Discovery and Validation of Biomarkers for Cancer: 15 Years of Experience with the Early Detection Research Network

Margaret Sullivan Pepe
Fred Hutchinson Cancer Research Center
What are Biomarkers?

- measured in body tissue or fluids
- diagnosis/screening e.g. PSA
- prognosis e.g. Genomic Health Recurrence Score
- risk prediction e.g. BRCA1 gene mutation
What is the Early Detection Research Network (EDRN)?

Created

- 2000 by NCI
- collaborative network to facilitate bench to bedside

Components

- 22 development + 8 reference laboratories
- 8 clinical validation centers
- data management and coordinating center
- organized around organ-specific collaborative groups
Early Detection of Ovarian Cancer

- symptomatic only in late stage
- hard to treat in late stage
- easy to treat with surgery in early stage
- incidence = 25/100,000
- seek blood based biomarker for ovarian cancer screening
Phases of Biomarker Development

- Biomarker Discovery
- Clinical Validation
- Implementation

Pepe et al. JNCI 2001 93:1054–1061

- Focus initially on design of clinical validation studies.
The research question: How well does biomarker detect presymptomatic ovarian cancer?

**Issues with design**

- biased samples: cases and controls from different settings
- biased samples: preclinical disease not addressed
- \( \text{AUC} = P(Y_{\text{case}} > Y_{\text{control}}) \) is not clinically relevant
- \( \bar{Y}_{\text{case}} - \bar{Y}_{\text{control}} \) is not clinically relevant
Rigorous Design for Clinical Validation

- PRoBE
- **Prospective** enrollment, sample collection and outcome ascertained for a clinically relevant population
- **Retrospective** random selection of cases and controls from the cohort
- **Blinded** specimen handling and assays
- **Evaluation** with relevant statistical methods

Components of the PRoBE Design

(i) Clinical Context
(ii) Clinical Performance Criteria
(iii) Biomarker Test
(iv) Data analysis and sample sizes

Detailed checklists for each aspect (Pepe et al. JNCI 2008 100:1432-1438).
**PRoBE for Ovarian Cancer Screening Biomarkers**

**Clinical Context** (Intended use drives design)

- cohort = healthy asymptomatic women
- definitions
  - case = ovarian cancer 6–18 months from sample
  - control = healthy cancer free 5 years from sample
  - other groups to account for whole population
- consequences of a positive test
  - ultrasound followed by surgery if indicated

⇒ stored blood samples from large healthy cohort, followed prospectively
Clinical Performance Criteria

• $\rho$ = case prevalence = 25/100,000 for age 55–59
  TPR = $P(Y \text{ positive} | \text{ case})$
  FPR = $P(Y \text{ positive} | \text{ control})$

• $B$ = benefit of work-up to a case
  $C$ = cost of work-up to a control

• Expected benefit
  $= B \cdot TPR \rho - C \cdot FPR(1 - \rho) > 0$

• $\frac{TPR}{FPR} > \frac{1 - \rho \cdot C}{\rho \cdot B}$
How to Solicit C/B

**Approach #1:** How many false positives are worth a true positive?

- e.g. 300 mammograms for 1 breast cancer detected

**Approach #2:** Risk Threshold (r)

- expected benefit: \( BP(D = 1|Y) - CP(D = 0|Y) \)
- risk > r \(\Rightarrow\) work-up warranted
  risk < r \(\Rightarrow\) work-up not warranted
  therefore \( Br - C(1 - r) = 0 \)
  \(\Rightarrow\) \( C/B = r/(1 - r) \)

e.g. \( r = 20\% \) \(\Rightarrow\) \( C/B = .20/.80 = 1/4 \)
Application to Ovarian Cancer

• In ovarian cancer: “10 surgeries should yield at least 1 cancer.”

• $r = 0.10$ for the Biomarker + Ultrasound test

• $TPR_{B+US} = P(Y\text{positive and USpositive}|\text{case})$
  $= P(Y\text{positive}|\text{case}) \times P(\text{US positive}|\text{case})$
  $= TPR \times 0.755$

• $FPR_{B+US} = FPR \times 0.018$

\[
\frac{TPR_{B+US}}{FPR_{B+US}} > \frac{1 - \rho}{\rho} \times \frac{r}{1 - r} = \frac{1 - .00025}{.00025} \times \frac{1}{9} = 444
\]

\[
\Rightarrow \frac{TPR}{FPR} > 444 \times \frac{0.018}{0.755} = 10.6
\]
Sample Size Calculations

• notation: $\text{ROC}(f) = \text{TPR}$ corresponding to biomarker positivity threshold that yields $\text{FPR} = f$

• conclude biomarker useful if $\text{ROC}(0.05) \geq 0.53$

• $H_0 : \text{ROC}(0.05) = 53\%$ versus $H_1 : \text{ROC}(0.05) = 0.73$
  – 0.73 is based on preliminary data
  – details in Pepe (2003) textbook

• $n_{cases} = 40$ and $n_{controls} = 160$ yields 71% power
  – Stata software
  – DABS FHCRC website
# Results of EDRN-PLCO Collaborative Study

### ROC(0.05)

<table>
<thead>
<tr>
<th>Marker</th>
<th>Phase 2 preliminary data (160 cases)</th>
<th>≤ 6 months (45 cases)</th>
<th>6 – 12 months (22 cases)</th>
<th>12 – 18 months (17 cases)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA-125</td>
<td>0.73</td>
<td>0.80</td>
<td>0.32</td>
<td>0.12</td>
</tr>
<tr>
<td>HE4</td>
<td>0.57</td>
<td>0.60</td>
<td>0.23</td>
<td>0.06</td>
</tr>
<tr>
<td>MMP7</td>
<td>0.47(?)</td>
<td>0.20</td>
<td>0.14</td>
<td>0.18</td>
</tr>
<tr>
<td>Spondin 2</td>
<td>0.28</td>
<td>0.11</td>
<td>0.14</td>
<td>0.06</td>
</tr>
<tr>
<td>CA72-4</td>
<td>0.40</td>
<td>0.44</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>MIF</td>
<td>0.15</td>
<td>0.18</td>
<td>0.09</td>
<td>0.00</td>
</tr>
</tbody>
</table>

When a Biomarker Test $X$ Already Exists

Examples: PSA, CA-125, mammography

Incremental value

- performance of $(X, Y)$ combined versus $X$ alone
- $\text{ROC}(0.05)$ improved from 0.68 to 0.71

- not possible if $X$ already in use (verification bias)
- cautions: independent data to evaluate improvement versus to combine markers
- use a “proper” statistics e.g., $\Delta \text{ROC}(0.05)$, not NRI
The NRI Statistic can be Misleading

\[ NRI = \{ P(\text{risk}(X, Y) > \text{risk}(X) | \text{case}) - P(\text{risk}(X, Y) < \text{risk}(X) | \text{case}) \} \]
\[ + \{ P(\text{risk}(X, Y) < \text{risk}(X) | \text{control}) - P(\text{risk}(X, Y) > \text{risk}(X) | \text{control}) \} \]

Table: Rates at which the null hypothesis of no performance improvement is rejected in favor of the one-sided alternative hypothesis that prediction is improved by adding the four biomarker panel to the baseline clinical score

<table>
<thead>
<tr>
<th>Dataset</th>
<th>NRI ‡</th>
<th>LR ‡</th>
<th>ΔAUC ‡</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training set (n = 420)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using training set risks, TR-TR</td>
<td>63.0%</td>
<td>5.3%</td>
<td>9.8%</td>
</tr>
<tr>
<td><strong>Test set (n = 420)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using training set risks, TR-TS</td>
<td>23.2%</td>
<td>—</td>
<td>1.1%</td>
</tr>
<tr>
<td>Using re-estimated risks, TS-TS</td>
<td>19.4%</td>
<td>4.7%</td>
<td>1.5%</td>
</tr>
<tr>
<td><strong>Test set (n = 840)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using training set risks, TR-TS</td>
<td>34.4%</td>
<td>—</td>
<td>0.6%</td>
</tr>
<tr>
<td>Using re-estimated risks, TS-TS</td>
<td>18.8%</td>
<td>5.1%</td>
<td>1.8%</td>
</tr>
</tbody>
</table>

* Because the biomarkers have no association with the outcome in the population, all rejections are false-positive results.

† AUC = change in the area under the receiver operating characteristic curve; LR = likelihood ratio; NRI = Net Reclassification Index; TR = training dataset; TS = test dataset.

‡ Five thousand simulated studies in which the biomarkers have no association with outcome. Nominal rejection rates are 5.0%.

Pepe et al JNCI 2014
Use a Clinically Relevant and Valid Statistic

- AUC — not relevant
- NRI — not relevant (usually)
- TPR at pre-specified low FPR — relevant in screening
- FPR at pre-specified high TPR — relevant in diagnosis

- Net Benefit = \( B \times TPR \times \rho - C \times FPR \times (1 - \rho) \)

Standardized NB = TPR - \( \left( \frac{C}{B} \right) FPR \frac{(1-\rho)}{\rho} \)

- meaningful as discounted TPR
Discovery Research

- Not producing biomarkers that validate
- Biased designs are common in discovery research
- Yield biomarkers of non-disease related differences between cases and controls
  - anesthesia, medication use, stress, . . . .
  - aging, other medical conditions, . . . .
- Yield biomarkers that look great
  - in severe cases, at diagnosis . . . .
Discovery Research

- Should use PRoBE designs too.

You need to do PRoBE too!

Pepe MS, Li CI, Feng Z. Improving the quality of biomarker discovery research: the right samples and enough of them. Cancer Epidemiol Biomarkers Prev. 2015
Sample Size Calculation for Colocare Study

Colocare

- stage 1 colon cancer
- markers for ‘high’ risk of recurrence within 2 years $\rho = \text{overall recurrence rate} = 10\%$
- ‘high risk’ = 30% = $r$, warrants chemotherapy
- useful marker: $\text{TPR/FPR} \geq \left(\frac{1-\rho}{\rho} \left(\frac{r}{1-r}\right)\right) \approx 3.9$
- fix FPR=10%
- # candidate biomarkers = 9,000
Operating Characteristics

- False leads expected: % FLE = proportion of null markers filtering in = 2% say
- Discovery power: proportion of useful markers filtering in = 95% say
- Filter in criterion: $p$-value for biomarker $< C$

Calculations

- Fix % FLR = 2% by choosing $C = 2$
- works in theory, not always in practice with small samples
- simulations to refine $C$
  - simulations to calculate discovery power
- not computationally intensive: vary # cases and # controls
- 40 cases, 160 controls, $C = 1\%$ yields FLE$\% = 2.3\%$ and Discovery Power = 95\%
Summary

- phases of research
- PRoBE ideal design for validation
- PRoBE ideal design for discovery
- many basic statistical issues
  - measures of performance
  - how to accommodate covariates?
  - is matching a good idea?
  - failure time event data?
  - etc.
- DMCC provides leadership and excellent implementation
Colleagues at EDRN

Ziding Feng       Ross Prentice       Jackie Dahlgren
Mark Thornquist   Ying Huang         Jackie Dahlgren
Yingye Zheng      Holly Janes        Sudhir Srivastava