Developing a Prediction Rule

Henry Glick

Epi 550

January 30, 2015

Pre-test Probability of Disease

• An important anchor for developing management strategies for patients
  – Can be adjusted to account for additional information (either from physician's experience or from patient's history)
• Unless evaluating a general screening program, population prevalence is inadequate for establishing pre-test probability
  – Depends instead 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 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

Prediction Rules vs Diagnostic Tests *

<table>
<thead>
<tr>
<th>Diagnostic Prediction Modeling</th>
<th>Prognostic Prediction Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanatory variables, predictors, covariates (X variables)</td>
<td>Prognostic factors or indicators</td>
</tr>
<tr>
<td>Diagnostic tests or index tests</td>
<td>Outcomes (Y variables)</td>
</tr>
<tr>
<td>Target disease/disorder (presence vs absence)</td>
<td>Event (future occurrence: yes no) Event definition measurement</td>
</tr>
<tr>
<td>Reference Standard</td>
<td>Missing Outcomes</td>
</tr>
<tr>
<td>Partial verification</td>
<td>Loss to follow-up/censoring</td>
</tr>
</tbody>
</table>


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 rule
VII. Test characteristics / Incremental information and cost in different specifications of a rule
VIII. Assessment of validity of rule
IX. Provision of information that helps clinicians identify a course of action
X. Assessment of whether 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 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 variables in prediction rule (otherwise, why not use gold standard)
  - i.e., What is 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 clinically important outcome?
- Gold standard should be understood by audience, considered appropriate, and replicable by audience
Tarnished Gold Standard

- Gold standard is tarnished when outcome is:
  - Indeterminate
  - 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 occurrence of predicted events to presence of predictors of events
  - What do we know about accuracy of radiologic readings in absence of information about 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 sample is obtained and plated
    • Posterior pharynx and tonsils increase sensitivity
  – Culture medium
  – Conditions in which 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
  – Organism can be cultured from pharynx in 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

• In Black et al. evaluation of 4 tests for diagnosis of chlamydia: ligase chain reaction (LCR), polymerase chain reaction (PCR), culture, and DNA probe (DNAP)
  – When culture=gold standard:
    • Sensitivities for LCR, PCR, and DNAP were 96.9, 89.9, and 78.1%; specificities were 97.5, 98.2, and 99.3
  – When 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 outcome as predictors of outcome
  - Suppose some components of gold standard are inexpensive to collect and utilize?
- Blind those deciding on presence of predictors to occurrence of predicted outcomes
- When reporting rule, indicate variables that were measured but not included in rule (because they did not add independent predictive information)
- Omission of a potentially important clinical variables does not alter value of rule as developed

Shared Predictor Variables. Pharyngitis Example

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

Distinct Predictor Variables. Pharyngitis Example

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

- Cough
- Tonsillar swelling or exudates
- Temperature >38°C (100.4°F)
- Swollen and tender anterior cervical nodes
- Age
  - 3-14
  - 15-44
  - 45+

IV. Study Sample

- “Was the spectrum of patients representative of the patients who will receive the test in practice?” (Whiting et al. The Development of QUADAS… BMC Medical Research Methodology. 2003;3:25)
- Best design: Consecutive sample of patients in whom you plan to use rule; i.e.,
  - Subjects should be demographically representative of patient population in which rule will be used
  - Subjects with and without disease should be included in “correct” proportions
- May want to ensure adequate samples of subgroups of interest (to see if rule has same operating characteristics in the subgroups)

IV. Study Sample (2)

- Potential for bias grows with case/control design or convenience samples, due to potential imbalances in 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
Spectrum Bias Setup: Rapid Antigen Test *

<table>
<thead>
<tr>
<th></th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>RADT+</td>
<td>368</td>
<td>87</td>
</tr>
<tr>
<td>RADT-</td>
<td>97</td>
<td>1124</td>
</tr>
<tr>
<td>Sens</td>
<td>79.1</td>
<td>91.3</td>
</tr>
</tbody>
</table>


Spectrum Bias, Rapid Antigen Test

<table>
<thead>
<tr>
<th></th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cent 0/1</td>
<td>Cent 3/4</td>
</tr>
<tr>
<td>RADT+</td>
<td>181</td>
<td>187</td>
</tr>
<tr>
<td>RADT-</td>
<td>66</td>
<td>31</td>
</tr>
<tr>
<td>Sens</td>
<td>73.3</td>
<td>85.8</td>
</tr>
<tr>
<td>p</td>
<td>0.001</td>
<td>0.000</td>
</tr>
</tbody>
</table>

• Differences in PPV and NPV neither necessary nor sufficient for spectrum bias

Spectrum Bias, RADT (2)

• Suppose we had used a common design:
  – Enroll diseased patients with lots of signs and symptoms (e.g., Centor 3/4) to construct D+ sample
  – Enroll nondiseased patients with few if any signs and symptoms (e.g., Centor 1/2) to construct D- sample?

<table>
<thead>
<tr>
<th></th>
<th>D+</th>
<th>D-</th>
</tr>
</thead>
<tbody>
<tr>
<td>RADT+</td>
<td>187</td>
<td>44</td>
</tr>
<tr>
<td>RADT-</td>
<td>31</td>
<td>813</td>
</tr>
<tr>
<td>Tot</td>
<td>218</td>
<td>857</td>
</tr>
<tr>
<td>Se/Sp</td>
<td>85.6%</td>
<td>94.9%</td>
</tr>
</tbody>
</table>
“Levels of Evidence” (Laupacis et al.)

• Prospective data collection specifically to develop or validate rule
• Data collected as part of another study, not specifically undertaken to develop or validate 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 desired error rate (e.g., confidence interval) for sensitivity and/or specificity
• As we said earlier in semester, we can estimate sample size required to estimate a sensitivity or specificity with a desired error rate using a formula for confidence interval around a single proportion
  – Proportion of positive tests among those with disease
  – Proportion of negative tests among those without disease

Consecutive Patients

• Discussion of study sample indicated that most robust design used a consecutive sample of patients in whom you plan to use rule
• In such a sample, p patients will have disease for every 1-p patients who do not (where p equals prevalence in sample)
• Using this design, total number of patients you will need to sample is larger of N_d/p and N nondis/(1-p)
• Rule of thumb: require a minimum of 10 patients with outcome and 10 patients without outcome for every predictor variable used in rule
  – Can only serve to increase sample size; can never serve to reduce 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 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 gold standard in everyone or adopt appropriate sampling / analytic techniques if gold standard is applied in only a subset of subjects

Additional Data

- In addition to gold standard and predictors, include:
  - Demographic and clinical characteristics
    - Test performance may depend on age, gender, and other patient characteristics that might make predictive value of 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 site was teaching or nonteaching
VI. Construction of Rule

- "Eyeball" - Useful to get sense of 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. simplified Walsh rule:
• Single points are assigned to 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>

• 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 best predictor of disease and dividing entire study population into two groups: one with 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 remaining predictors
• Each path along tree represents a sequence of clinical findings and defines a patient subgroup (and associated probability of disease)
• Software is available
If 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 first branching point?

Clinical Information

- Clinical information supplied by a predictor depends on:
  - Odds ratio / relative risk / discriminant score AND
  - Fraction of people in whom the predictor is present
- Predictor that indicates a 95% probability of disease but is present in only 1% of population generally has less information than a predictor that indicates a 50% probability of disease but is present in 10% of population

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 Probability of Disease

- Proceeds in 2 Steps
  - Step 1. Use Estimated Coefficients and Explanatory Variables to Calculate a Risk Score "S"
    - Where \( S = \alpha + \sum \beta_i X_i \)
      - \( \alpha \) = Intercept
      - \( \beta_i \) = Coefficients from logistic regression
      - \( X_i \) = 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 \( -\infty \), predicted probability approaches 0; when \( S \) approaches \( +\infty \), predicted probability equals 1
- When \( S = 0 \), predicted probability = 0.5
Calculating 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 absolute value of risk scores for probabilities (p) less than 0.5 (e.g., risk score representing a probability of 90% is 2.1972246)

Create a Risk Scoring System

• Creating a risk scoring system based on values of 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 we develop a prediction rule in population with a probability of disease of 10% and want to use it in population with a probability of disease of 5%
- Would it be accurate in latter population?
  - Could be, if primary reason for difference was a difference in distribution of risk factors (e.g., fewer patients develop tonsillar exudates, fewer have a fever history, etc.)

Moving Rule When Prevalence Differs

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

- Centor et al. subtracted approximately 1.3 from intercept to modify rule for 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>Tosillar 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>(\leq 0)</td>
<td>1 - 2.5%</td>
<td>No further testing or Rx</td>
</tr>
<tr>
<td>1</td>
<td>5 - 10%</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>11 - 17%</td>
<td>Culture all; Rx for positives</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 †

• Calculating $\Sigma \beta_i X_i$ tedious and likely a disincentive to use of prediction rules
• Often avoided by constructing a point system
  – System assigns integer points to each level of each risk factor to approximate (relative) $\Sigma \beta_i X_i$
• Risk estimates derived from reference table that 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 risk factors and calculate $\beta_i X_i$ for reference values
   – Calculate scores
2. Translate scores into points
3. Determine risks associated with point totals

Reporting on Construction of Rule (Laupacis et al.)

• Adequately describe and justify mathematical technique used to derive rule
• Address whether or not you avoided problem of overfitting data with too few events per predictor variable
Reporting on Construction of Rule (2)

- Specify how variables were selected (e.g., did you use a preliminary screen based on univariate association and reliability?)
  - Prespecify predictors that will be used in 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 kappa statistic or correlation coefficient
  - Values less than 0.6 generally represent lack of agreement
  - Predictors with low reproducibility should not be included in 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

• Ability to give different scores to those with and without disease
  – e.g., to assign generally lower scores to those without disease and generally higher scores to those with disease
• Measures of discrimination
  – Sensitivity and Specificity
  – ROC Analysis

McGinn et al ROC Curve

<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 Only Criterion for a Good Prediction

• Example of a perfectly discriminating, but in some sense mistaken, prediction
  – If 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 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 same score to every day, whether it rains or not)

Calibration

• Calibration is a measure of accuracy of predicted probabilities of disease
  – e.g., degree to which observed and predicted probabilities are equal

• Because it is a property of predicted probabilities and not scores like serum creatinine or hemoglobin levels, does not play a role in evaluation of diagnostic test characteristics

• Could play a role in evaluation of a physician’s pre-test probabilities or of post test probabilities

Types of Calibration

• (At least) two types of calibration:
  – Calibration in the large
  – Calibration in the small
Calibration in the Large

• Property of full sample
• Calculated by comparing observed probability in full sample with average predicted probability in full sample (i.e., average of each of predictions)
  – e.g., if 100 out of 1000 patients have outcome being predicted and average predicted probability is 10%, prediction rule is perfectly calibrated in the large
• For sample in which logistic regression is estimated, results are always perfectly calibrated in the large (i.e., average of predicted probabilities equals average probability in 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>
</tbody>
</table>

Avg Prob: 30% 30% 30%

• 2 rules have identical calibration in large, but rule 2 is better than rule 1

Calibration in the Small

• Property of subsets of sample
  – Calculated by comparing observed probability in each subset with average predicted probability in 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 5% predicted probability, 5% of time it rains;
  – On days with 50% probability, 50% of times it rains;
  – On days with 95% probability, 95% of times it rains
Why is calibration in the small a property of subsets of sample rather than of individual observations in sample?

![Calibration Curve: Calibration in the Small](image)

- Horizontal axis
- Vertical axis
- 45° line

Steps in Plotting Calibration in the Small

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

Roc area: 0.72

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

Analogs of Calibration in the Large and Small

- In describing “stability” as a good property for judging accuracy of microwave/caesium vs laser/strontium atomic clocks:
  - “…if you have your wristwatch, and one day you are one second late, and one day one second early [analogs of calibration in the small], then your clock is not stable. But it could still have good accuracy if over a million days the time is correct [analogs of calibration in the large]”


- Microwave/caesium lose 1 second / 100m years;
laser/strontium lose 1 second / 300m years
- On the horizon: ion clocks, 1 second / few billion years
Calibration Statistics

- Logistic regression - Hosmer and Lemeshow or Pearson
- Yates Decomposition

Example Data

<table>
<thead>
<tr>
<th>Obs#</th>
<th>Cure</th>
<th>Inftype</th>
<th>Severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs#</th>
<th>Cure</th>
<th>Inftype</th>
<th>Severe</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

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

Logit estimates

<table>
<thead>
<tr>
<th></th>
<th>Number of obs</th>
<th>LR chi² (2)</th>
<th>Prob &gt; chi²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>10.92</td>
<td>0.0042</td>
</tr>
</tbody>
</table>

Log likelihood = -8.40  Pseudo R² = 0.3939

cure | OR   | Std Err. | Z   | P>|z|
Inftype | 22.07 | 35.94    | 1.90 | 0.057 |
Severity | 0.3240 | 0.1757   | -2.08 | 0.038 |

Iroc,nograph

Logistic model for cure

number of observations = 20
area under ROC curve = 0.8750

ROC Curve for Cure Example

logit cure inftype severity
Iroc
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 chi2(2) = 1.95
- Prob > chi2 = 0.3768

---

Pearson Chi² Statistic

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

- Number of observations = 20
- Number of covariate patterns = 12
- Pearson chi2(9) = 2.54
- Prob > chi2 = 0.9798

---

Calibration and Discrimination, Examples

- Is following weather person well discriminating and well calibrated?
  - Example 1: Every day, 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
Calibration and Discrimination, Example 2

- Is following weather person well discriminating and well calibrated?
  - Example 2: Every day, 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, Example 3

- Is the following weather person well discriminating and well calibrated?
  - Example 3: Every day, weatherperson predicts there is a 50% chance of rain (and in truth it rains 5 out of every 10 days).

Calibration and Discrimination, Example 4

- Is the following weather person well discriminating and well calibrated?
  - Example 4: Every day, 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.
  - What would you say if on 16.7% of all days weatherperson said 95%? 
    \[
    0.167 \times 0.95 + 0.833 \times 0.05 = 0.2
    \]
Calibration and Discrimination

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

<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>Rule 1</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>Rule 2</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>Rule 3</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 points on a calibration curve are clustered together, discrimination cannot be very good
• When points are pushed towards both ends of calibration curve (large fractions of predictions between 0 and 20% and large fractions between 80 and 100%, discrimination will be reasonably good

Calibration and Discrimination (2)

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

Example: Location of 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>
Conclusions: Location of Cluster

- Even though rules appear similar, in that there are absolute 5% differences between each stratum:
  - ROC area reaches a minimum when they are centered at 50%
  - ROC areas are equal as clusters symmetrically approach two ends of probability distribution, and are increasing
  - When a rule is well calibrated, and when points on calibration curve are clustered within a small region, rule’s discriminating ability will be small

Calibration In Diseased and Nondiseased Individuals

- When we construct calibration plot, we rank order observations by predicted probability
- Why don’t we rank order them by observed outcome??
Diseased and Nondiseased Individuals

- Calibration is a property of predicted probabilities, not of known disease status
  - If we accurately characterize a probability of disease as 5%, 95 of 100 will not have disease; if we accurately characterize a probability of diseases as 95%, 5 of 100 will not have disease
  - Implication: For very good prediction rules:
    - Subjects without disease should not be expected to have a predicted probability of disease of 0%
    - Subjects with disease should not be expect to have a predicted probability of disease of 100%

Incremental Information and Costs in Different Specifications of a Rule

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

Steyerberg et al.*: Prediction Model Performance

- "reporting discrimination and calibration will always be important for a prediction model"
- Model discrimination “will commonly be most relevant for research purposes”
- "calibration is important if model predictions are used to inform...making decisions"
  - Cox "recalibration parameters“ and validation plots may be better than H&M test
- "novel measures for reclassification and clinical usefulness can provide valuable additional insight"

VIII. Assessment of Validity *

• Predictive validity refers to quality of rule's predictions in sample in which it was developed and in new samples
• Most prediction rules lose accuracy when used in patients who were not included in 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 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 derivation and validation population (case-mix)
• Differences in definitions of predictors and outcome variable and measurement methods between derivation and validation populations
• Improvement over time in measurement techniques, which may affect strength of a predictor

Apparent Differences

• Apparent differences may be due to 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, suggested that validation sample should contain at least 100 events and 100 nonevents to detect substantial changes in accuracy (for example, a 0.1 change in ROC area) with 80% power
Internal Validity

• Quality of prediction in 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 derivation dataset) used to address external validation concerns of overfitting or "optimism"

Internal Validity (2)

• Overfitting is modeling of relationships that are specific to 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 resulting model in bootstrap dataset as well as full derivation dataset and calculate each area under ROC curve
  – Interpret difference in mean areas between bootstrap and derivation datasets as a measure of "optimism"

External Validity

• Quality of prediction in a new validation dataset
• Report information about study population so its generalizability can be assessed. Data include:
  – Medical setting from which patients were drawn,
  – Age, gender, and clinical characteristics of patients
"Levels" of External Validation

- **Temporal validation**
  - Tests generalizability of a prediction rule over time, typically using same physicians or investigators as in development study, in same institution(s), and in similar patients

- **Geographical validation**
  - Tests generalizability of a prediction rule in a patient population that is similarly defined as development population, but in hospitals or institutions of other geographical areas

"Levels" of External Validation (2)

- **Domain validation**
  - Evaluates 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)

- Level of evidence of validation increases as we go down 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 rule by combining information from original rule with information from validation population

• Six general strategies
  1. If prevalence differs dramatically between study populations, adjust intercept of original prediction rule (e.g., updating Centor strep rule)
  2. "Logistic recalibration": Adjust Intercept / coefficients with a single correction factor estimated from data of new patients in validation set
     * These two methods may improve calibration, but cannot improve discrimination, because recalibration does not affect rankings

Updating Prediction Rules (2)

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

IX. Provision of Information That Helps Clinicians Identify a Course of Action after Applying 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 $< 20\%$; 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, $11\% < p \leq 35\%$: Obtain cultures and treat if positive culture
  - For scores $\geq 4$, $p > 50\%$: Treat without culture

Course of Action (II)

- McGinn (revised Walsh algorithm) recommends:
  - Empiric therapy for all patients with a score of 2+ ($\geq 55\%$ post-test probability) and rapid testing for patients with scores of 0 or 1 (post-test probability $> 15\%$)

ACP Guidelines

- ACP, AAFP, and CDC consider it reasonable not to perform a throat culture or rapid antigen-detection test if all 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 positive rapid test if 2 or 3 features are present
  - Treat without testing if three or four features are present
- More concerned with cost and loss to follow-up than with resistance
IDSA Guidelines

- Do not test if there are "clinical and epidemiological features that strongly suggest a viral etiology (e.g., cough, rhinorrhea, hoarseness, and oral ulcers"
  - Only use of Centor rule
- Rapid test and/or culture should be performed before any treatment is initiated. Negative RADT test should be backed up with culture in children ≥3 and adolescents, but not in adults (under usual circumstances)
- Positive rapid tests do not need to be backed up
- Therapy should not be initiated until either rapid test or culture is positive
- Penicillin or amoxicillin is recommended drug of choice for those non-allergic to these agents

X. Assessment of Whether 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 probability calculated with a prediction rule
    - e.g., Sinuff et al. found that ICU physicians more accurately discriminated between survivors and nonsurvivors in first 24 hours of ICU admission than did ICU survival prediction rules
  - They believe their patients are different from those used in development of rule
  - They are afraid they won't apply rule correctly
  - They feel false negative rate is too high

Sensibility

- Physicians also may not use 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 purpose of rule
  - No obvious items should be missing (or their absence is adequately explained)
  - method for aggregating component variables should appear reasonable
Assessment of Whether Rule Affects Practice (II)

• Providers may not use a rule because:
  – Rule is not user-friendly or significantly extends time of usual clinical encounter, e.g., rule:
    • Includes variables that are not collected in daily practice
    • Require 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 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 Ottawa ankle rule
  – Best predictors whether a rule would be used in practice were 1) familiarity acquired during training, 2) confidence in usefulness of rule, and 3) user-friendliness of 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
  - Ideal design use a cluster randomized trial in which physicians or care units are randomized to either use of rule or use of "care or clinical judgment as usual"
  - Alternate design: before/after study within same physicians or care units (temporal changes may compromise 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 randomization groups

Perform Impact Analysis (II)

- Consider inclusion criteria
- Ideal endpoints are clinically relevant process parameters, patient outcomes, and cost-effectiveness
- Use blinding
- Estimate sample size
- Understand 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

TRIPOD Reporting GUIDELINES

- TRIPOD reporting checklist (distributed on CANVAS)

- TRIPOD explanation and elaboration