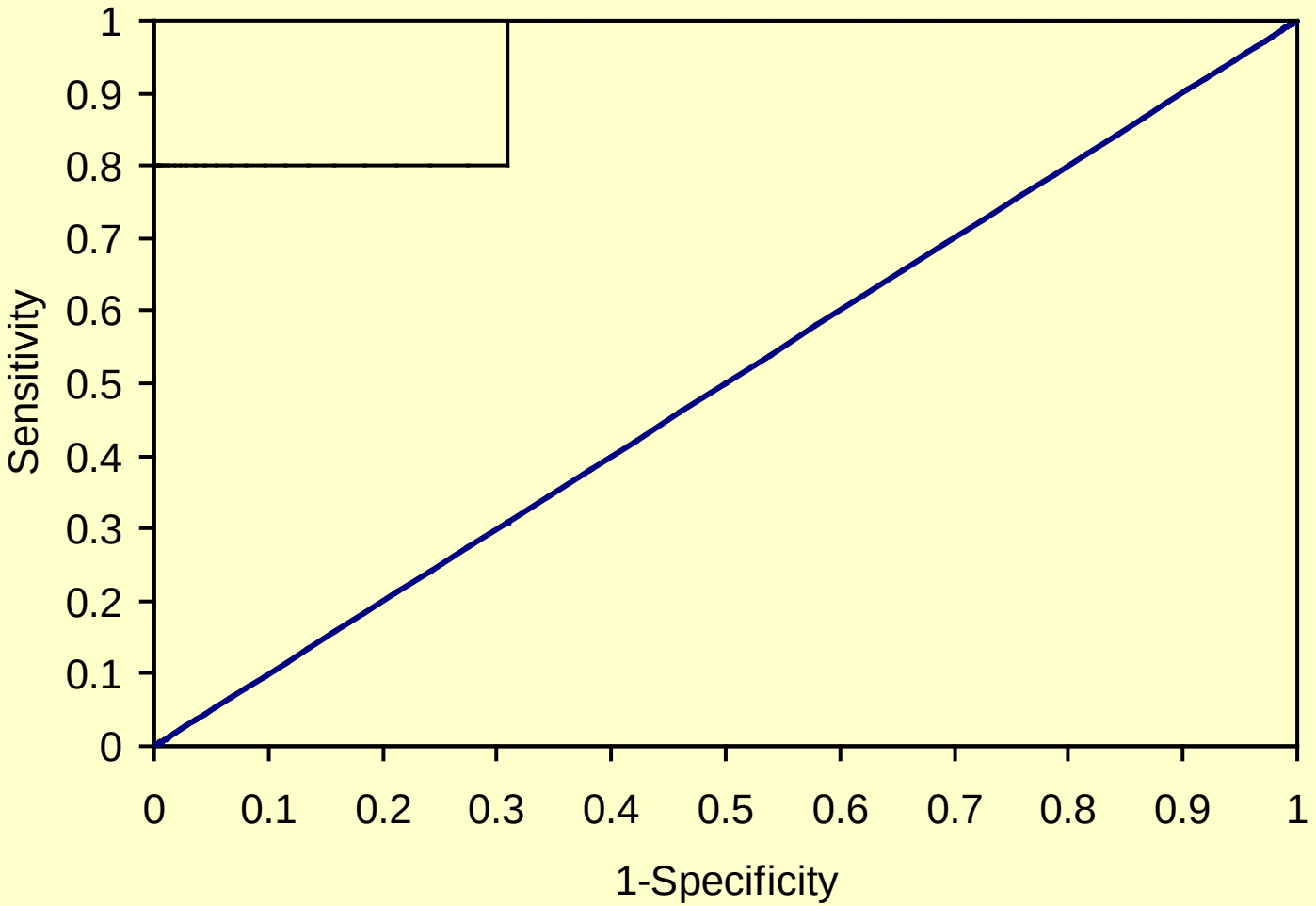
- Panel consists of a set of markers and a decision rule for declaring the status of an individual based on her or his set of marker values
- Panels will be constructed on a training set of specimens and the performance evaluated on an independent test set of specimens
- Studies in this step will typically be Phase II

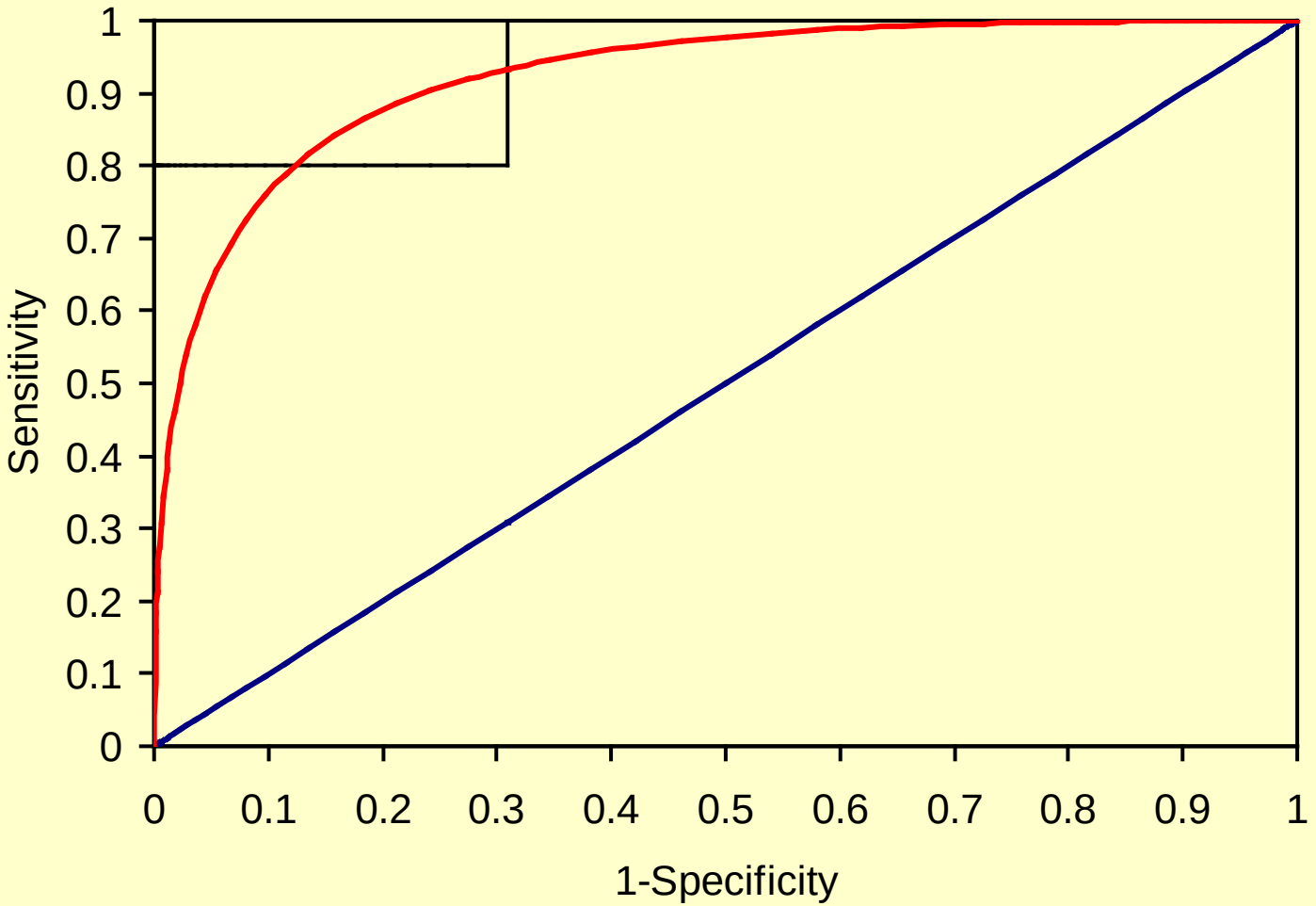
Design of Studies in Step 2

- Parameters of the design (e.g., sensitivity and specificity combination required to declare the panel a success) are driven by the intended screening context
- Specimens must be representative of the population relevant to the screening context
- EDRN has published literature on statistical design issues (e.g., sample size determination) for these studies

Sample Size

- Sensitivity and specificity are estimated from independent samples (cases and controls, respectively)
- So sample size for sensitivity and specificity need to be assessed separately (rather than, say, as a ratio of controls to cases)
- The screening context determines the null and alternative hypotheses for each
- Overall p-value needs to be apportioned between sensitivity and specificity


Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1-Specificity


Sensitivity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1-Specificity

Specimen Selection

- The most common problem seen in EDRN phase I studies is incomparability of case and control specimens
- Unless you are using a design such as the PRoBE design that I'll discuss later, there is a strong likelihood that there may be systematic differences between case and control specimens that may appear as signals in models (particularly if doing unsupervised high-dimensional searching)
- Before proceeding to more expensive studies, it is often a good idea to test potential markers in a new set of specimens obtained from a different source

Cross-Validation

- When describing how cross-validation was done, be sure that all steps in model selection and fitting were cross-validated
  - Selection of markers
  - Fitting of models
  - Exclusion of samples

Cross-Validation vs. Training/Test

- Cross-validation is a valid way of correcting for over-fitting, but I prefer training/test for the following reasons:
  - The models fit in the cross-validations are not the same as the final model, so the interpretation of the results is less clean
  - If there are any systematic differences in the collection/handling of case and control specimens, then cross-validation will not necessarily remove all bias

Preferable Training/Test Design

- Training and test sets are obtained independently from different sources
- Develop model on training set and estimate its sensitivity and specificity using cross-validation (frequently 10-fold cross-validation is used)
- If the cross-validated sensitivity and specificity are sufficiently strong, apply the final model, as is, on the test set
- This design truly tests the portability of the panel

Avoid Work-Up Bias

- A common temptation if there are particular individuals who appear to be outliers in a model (e.g., normals whose marker pattern make them look like cases) is to seek further information about them to see whether they might be misclassified
- But if this is done *only* on the unusual cases, you will have biased your data toward your marker model

Do All Lab Work Blinded

- Enough said

Step 3: Try Again If It Fails

- Let the library accumulate more markers and try again
- Do *not* discard a marker if it failed to contribute to the panel in Step 2

Frequency of Doing Step 2

- Clearly there are time and resource issues with conducting repeated Step 2 studies
- There are also multiple testing issues with conducting repeated Step 2 studies
- Repeating Step 2 preferably done only if there are cogent reasons to think that the additional markers deposited in the library could improve the panel enough to meet the target performance criteria needed for the screening context

Design of Biomarker Studies

- EDRN's experience has been that a lot of markers that have very promising performance in Phase I studies performed very poorly in Phase II studies
- In consequence, EDRN statisticians have developed study design principles for Phase I/II studies to address the causes of the poor agreement between Phase I and II results

The PRoBE Design

Prospective collection of specimens from the target population prior to outcome status ascertainment.

After outcome status ascertained, **R**etrospective random sampling of cases and controls.

Blinded to case-control status when specimens were **E**valuated on biomarker assay.

Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design. JNCI (2008;100: 1432-1438)

Components of PRoBE

- The screening context (drives everything)
- Criteria for biomarker performance
- The biomarker itself
- Study size

Components of Design Related to Screening Context

- Clinical application (settings of sample collection)
- Outcome definition (prospective)
- Case/control status definition (all subjects)
- Selection (random sampling, matching)

Components of Design Related to Performance Criteria

- Sensitivity/specificity (subgroup, time)
- Minimum acceptable performance (cost analysis)
- Comparisons to existing modality

Components of design relating to the biomarker

- Procedures (specimen collection, processing, storage, assay procedures and reporting)
- Blinding (specimen handling, assay, reporting)
- Combinations (algorithm, data collection, blinding)
- Cutoffs do not have to be predetemined but analysis and sample size have to take into account the variation due to estimating cutoff

Study size

- Null hypothesis (minimum acceptable performance)
- Alternative hypothesis (rationale, pilot data)
- Sample size (cutoff)
- Early termination

Compare to Alternative Designs

- Common bias in biomarker studies
  - Systematic differences between cases and controls in subject or specimen collections.
  - Case/control selection or specimen collection was not in the same setting as the intended clinical use.
  - Overfitting
- PRoBE design eliminates these biases.

Compare to simple case-control studies

- Spectrum bias: e.g. cases are more severe and controls are more healthy than the target screening context
  - PRoBE defines all potential cases and controls, randomly selected from each group.
- Knowledge of disease status biases the interpretation of the assay or the handling of specimens or the patient's behavior
  - PRoBE requires storing specimens before outcome ascertainment and **B**linding of disease status.

Compare to prospective studies

- Cost
- Ethical dilemma
- Verification bias
- Knowledge of biomarker values may influence outcome determination

- PRoBE eliminates all above drawbacks (**P**rospective collection, **R**etrospective **B**linded **E**valuation)

Implications for Discovery Studies

- All key elements of the PRoBE design can and should be applied to discovery studies
- Not able to follow all in discovery? Suggest the following exercise:
  - Make a list of the elements that violate PRoBE
  - Ask yourself what could go wrong with each violation?
  - If you are scared, look for ways to satisfy them. Otherwise, in reporting acknowledge those you could not satisfy.
  - When we review biomarkers for validation, we should check these elements, particular for full scale validation studies.

“You are killing my discovery work! How could I get these ideal specimens?”

- The PRoBE design has been presented to the EDRN investigators and PRoBE-design studies are being performed by EDRN
- Biomarker research needs to ***learn what works and what does not work quickly*** (not spend years of discovery using poor quality specimens and then spend years to show this marker didn’t work)
- It is impossible for reviewers to fully figure out whether these principles have been followed. We need to do it by ourselves so we are certain about it.
- Reference sets are useful but we need to expand them.

Questions?