Developing a Prediction Rule

Henry Glick

Epi 550

January 25, 2012

Pre-test Probability of Disease

• An important anchor for developing management strategies for patients
  – Can be adjusted to account for additional information (either from the physician's experience or from the patient's history)
• Unless you are evaluating a general screening program, the population prevalence is inadequate for establishing the pre-test probability
  – It instead depends on prevalence in patients with particular sets of clinical findings

Clinical Prediction Rules

• Models for assigning patients to subgroups for whom probabilities of disease are known or for suggesting a diagnostic or therapeutic course of action (e.g., who should receive a radiograph and who should not)
  – Laupacis et al.: a decision making tool that includes 3 or more variables obtained from the history, physical examination, or simple diagnostic tests
• Based on clinical studies in which specified data are obtained from patients with and without disease
• Toll et al.*: number of articles discussing prediction rules doubled from 6700 in 1995 to 15,700 in 2005

Steps In Developing Prediction Rules

I. Hypothesis generation
II. Choice of gold standard
III. Choice of predictor variables
IV. Study Sample / Sample size
V. Data collection
VI. Construction of the rule
VII. Test characteristics / Incremental information and cost in different specifications of a rule
VIII. Assessment of the validity of the rule
IX. Provision of information that helps clinicians identify a course of action
X. Assessment of whether the rule affects practice

Illustration

• 2+ prediction rules for strep pharyngitis *
  – Walsh
    • Revised Walsh
  – Centor
    • Modified Centor


Clinical Problem

• 105 outpatient office visits per 1000 children <15 for acute pharyngitis in 2008 (NAMCS)
• Illness generally both benign and self-limited, but antibiotics prescribed in a high percentage of visits
• Caused by a multitude of microbial agents
  – Most cases have a viral etiology
  – Of those with bacterial causes, β-hemolytic group A strep (GABHS) is the commonest: 20-30% of cases among children; 5-15% of cases among adults
• "Given the frequency of strep throat and the voluminous medical literature devoted to this infection…, it is indeed surprising that so much controversy persists regarding the appropriate diagnosis and management of this common and ubiquitous infection." (Bisno)
I. Hypothesis Generation

• Consider a clinically relevant, measurable outcome
• Generate potential predictors of event being predicted
• Potential sources
  – Clinical experience
  – Literature

II. Choice of Gold Standard

• Gold standard should be well specified, objective, and defined by reproducible criteria that are more costly to assess than are the variables in the prediction rule (otherwise, why not use the gold standard)?
  – i.e., What is the outcome (e.g., surrogate or final outcome)?
  – How will it be measured?
  – When will it be measured?
  – If it is a surrogate outcome, does it have a well-established relationship with the clinically important outcome?
• Gold standard should be understood by audience, considered appropriate, and replicable by audience

Tarnished Gold Standard

• Gold standard is tarnished when the outcome is:
  – Indeterminant
  – Incorrect
  – Verified in a nonrandom sample
• Evaluate potential problems associated with tarnishing
• Develop strategy for assigning outcome status
Assessment of Gold Standard

- Blind those deciding on the occurrence of predicted events to the presence of predictors of events
  - What do we know about the accuracy of radiologic readings in the absence of information about the patient?

Gold Standard. Pharyngitis Example

- Walsh: Positive culture for "group A" β-hemolytic streptococci (accuracy = 90%)
- Centor: Positive culture for β streptococcus specifically typed with a rapid latex test
- Modified Centor*: Positive culture for β streptococcus specifically typed with a latex agglutination test

Gold Standard Concerns

- Cooper et al. "Diagnosis of GABHS remains a subject of controversy, partly because the best standard for diagnosis has not been definitively established... Results of throat swab cultures vary according to:"
  - Technique
  - Site in which the sample is obtained and plated
    - Posterior pharynx and tonsils increase sensitivity
  - Culture medium
  - Conditions in which the culture is incubated
  - If results are checked at 24 or 48 hours.

Gold Standard Concerns (2)

- Throat swabs also fail to distinguish acute infection from chronic carrier state
  - The organism can be cultured from the pharynx in the absence of symptoms or signs of infection during winter months:
    - In approximately 10% of school-age children
    - Less frequently in persons in other age groups


Growing Complexities

- Determining an appropriate gold standard growing more complicated
- Black et al. evaluated 4 tests for diagnosis of chlamydia: ligase chain reaction (LCR), polymerase chain reaction (PCR), culture, and DNA probe (DNAP)
  - IF culture=gold standard, sensitivities for LRC, PCR, and DNAP were 96.9, 89.9, and 78.1%; specificities were 97.5, 98.2, and 99.3
  - IF LCR=gold standard, sensitivities for [culture,] PCR and DNAP were [80.1,] 75.8 and 60.8%; specificities were [98.4,] 99.0 and 99.6%


III. Choice of Predictor Variables

- Disease predictors should be well specified, objective, clinically sensible, and reproducible
- Don’t use criteria that are used to define the outcome as predictors of the outcome
  - Suppose some components of the gold standard are inexpensive to collect and utilize?
- Blind those deciding on the presence of predictors to the occurrence of predicted outcomes
- When reporting the rule, indicate variables that were measured but not included in the rule (because they did not add independent predictive information)
- Omission of a potentially important clinical variables does not alter the value of the rule as developed
### Shared Predictor Variables. Pharyngitis Example

<table>
<thead>
<tr>
<th></th>
<th>Walsh et al.</th>
<th>Centor et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cough</td>
<td>Cough</td>
<td>Exudates on the tonsils</td>
</tr>
<tr>
<td>Pharyngeal/tonsillar exudate</td>
<td>Exudates on the pharynx</td>
<td></td>
</tr>
<tr>
<td>Oral temperature</td>
<td>Temperature $\geq 101^\circ$ F</td>
<td></td>
</tr>
<tr>
<td>Pharyngeal erythema</td>
<td>Injection of pharynx</td>
<td></td>
</tr>
<tr>
<td>Swollen tonsils</td>
<td>Tonsil swelling</td>
<td></td>
</tr>
<tr>
<td>Enlarged/tender cervical nodes</td>
<td>Swollen tender anterior or posterior cervical notes</td>
<td></td>
</tr>
<tr>
<td>Recent contact with someone with streptococcal infection</td>
<td>Exposure history</td>
<td></td>
</tr>
<tr>
<td>Rhinorrhea</td>
<td>Coryza</td>
<td></td>
</tr>
</tbody>
</table>

### Distinct Predictor Variables. Pharyngitis Example

<table>
<thead>
<tr>
<th></th>
<th>Walsh et al.</th>
<th>Centor et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss of hearing</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Tinnitus</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>Ear or sinus pain</td>
<td>--</td>
<td></td>
</tr>
<tr>
<td>--</td>
<td>Duration of symptoms</td>
<td></td>
</tr>
<tr>
<td>--</td>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>--</td>
<td>Fever history</td>
<td></td>
</tr>
<tr>
<td>--</td>
<td>Difficulty swallowing</td>
<td></td>
</tr>
</tbody>
</table>

### McIsaac et al. Predictor Variables

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cough</td>
</tr>
<tr>
<td>Tonsillar swelling or exudates</td>
</tr>
<tr>
<td>Temperature $&gt;38^\circ$C (100.4$^\circ$F)</td>
</tr>
<tr>
<td>Swollen and tender anterior cervical nodes</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>3-14</td>
</tr>
<tr>
<td>15-44</td>
</tr>
<tr>
<td>45+</td>
</tr>
</tbody>
</table>
IV. Study Sample

- Best design: Consecutive sample of patients in whom you plan to use rule; i.e.,
  - Subjects should be demographically representative of the patient population in which the rule will be used
  - Subjects with and without disease should be included in the correct proportions
- Potential for bias grows with case/control design or convenience samples, due to potential imbalances in the pre-test probabilities among diseased and nondiseased subjects
  - e.g., bias will occur if all subjects with disease have very high pre-test probabilities and all subjects without disease have very low probabilities

Carcinoembryonic Antigen for Colorectal Cancer

- ROC curve for carcinoembryonic antigen as a diagnostic test for colorectal cancer, by stage of disease. (Redrawn from Fletcher, Fletcher, and Wagner, Clinical Epidemiology, p. 55)

“Levels of Evidence” (Laupacis et al.)

- Prospective data collection specifically to develop or validate the rule
- Data collected as part of another study, not specifically undertaken to develop or validate the rule
- Data collected retrospectively
  - Because of lack of uniform coding of data
  - Because of lack of blinding of potential risk factors and outcome (i.e., those recording signs and symptoms may have done so based on some set of hypotheses they had)
Sample Size

- One approach to sample size for a prediction rule is to base it on the desired error rate (e.g., confidence interval) for sensitivity and/or specificity.
- As we said earlier in the semester, we can estimate the sample size required to estimate a sensitivity or specificity with a desired error rate using a formula for the confidence interval around a single proportion.
  - The proportion of positive tests among those with disease.
  - The proportion of negative tests among those without disease.

Consecutive Patients

- The discussion of study sample indicated that the most robust design used a consecutive sample of patients in whom you plan to use the rule.
- In such a sample, \( p \) patients will have disease for every \( 1-p \) patients who do not (where \( p \) equals the prevalence in the sample).
- Using this design, the total number of patients you will need to sample is the larger of \( N_{dis}/p \) and \( N_{non-dis}/(1-p) \).
- Rule of thumb: require a minimum of 10 patients with the outcome and 10 patients without the outcome for every predictor variable used in the rule.
  - Can only serve to increase the sample size; it can never serve to reduce the sample size!!!

Sample Size. Pharyngitis Example

- Walsh et al.:
  - 418 adult patients presenting with a sore throat at an HMO ambulatory clinic who had a throat culture.
- Centor et al.:
  - 222 out of 286 consecutive adults presenting in the Medical College of Virginia emergency room with complaints of sore throat and were not positive for non-Group A beta streptococcus.
- McIsaac et al.:
  - 787 out of 918 screened persons aged 3 to 69 years of age who participated in a randomized trial comparing 2 different antibacterial therapies for Group A beta streptococcus.
V. Data Collection

- Uniform data collection in all patients in sample
- Either perform the gold standard in everyone or adopt appropriate sampling / analytic techniques if the gold standard is applied in only a subset of subjects

Additional Data

- In addition to the gold standard and predictors, include:
  - Demographic and clinical characteristics
    - Test performance may depend on age, gender, and other patient characteristics that might make the predictive value of the rule different in different populations (e.g., whether it's an asymptomatic population vs. symptomatic population, etc.)
  - Setting in which data were collected
    - Test performance may depend on referral characteristics; type of institution (primary, secondary, or tertiary); whether it was an office, clinic, emergency department, or hospital ward; and whether the site was teaching or nonteaching

VI. Construction of Rule

- "Eyeball" - Useful to get sense of the data
- Univariate (e.g., two by two tables)
- Multivariable
  - Discriminant analysis
  - Branching algorithms / Recursive partitioning
  - Logistic/OLS regression
    - Laupacis et al.: logistic regression/ discriminant analysis maximize accuracy while recursive partitioning results in 1 or more strata that include only patients with a particular outcome
    - Neural networks
Discriminant Analysis. Pharyngitis Example

Walsh et al.
- +3 for each degree of temperature over 36.1°
- +17 for recent exposure to strep infection
- -7 for recent cough
- +6 for pharyngeal exudate
- +11 for enlarged or tender cervical lymph nodes

Translation of Scores to Probabilities

<table>
<thead>
<tr>
<th>Score</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-10 - 0</td>
<td>1.8</td>
</tr>
<tr>
<td>1 - 10</td>
<td>4.6</td>
</tr>
<tr>
<td>11-20</td>
<td>18.0</td>
</tr>
<tr>
<td>21-30</td>
<td>19.0</td>
</tr>
<tr>
<td>31-40</td>
<td>22.0</td>
</tr>
<tr>
<td>41+</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Revised Walsh Risk Scoring System

- McGinn et al. have simplified the Walsh rule:
- Single points are assigned to the five risk factors:

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature &gt;38.3°C</td>
<td>+1</td>
</tr>
<tr>
<td>Exposure to known strep contact</td>
<td>+1</td>
</tr>
<tr>
<td>Pharyngeal or tonsillar exudates</td>
<td>+1</td>
</tr>
<tr>
<td>Enlarged or tender nodes</td>
<td>+1</td>
</tr>
<tr>
<td>Recent cough</td>
<td>-1</td>
</tr>
</tbody>
</table>

- The total score ranges between -1 and +4
Translation of Scores to Probabilities

<table>
<thead>
<tr>
<th>Score</th>
<th>LR</th>
<th>95% CI</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0.16</td>
<td>0.05 - 0.42</td>
<td>4.6</td>
</tr>
<tr>
<td>0</td>
<td>0.62</td>
<td>0.29 - 1.20</td>
<td>15.9</td>
</tr>
<tr>
<td>1</td>
<td>2.61</td>
<td>1.49 - 4.44</td>
<td>44.4</td>
</tr>
<tr>
<td>2</td>
<td>4.35</td>
<td>1.65 - 11.26</td>
<td>57.1</td>
</tr>
<tr>
<td>3+</td>
<td>8.14</td>
<td>1.88 - 35.23</td>
<td>71.4</td>
</tr>
</tbody>
</table>

Branching Algorithms / Recursive Partitioning

- Builds an empirical tree diagram by:
  - Identifying the best predictor of disease and dividing the entire study population into two groups: one with the predictor (and a relatively high risk of disease) and one without it (and with a relatively low risk of disease)
  - Sequentially dividing each group into subgroups with each of the remaining predictors
- Each path along the tree represents a sequence of clinical findings and defines a patient subgroup (and associated probability of disease)
- Software is available

Recursive Partitioning Algorithm Adults with Sore Throat

If the score for recent exposure to strep infection is +17 and that for enlarged or tender cervical lymph nodes is +11, why isn’t recent exposure the first branching point?
Logistic Regression. Pharyngitis Example

Centor et al.

• Four clinical features
  – Tonsillar exudates
  – Swollen and tender anterior cervical lymph nodes
  – Lack of cough
  – History of fever

Calculating the Probability of Disease

• Proceeds in 2 Steps
  Step 1. Use the Estimated Coefficients and
  Explanatory Variables to Calculate a Risk Score "S"
  Where S = α + ∑ βi Xi
  α = the Intercept
  βi = the coefficients from the logistic Regression
  Xi = the Predictor Variables

Strep Pharyngitis Coefficients

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-2.69</td>
</tr>
<tr>
<td>Tonsillar exudates</td>
<td>1.04</td>
</tr>
<tr>
<td>Swollen/tender anterior cervical nodes</td>
<td>1.00</td>
</tr>
<tr>
<td>Cough</td>
<td>-0.95</td>
</tr>
<tr>
<td>Fever history</td>
<td>0.89</td>
</tr>
</tbody>
</table>

• For a person with tonsillar exudates and fever history:
  S = -2.69 + 1.04 + 0.89 = -0.76
Risk Score S

- S ranges between $-\infty$ and $+\infty$
- When S approaches $-\infty$, the predicted probability approaches 0; when S approaches $+\infty$, the predicted probability equals 1
- When S = 0, the predicted probability = 0.5

Calculating the Probability of Disease (II)

- Step 2. Transform S into a probability
  \[ p = \frac{e^S}{1 + e^S} \]
- For a person with tonsillar exudates and a fever history
  \[ p = \frac{e^{-0.76}}{1 + e^{-0.76}} = \frac{0.46767}{1 + 0.46767} = 0.3184 \]

Other Risk Scores / Probabilities

<table>
<thead>
<tr>
<th>Probability (%)</th>
<th>Risk Score S</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-2.1972246</td>
</tr>
<tr>
<td>20</td>
<td>-1.3862944</td>
</tr>
<tr>
<td>30</td>
<td>-0.84729786</td>
</tr>
<tr>
<td>40</td>
<td>-0.40546511</td>
</tr>
<tr>
<td>50</td>
<td>0</td>
</tr>
</tbody>
</table>

- Risk scores for probabilities greater than 0.5 (1-p) are the absolute value of the risk scores for probabilities (p) less than 0.5 (e.g., the risk score representing a probability of 90% is 2.1972246)
Create a Risk Scoring System

• Creating a risk scoring system based on values of the independent variables and coefficients
  – Centor coefficients
    1.04 Tonsillar exudates
    1.00 Swollen/tender anterior cervical nodes
    0.95 Absence of Cough
    0.89 Fever history
  – Appropriate to assume equal weighting

Equally Weighted Risk Scoring System

• Centor et al.

<table>
<thead>
<tr>
<th>Number of Features Present</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.5</td>
</tr>
<tr>
<td>1</td>
<td>6 - 6.9</td>
</tr>
<tr>
<td>2</td>
<td>14.1 - 16.6</td>
</tr>
<tr>
<td>3</td>
<td>30.1 - 34.1</td>
</tr>
<tr>
<td>4</td>
<td>55.7</td>
</tr>
</tbody>
</table>

Moving a Rule to a Practice with a Very Different Prevalence of Disease?

• Suppose you developed a prediction rule in a population with a probability of disease of 10% and wanted to use it in a population with a probability of disease of 5%. Would it be accurate in the latter population?
  – Yes, if the primary reason for the difference was a difference in the distribution of the risk factors (e.g., fewer patients develop tonsillar exudates, fewer have a fever history, etc.)
Moving Rule When Prevalence Differs

- No, if the individuals without any risk factors in the new population (e.g., the ones with a score of 0) have a risk for disease that is less than 2.5%
  - If the odds ratios are unaffected between the two populations, we can adjust for this difference by changing the risk for disease in those without risk factors (i.e., a change in intercept from the logistic regression)
- No, if the odds ratios for the risk factors differ (i.e., changes in the coefficients from the logistic regression)

Intercept Shift Revision of Rule

- Centor et al. subtracted approximately 1.3 from the intercept to modify the rule for the new setting

<table>
<thead>
<tr>
<th>Score</th>
<th>-S</th>
<th>Probability (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>-4.9636</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>-3.9774</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>-3.0074</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>-2.0492</td>
<td>11</td>
</tr>
<tr>
<td>3+</td>
<td>-1.071</td>
<td>25</td>
</tr>
</tbody>
</table>

McIsaac Score

<table>
<thead>
<tr>
<th>Feature</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature &gt;38°C</td>
<td>1</td>
</tr>
<tr>
<td>Absence of cough</td>
<td>1</td>
</tr>
<tr>
<td>Swollen and tender anterior cervical nodes</td>
<td>1</td>
</tr>
<tr>
<td>Tonsillar swelling or exudates</td>
<td>1</td>
</tr>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>3-14</td>
<td>1</td>
</tr>
<tr>
<td>15-44</td>
<td>0</td>
</tr>
<tr>
<td>45+</td>
<td>-1</td>
</tr>
</tbody>
</table>
McIsaac Scoring System

<table>
<thead>
<tr>
<th>Score</th>
<th>Probability of Disease</th>
<th>Suggested Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>≤0</td>
<td>1 - 2.5%</td>
<td>No further testing or Rx</td>
</tr>
<tr>
<td>1</td>
<td>5 - 10%</td>
<td>Culture all; Rx for positives</td>
</tr>
<tr>
<td>2</td>
<td>11 - 17%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>28 - 35%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>51 - 53%</td>
<td>Rx all and/or culture</td>
</tr>
</tbody>
</table>

General Principles for Generating Risk Scoring Systems †

- Because calculating $\Sigma \beta_i x_i$ is tedious and likely to be a disincentive to use of prediction rules, we routinely develops a point system for the rule
  - Achieved by assigning integer points to each level of each risk factor to approximate $\Sigma \beta_i x_i$
  - Resulting risk estimate is derived from a reference table which reports risks for different point totals
- Point systems usually break continuous variables into categories
  - May want categories to mirror clinically meaningful risk factor categories
  - e.g., JNC VI blood pressure categories


Steps in Generating Risk Scoring Systems

1. Categorize the risk factors and calculate $\beta_i x_i$ for the reference values
   - Calculate scores
2. Translate scores into points
3. Determine risks associated with point totals
   (See detailed example in Appendix 1)
Reporting on Construction of the Rule (Laupacis et al.)

- Adequately describe and justify the mathematical technique used to derive the rule
- Address whether or not you avoided the problem of overfitting the data with too few events per predictor variable

Reporting on Construction of the Rule (2)

- Specify how the variables were selected (e.g., did you use a preliminary screen based on univariate association and reliability?)
  - Prespecify the predictors that will be used in the model
  - Develop prespecified criteria for selecting predictors (i.e., all variables with correlations of 0.15 or greater are candidates; backward stepwise procedure; reassess correlation of regression residuals and non-candidates)
- Specify regression diagnostics utilized (influential observations and multicollinearity)

Reproducibility

- Reproducibility (interobserver agreement) applicable both for assessment of predictor variables and of rule
- Measured either with the kappa statistic or correlation coefficient
- Values less than 0.6 generally represent lack of agreement
- Predictors with low reproducibility should not be included in the rule
- Given costs involved with assessment, can be assessed for a representative subset
VII. Test Characteristics

- Discrimination
- Calibration
- Deal with patients with indeterminant disease status

Discrimination

- Discrimination represents the ability to give different scores to those with and without disease (e.g., to assign generally lower scores to those without disease and to assign generally higher scores to those with the disease)
- Measures of discrimination
  - Sensitivity and Specificity
  - ROC Analysis

McGinn et al ROC Curve

Figure 1. Receiver operating characteristic curves for Walsh and simplified clinical prediction rules (CPRs).
**McIsaac* Sensitivity and Specificity**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>100</td>
<td>93.2 (90.8 - 95.1)</td>
</tr>
<tr>
<td>&lt;18</td>
<td>100</td>
<td>90.3 (86.4 - 93.4)</td>
</tr>
<tr>
<td>18+</td>
<td>100</td>
<td>96.5 (93.5 - 98.4)</td>
</tr>
</tbody>
</table>

* Test characteristics for the proposed testing and treatment algorithm

---

**Discrimination Not the Only Criterion for a Good Prediction**

- Example of a perfectly discriminating, but in some sense mistaken, prediction
  - If a weatherperson always says there is a 51% chance of rain on days when it rains and always says there is a 49% chance of rain on days when it does not rain, he/she is perfectly discriminating (sensitivity = 1.0; specificity = 1.0)

- Example of a totally nondiscriminating, but in some sense accurate, prediction
  - If a weatherperson always says there is a 30% chance of rain, and in truth it rains 3 out of every 10 days (i.e., he/she gives the same score to every day, whether it rains or not)

---

**Calibration**

- Calibration is a measure of the accuracy of the predicted probabilities of disease (e.g., the degree to which observed and predicted probabilities are equal)
  - Because it is a property of predicted probabilities and not of scores like serum creatinine or hemoglobin levels, it does not play a role in the evaluation of diagnostic test characteristics
  - It could, however, play a role in the evaluation of a physician’s pre-test probabilities or of post-test probabilities
Types of Calibration

- There are at least two types of calibration, calibration in the large and calibration in the small.
- Calibration in the large, is a property of the full sample.
  - It is calculated by comparing the observed probability in the full sample (e.g., 10% if there are 1000 patients and 100 have the outcome being predicted) with the average predicted probability in the full sample (i.e., the average of each of the predictions).
  - e.g., if the latter is 10%, the prediction rule is perfectly calibrated in the large.
  - For the sample in which a logistic regression is estimated, the results are perfectly calibrated in the large (i.e., the average of the predicted probabilities equals the average probability in the sample).

Calibration in the Large: Necessary But Not Sufficient

<table>
<thead>
<tr>
<th>Obs #</th>
<th>Truth</th>
<th>Pred Rule 1</th>
<th>Pred Rule 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Avg Prob</td>
<td>30%</td>
<td>30%</td>
<td>30%</td>
</tr>
</tbody>
</table>

- Although the 2 rules have identical calibration in the large, rule 2 is better than rule 1.

Calibration in the Small

- Calibration in the small is a property of subsets of the sample.
  - It is calculated by comparing the observed probability in each subset with the average predicted probability in the subset.
  - A weatherperson who makes 3 kinds of predictions (e.g., 5% chance of rain today, 50% chance of rain today, and 95% chance of rain today) is well calibrated in the small if:
    - On days with a 5% predicted probability, 5% of the time it rains;
    - On days with a 50% probability, 50% of the times it rains;
    - On days with a 95% probability, 95% of the times it rains.
Why is calibration in the small a property of subsets of the sample rather than of individual observations in the sample?

**Calibration Curve: Calibration in the Small**
- Horizontal axis
- Vertical axis
- 45° line

**Steps in Plotting Calibration in the Small**
1. Obtain the required 2 data items for each individual
   - Predicted probability of the outcome
   - The gold standard determination
2. Using predicted probability, rank order observations from lowest to highest
3. Divide the rank-ordered observations into groups (e.g., if there are 1000 observations, 20 groups of 50 observations)
4. Calculate the observed probability / group (number of outcomes coded a 1 divided by total observations / group
5. Calculate the mean predicted probability in each group
6. Plot observed and mean predicted probabilities for each group (e.g., 20 points on the calibration plot)
Calibration in the Small

<table>
<thead>
<tr>
<th>Obs #</th>
<th>Pred Prob (%)</th>
<th>Truth</th>
<th>Pred / Obs (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>1</td>
<td>30.3 / 33.3</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>46</td>
<td>1</td>
<td>50.5 / 50</td>
</tr>
<tr>
<td>6</td>
<td>55</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>56</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>61</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>66</td>
<td>1</td>
<td>66.3 / 66.7</td>
</tr>
<tr>
<td>10</td>
<td>72</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Avg Prob</td>
<td>49.2%</td>
<td>50%</td>
<td></td>
</tr>
</tbody>
</table>

Mortality Prediction, Suspected Alzheimer’s Disease

- Based on data from 2023 and 590 elderly persons for whom data on mortality after 5 and 10 years of follow-up, respectively, were available.

Calibration Statistics

- Logistic regression - Hosmer and Lemeshow or Pearson
- Yates Decomposition
### Example Data

| Obs# | Cure | Inftype | Severe | Cured | | Obs# | Cure | Inftype | Severe |
|------|------|---------|--------|-------|------|------|---------|--------|
| 1    | 0    | 0       | 0      | 11    | 1    | 0    | 0       |
| 2    | 0    | 0       | 1      | 12    | 1    | 0    | 1       |
| 3    | 0    | 0       | 2      | 13    | 1    | 0    | 2       |
| 4    | 0    | 0       | 3      | 14    | 1    | 1    | 0       |
| 5    | 0    | 0       | 3      | 15    | 1    | 1    | 0       |
| 6    | 0    | 0       | 4      | 16    | 1    | 1    | 1       |
| 7    | 0    | 0       | 5      | 17    | 1    | 1    | 1       |
| 8    | 0    | 1       | 3      | 18    | 1    | 1    | 2       |
| 9    | 0    | 1       | 4      | 19    | 1    | 1    | 3       |
| 10   | 0    | 1       | 5      | 20    | 1    | 1    | 4       |

### Sample Statistics

<table>
<thead>
<tr>
<th></th>
<th>Cured</th>
<th>Not Cured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infection type=1</td>
<td>70%</td>
<td>30%</td>
</tr>
<tr>
<td>Severity</td>
<td>1.4 (1.35)</td>
<td>3 (1.63)</td>
</tr>
</tbody>
</table>

### logistic cure inftype severity

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Logit estimates</td>
<td>Number of obs</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR chi² (2)</td>
<td>=</td>
<td>10.92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prob &gt; chi²</td>
<td>=</td>
<td>0.0042</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log likelihood = -8.40</td>
<td>Pseudo R²</td>
<td>= 0.3939</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OR Std Err Z</td>
<td>P&gt;</td>
<td>z</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cure</td>
<td>Inftype</td>
<td>Severity</td>
<td>22.07</td>
<td>35.94</td>
<td>1.90</td>
</tr>
<tr>
<td></td>
<td>0.3240</td>
<td>0.1757</td>
<td>-2.08</td>
<td>0.038</td>
<td></td>
</tr>
</tbody>
</table>
Logistic model for cure

number of observations = 20
area under ROC curve = 0.8750

ROC Curve for Cure Example

logit cure inftype severity

Hosmer and Lemeshow Statistic

estat gof, group(4) table
Logistic model for cure, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

<table>
<thead>
<tr>
<th>Group</th>
<th>Prob</th>
<th>Obs_1</th>
<th>Exp_1</th>
<th>Obs_0</th>
<th>Exp_0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1936</td>
<td>0</td>
<td>0.4</td>
<td>5</td>
<td>4.6</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>0.4595</td>
<td>3</td>
<td>2.1</td>
<td>3</td>
<td>3.9</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>0.7914</td>
<td>2</td>
<td>2.8</td>
<td>2</td>
<td>1.2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0.9830</td>
<td>5</td>
<td>4.7</td>
<td>0</td>
<td>0.3</td>
<td>5</td>
</tr>
</tbody>
</table>

Number of observations = 20
Number of groups = 4
Hosmer-Lemeshow ch_2(2) = 1.95
Prob > chi2 = 0.3768
Pearson Chi$^2$ Statistic

estat gof
Logistic model for cure, goodness-of-fit test

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>20</td>
</tr>
<tr>
<td>Number of cpvariate patterns</td>
<td>12</td>
</tr>
<tr>
<td>Pearson chi2(9)</td>
<td>2.54</td>
</tr>
<tr>
<td>Prob &gt; chi2</td>
<td>0.9798</td>
</tr>
</tbody>
</table>

Calibration and Discrimination, Examples

- Are the following weather people well discriminating and well calibrated?
  - Example 1: Every day, the weatherperson makes 1 of only 2 predictions, either a 49% chance of rain or a 51% chance of rain. On all days when she says there is a 49% chance of rain, it fails to rain; on all days when she says there is a 51% chance of rain, it rains
  - Example 2: Every day, the weatherperson makes 1 of only 2 predictions, either a 5% chance of rain or a 95% chance of rain. On days when she says there is a 5% chance of rain, it rains 5 of every 100; on days when she says there is a 95% chance of rain, it rains 95 of every 100

Calibration and Discrimination, Examples (2)

- Are the following weather people well discriminating and well calibrated?
  - Example 3: Every day, the weatherperson predicts there is a 50% chance of rain (and in truth it rains 5 out of every 10 days)
  - Example 4: Every day, the weatherperson makes 1 of only 2 predictions, either a 5% chance of rain or a 95% chance of rain. On days when she says there is a 5% chance of rain, it rains 2 of every 10; on days when she says there is a 95% chance of rain, it also rains 2 of every 10 (would your modify your conclusion if on 16.7% of all days the weatherperson said 95%?)
Calibration and Discrimination

- What does the calibration curve look like for a highly discriminating and well calibrated prediction rule?
  - As the following examples demonstrate, It will have observations that all fall on the 45° line and that are clustered near 0,0 and 1,1
  - Consider the following 3 prediction rules:

Example: Highly Discriminating and Well Calibrated

<table>
<thead>
<tr>
<th>Strata</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rule 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D+</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td>55</td>
<td>60</td>
<td>250</td>
</tr>
<tr>
<td>D-</td>
<td>60</td>
<td>55</td>
<td>50</td>
<td>45</td>
<td>40</td>
<td>250</td>
</tr>
<tr>
<td>Pred P</td>
<td>40%</td>
<td>45%</td>
<td>50%</td>
<td>55%</td>
<td>60%</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Rule 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D+</td>
<td>10</td>
<td>30</td>
<td>50</td>
<td>70</td>
<td>90</td>
<td>250</td>
</tr>
<tr>
<td>D-</td>
<td>90</td>
<td>70</td>
<td>50</td>
<td>20</td>
<td>10</td>
<td>250</td>
</tr>
<tr>
<td>Pred P</td>
<td>10%</td>
<td>30%</td>
<td>50%</td>
<td>70%</td>
<td>90%</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Rule 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D+</td>
<td>5</td>
<td>10</td>
<td>50</td>
<td>90</td>
<td>95</td>
<td>250</td>
</tr>
<tr>
<td>D-</td>
<td>95</td>
<td>90</td>
<td>50</td>
<td>10</td>
<td>5</td>
<td>250</td>
</tr>
<tr>
<td>Pred P</td>
<td>5%</td>
<td>10%</td>
<td>50%</td>
<td>90%</td>
<td>95%</td>
<td>50%</td>
</tr>
</tbody>
</table>

- All 3 rules are perfectly calibrated in the small
- Thus, all 3 rules are perfectly calibrated in the large

ROC and Calibration Curves, 3 Rules
Conclusions: Highly Discriminating and Well Calibrated

- All 3 rules are well calibrated in the large and in the small
- When the points on a calibration curve are clustered together, discrimination cannot be very good
- When the points are pushed towards both ends of the calibration curve (large fractions of the predictions between 0 and 20% and large fractions between 80 and 100%), discrimination will be reasonably good

Calibration and Discrimination (2)

- Does the location of the cluster of points on the calibration curve affect discrimination
  - e.g., If they are clustered between 5% and 25%; between 40% and 60%; and between 75% and 95%
- Consider the following 3 prediction rules:

Example: Location of the Cluster

<table>
<thead>
<tr>
<th>Strata</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>D+</td>
<td>40</td>
<td>45</td>
<td>50</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>D-</td>
<td>60</td>
<td>55</td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>Pred P</td>
<td>40%</td>
<td>45%</td>
<td>50%</td>
<td>55%</td>
</tr>
<tr>
<td>Rule 2</td>
<td>D+</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>D-</td>
<td>95</td>
<td>90</td>
<td>85</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Pred P</td>
<td>5%</td>
<td>10%</td>
<td>15%</td>
<td>20%</td>
</tr>
<tr>
<td>Rule 3</td>
<td>D+</td>
<td>75</td>
<td>80</td>
<td>85</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>D-</td>
<td>25</td>
<td>20</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Pred P</td>
<td>75%</td>
<td>80%</td>
<td>85%</td>
<td>90%</td>
</tr>
</tbody>
</table>
ROC and Calibration Curves, 3 Rules

Conclusions: Location of the Cluster

• Even though the rules appear similar, in that there are absolute 5% differences between each stratum:
  – The ROC area reaches a minimum when they are centered at 50% (because for both D+ and D- there is only a 150% difference between the Ns in the two extreme strata)
  – The ROC areas are equal as the clusters symmetrically approach the two ends of the probability distribution, and are increasing (because for either D+ or D-, there are larger percentage differences between the Ns in the two extreme strata [e.g., 500%])
  – When a rule is well calibrated, and when the points on the calibration curve are clustered within a small region, the rule’s discriminating ability will be small.

Calibration In Diseased and Nondiseased Individuals

• When we construct the calibration plot, we rank order observations by the predicted probability. Why don’t we rank order them by observed outcome
  – Suppose we looked at the predicted probabilities of disease among people in whom we know disease did not occur for two prediction rules
  – Would we know if the rule that had the lower predicted probability among those without disease was better calibrated or discriminating (see rules 1b and 1c above, or alternatively, rules 1a and 1c)?
Diseased and Nondiseased Individuals

- Calibration is a property of predicted probabilities, not of known disease status
  - If you correctly state the probability of disease is 5%, 95 of 100 will not have disease; if you correctly state it is 95%, 5 of 100 will not have disease
  - Implication: For very good prediction rules, if you looked only at subjects without disease, you would not expect the predicted probability of disease to be 0%

Incremental Information and Costs in Different Specifications of a Rule

- Clinical information
  - Differences in intercepts
  - Differences in area
- Costs

VIII. Assessment of Validity *

- Predictive validity refers to the quality of the rule's predictions in the sample in which it was developed and in new samples
- Most prediction rules lose accuracy when used in patients who were not included in the derivation sample
  - e.g., ROC area for prediction rule diagnosing serious bacterial infection in children presenting with fever without apparent source equaled 0.76 (95% CI 0.66 to 0.88) in the derivation data set, but equaled 0.57 (95% CI, 0.47 to 0.67) when applied to new patients from another hospital in a later period

Sources of Reduced Accuracy

- Differences between the derivation and validation population (case-mix)
- Differences in the definitions of predictors and the outcome variable and the measurement methods between the derivation and validation populations
- Improvements over time of measurement techniques for predictors may improve, which may affect the predictive strength of a predictor

Apparent Differences

- Apparent differences may be due to the fact that validation studies commonly include fewer individuals than development studies
  - Apparent differences may be due to random variation
- For prediction rules that predict dichotomous outcomes, it has been suggested that the validation sample should contain at least 100 events and 100 nonevents to detect substantial changes in accuracy (for example, a 0.1 change in the ROC area) with 80% power

Internal Validity

- Quality of prediction in the derivation dataset
- Calibration and discrimination are 2 measures of internal validity
- Bootstrapping, split-samples, and training/test datasets are internal validation techniques (because they are performed on the derivation dataset) used to address the external validation concerns of overfitting or "optimism"
Internal Validity (2)

- Overfitting is the modeling of relationships that are specific to the derivation dataset, and would not hold in other datasets
- One approach for addressing overfitting is to:
  - Draw repeated bootstrap samples
  - Perform variable selection in each
  - Use the resulting model in the bootstrap dataset as well as the full derivation dataset and calculate each area under the ROC curve
  - Interpret the difference in the mean areas between the bootstrap and derivation datasets as a measure of "optimism"

External Validity

- Quality of prediction in a new validation dataset
- Report information about the study population so its generalizability can be assessed. Data include:
  - The medical setting from which the patients were drawn,
  - The age, gender, and clinical characteristics of the patients

"Levels" of External Validation

- Temporal validation
  - Tests the generalizability of a prediction rule "over time, typically using the same physicians or investigators as in the development study, in the same institution(s), and in similar patients
- Geographical validation
  - Tests the generalizability of a prediction rule in a patient population that is similarly defined as the development population, but in hospitals or institutions of other geographical areas
"Levels" of External Validation (2)

- Domain validation
  - Evaluates the generalizability of a prediction rule across patients from different settings (primary, secondary, or tertiary care / inpatients versus outpatients), patients of different ages or genders, and perhaps from a different type of hospital (academic vs. general hospital)
- The level of evidence of validation increases as we go down the list

Assessment of Validity Pharyngitis Example

- Walsh et al.
  - "[The rules were] developed on the basis of the data collected in the first five months of the study (246 patients) and then shown to perform as effectively on the next 172 patients."

Updating Prediction Rules

- When a validation study shows disappointing results, we may want to consider updating the rule by combining the information from the original rule with the information from the validation population
- Six general strategies
  1. If the prevalence differs dramatically between study populations, adjust the intercept of the original prediction rule (e.g., updating the Centor strep rule)
  2. "Logistic recalibration": Adjust Intercept / coefficients with a single correction factor estimated from the data of the new patients in the validation set
     - These two methods may improve calibration, but cannot improve discrimination, because the recalibration does not affect the rankings
Updating Prediction Rules (2)

- Model revisions that modify discrimination and calibration:
  3. Restimate regression coefficients that for those variables that differ by use of the validation data
  4. Use the validation data to estimate coefficients for predictors that were omitted from the original rule
  5. Reestimate intercept and all predictors by use of the validation data
  6. Reestimate intercept and predictors and estimate coefficients for predictors that were omitted from the original rule by use of the validation data

IX. Provision of Information That Helps Clinicians Identify a Course of Action after Applying the Prediction Rule

- Laupacis et al.: “Rules are more likely to be used if they suggest a course of action rather than provide a probability of disease. This is likely to be particularly true in situations where a decision must be made quickly.”

Courses of Action: Pharyngitis Example

  - Withhold treatment and do not obtain cultures when $P < 5\%$
  - Obtain cultures when $P \geq 5\%$ and $\leq 20\%$; treat if positive culture
  - Treat without culture when $P > 20\%$

- McIsaac’s Decision Rule (JAMA, 2004)
  - For scores $\leq 1$, $p \leq 10\%$: Withhold treatment and do not obtain cultures
  - For scores of 2 or 3, $11\% \leq p \leq 35\%$: Obtain cultures and treat if positive culture
  - For scores $\geq 4$, $p > 50\%$: Treat without culture
Course of Action (II)

- ACP, AAFP, and CDC consider it reasonable not to perform a throat culture or rapid antigen-detection test if none of the 4 "Centor" clinical features are present.
- Endorse three strategies for adults with two or more features:
  - Treat patients with a positive rapid test
  - Treat without testing if all 4 clinical features are present or after a rapid test if 2 or 3 features are present
  - Treat without testing if three or four features are present
- The IDSA and AHA do not endorse the 2nd and 3rd strategies because these approaches result in higher rates of prescribing unnecessary antibiotics.

Course of Action (III)

- McGinn (revised Walsh algorithm) recommends:
  - Empiric therapy for all patients with a score of 2+ (>55% post-test probability) and rapid testing for patients with scores of 0 or 1 (post-test probability > 15%)
- Bisno (2002) raises concerns about the ACP recommendation for empiric therapy for Centor scores of 3
  - Leads to misuse of antibiotics in 60-70% of patients with this finding

"Yet, the major strategic consideration, surely, should be limitation of excessive and unnecessary prescription of antimicrobials...."

X. Assessment of Whether the Rule Affects Practice

- Providers may not use a rule's predictions because:
  - They believe, or it has been demonstrated, that their predicted probability is at least as good as the probability calculated with a prediction rule
    - e.g., Sinuff et al. found that ICU physicians more accurately discriminated between survivors and nonsurvivors in the first 24 hours of ICU admission than did the ICU survival prediction rules
  - They believe their patients are different from those used in the development of the rule
  - They are afraid they won't apply the rule correctly
  - They feel the false negative rate is too high
Sensibility

- Physicians also may not use the rule if they don’t find it to be sensible (to have face validity) even if it can be shown to be effective
  - Items included in rule should clinical sense and seem appropriate for the purpose of the rule
  - No obvious items should be missing (or their absence is adequately explained)
  - The method for aggregating component variables should appear reasonable

Assessment of Whether the Rule Affects Practice (II)

- Providers may not use a rule because:
  - The rule is not user-friendly or significantly extends the time of the usual clinical encounter, e.g., the rule:
    - Includes variables that are not collected in daily practice
    - Requires extensive calculations or use of a calculator
  - They believe there are practical barriers to its use, such as fear of malpractice litigation

Assessment of Whether the Rule Affects Practice (III)

- Adoption may depend on age and training
  - Brehaut et al. found that older physicians and part-time working physicians were less likely to be familiar with the Ottawa ankle rule
  - The best predictors whether a rule would be used in practice were 1) familiarity acquired during training, 2) confidence in the usefulness of the rule, and 3) user-friendliness of the rule
Impact Analysis *

- Ascertainment of whether a rule is used by clinicians, changes or directs physicians' decisions and improves clinically relevant process parameters, patient outcomes, or cost-effectiveness
- Prepare for impact analysis
  - Translate predictions into decisions
  - Get clinicians' input
  - Anticipate potential obstacles
  - Define impact


Perform Impact Analysis

- Use appropriate study design
  - The ideal design use a cluster randomized trial in which physicians or care units are randomized to either use of the rule or use of "care or clinical judgment as usual"
  - Alternate design: before/after study within the same physicians or care units (temporal changes may compromise the validity of this design)
  - Randomization of patients rather than physicians or care units is not advised
    - Learning effects and contamination may lead to a reduced contrast between the randomization groups

Perform Impact Analysis (II)

- Consider inclusion criteria
- The ideal endpoints are clinically relevant process parameters, patient outcomes, and cost-effectiveness
- Use blinding
- Estimate sample size
- Understand the potential versus actual impact: efficacy versus efficiency
### Standards of Evidence for Prediction Rules *

<table>
<thead>
<tr>
<th>Level of Evidence</th>
<th>Standard of Evaluation</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1: Derivation</td>
<td>Identification of predictors using multivariate model; blinded assessment of outcomes</td>
<td>Needs validation and further evaluation before using clinically in actual patient care</td>
</tr>
<tr>
<td>Level 2: Narrow validation</td>
<td>Verification of predictors when tested prospectively in 1 setting; blinded assessment of outcomes</td>
<td>Needs validation in varied settings; may use predictions cautiously in patients similar to sample studied</td>
</tr>
<tr>
<td>Level 3: Broad validation</td>
<td>Verification of predictive model in varied settings with wide spectrum of patients and physicians</td>
<td>Needs impact analysis; may use predictions with confidence in their accuracy</td>
</tr>
</tbody>
</table>

EBM Working Group cited in Reilly & Evans

---

### Standards of Evidence for Prediction Rules *

<table>
<thead>
<tr>
<th>Level of Evidence</th>
<th>Standard of Evaluation</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 4 Narrow impact analysis</td>
<td>Prospective demonstration in 1 setting that use of prediction rule improves physicians' decisions (quality or C-E of patient care)</td>
<td>May use cautiously to inform decisions in settings similar to that studied</td>
</tr>
<tr>
<td>Level 5 Broad impact analysis</td>
<td>Prospective demonstration in varied settings that use of prediction rule improves physicians' decisions for wide spectrum of patients</td>
<td>May use in varied settings with confidence that its use will benefit patient care quality or effectiveness</td>
</tr>
</tbody>
</table>

EBM Working Group cited in Reilly & Evans