

Developing a Prediction Rule

Henry Glick

Epi 550

January 30, 2020



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)
- 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

* Toll DB, Janssen KJM, Vegouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: A review. J Clin Epi. 2008;61:1085-94



Diagnostic Tests vs Prediction Rules *

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

* Collins GS, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *Ann Intern Med.* 2015; 162: 55-63. doi:10.7326/M14-0697.

- ### 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
- * Walsh BT et al. Recognition of streptococcal pharyngitis in adults. *Arch Intern Med.* 1975;135:1493-7.
 * McGinn TG, et al. Validation and modification of streptococcal pharyngitis clinical prediction rules. *Mayo Clin Proc.* 2003;78:289-93.
 * Centor et al. The diagnosis of strep throat in adults in the emergency room. *Med Decis Making.* 1981;1:239-46.
 * McIsaac et al. Empirical validation of guidelines for the management of pharyngitis in children and adults. *JAMA.* 2004;291:1587-95.
-

Clinical Problem

- 105 outpatient office visits per 1000 US 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.

Cooper RJ, et al. Principles of appropriate antibiotic use for acute pharyngitis in adults: background. Ann Int Med. 2001;134:509-17



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

Cooper RJ, et al. Principles of appropriate antibiotic use for acute pharyngitis in adults: background. Ann Int Med. 2001;134:509-17



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 LRC, 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%

Black CM, et al. Head-to-head multicenter comparison of DNA Probe and nucleic acid amplification tests for chlamydia trachomatis infection... J Clin Microbiol. 2002;40:3757-63



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

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



Distinct Predictor Variables. Pharyngitis Example

Walsh et al.	Centor et al.
Loss of hearing	--
Tinnitus	--
Ear or sinus pain	--
--	Duration of symptoms
--	Age
--	Fever history
--	Difficulty swallowing



Mclsaac 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 among 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 more likely if all subjects with disease have very high pre-test probabilities (e.g., patients with many signs and symptoms) and all subjects without disease have very low probabilities (e.g., undergraduates or medical students) with no signs and symptoms



Study Sample Issues

- Spectrum bias
- Levels of evidence
- Sample size



IVa. Spectrum Bias

- Inclusion of a nonrepresentative sample of patients in whom the test will be used in cases where sensitivity and specificity are not independent of prevalence



Spectrum Bias Setup: Rapid Antigen Test *

	D+	D-
RADT+	368	87
RADT-	97	1124
	465	1231
Se & Sp	79.1	91.3

* Rimoin AW, et al. The utility of rapid antigen detection testing for the diagnosis of streptococcal pharyngitis in low-resource settings. *Int J Inf Dis.* 2010;14:e1048-53



Spectrum Bias, Rapid Antigen Test

	D+		D-	
	Cent 0/1	Cent 3/4	Cent 0/1	Cent 3/4
RADT+	181	187	44	43
RADT-	66	31	813	311
	247	218	857	354
Sens	73.3	85.8	Spec	94.9
p =	0.001		0.000	

- Detection of spectrum bias based on assessment of sensitivity and specificity, not on PPV or NPV
 - i.e., differences in PPV and NPV between Centor 0/1 and 3/4 neither necessary nor sufficient



Spectrum Bias, RADT (2)

- Suppose we'd used a common design for assessing test characteristics (e.g., using people with obvious cancer as cases and medical students as controls):
 - 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 0/1) to construct D- sample?

	D+	D-
RADT+	187	44
RADT-	31	813
Tot	218	857
Se/Sp	85.8%	94.9%



Spectrum Bias, RADT (3)

- Test characteristics from following feasible tables "better" than table derived using "common design"

	Mixed Pop		Centor 0/1		Centor 3/4	
	D+	D-	D+	D-	D+	D-
RADT+	368	87	181	44	187	43
RADT-	97	1124	66	813	31	311
Tot	465	1231	247	857	218	354
Se/Sp	79.1	91.3	73.3	94.9	85.8	87.9



More Generally.....

Centor Score	Prevalence N (%)	Rapid test Sensitivity % (95% CI)
0,1 (n=169)	23 (14)	61 (54-66)
2 (n=143)	29 (20)	76 (69-83)
3 (n=122)	53 (43)	90 (93-100)
4 (n=64)	33 (52)	97 (93-100)
Overall	137 (28)	84 (81-87)

Mantel-Haenszel trend test, p=0.001

DiMatteo, et al. The relationship between clinical features of pharyngitis and the sensitivity of a rapid antigen test: evidence of spectrum bias. Ann Emerg Med. 2001; 38:648-52.



IVb. "Levels of Evidence" (Laupacis et al.)

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



IVc. 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
- 2x2 topics lecture notes show how to estimate sample size required for measuring 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



Sample Size and Consecutive Patients

- Discussion of study sample indicated that most robust design uses a consecutive sample of patients in whom you plan to use rule
- In such a sample, approximately 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_{dis}/p and $N_{nondis}/(1-p)$
- Separate 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

- Walsh et al.

Score	Probability (%)
-10 - 0	1.8
1 - 10	4.6
11-20	18.0
21-30	19.0
31-40	22.0
41+	100.0



Revised Walsh Risk Scoring System

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

Risk factor	Score
Temperature >38.3°C	+1
Exposure to known strep contact	+1
Pharyngeal or tonsillar exudates	+1
Enlarged or tender nodes	+1
Recent cough	-1

- Total score ranges between -1 and +4



Translation of Scores to Probabilities

Score	LR	95% CI	Probability (%)
-1	0.16	0.05 - 0.42	4.6
0	0.62	0.29 - 1.20	15.9
1	2.61	1.49 - 4.44	44.4
2	4.35	1.65 - 11.26	57.1
3+	8.14	1.88 - 35.23	71.4

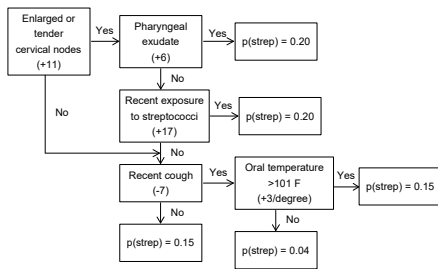


Branching Algorithms / Recursive Partitioning

- 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



Recursive Partitioning Algorithm Adults with Sore Throat



- 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:
 - Likelihood ratio / Odds ratio / relative risk / discriminant score AND
 - Fraction of people in whom predictor is present
- Predictor that indicates a 95% probability of disease but is present in only 1% of population generally less informative than predictor that indicates a 50% probability of disease and is present in 15% 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 Predicted Probability of Disease

- Proceeds in 2 Steps

Step 1. Use Estimated Coefficients and Explanatory Variables to Calculate a Risk Score "S"

$$\text{Where } S = \alpha + \sum \beta_i X_i$$

α = Intercept

β_i = Coefficients from logistic regression

X_i = Predictor Variables



Strep Pharyngitis Coefficients

Variable	Coef.
Intercept	-2.69
Tonsillar exudates	1.04
Swollen/tender anterior cervical nodes	1.00
Cough	-0.95
Fever history	0.89

- For a person with tonsillar exudates and fever history risk score S equals:

$$-2.69 + 1.04 + 0.89 = -0.76$$



Risk Score S

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



Calculating Predicted 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

Probability (%)	Risk Score S
10	-2.1972246
20	-1.3862944
30	-.84729786
40	-.40546511
50	0

- 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
 - Reasonable to assume equal weighting



Equally Weighted Risk Scoring System

- Centor et al.

Number of Features Present	Probability (%)
0	2.5
1	6 - 6.9
2	14.1 - 16.6
3	30.1 - 34.1
4	55.7



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 the predicted probability of disease be accurate in latter population?
 - Could be accurate if risk among patients without risk factors remains 2.5% and if primary reason for difference was a lower prevalence of risk factors (e.g., fewer patients develop tonsillar exudates, fewer have a fever history, etc.)



Moving Rule When Prevalence Differs

- Unlikely to be accurate if individuals without any risk factors in new population (e.g., ones with a score of 0) have a risk for disease that differs from 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)
- Unlikely to be accurate if odds ratios for risk factors differ (i.e., changes in coefficients from logistic regression)



Intercept Shift Revision of Rule

- When moving rule to a population with a lower prevalence of disease, Centor et al. subtracted approximately 1.3 from intercept to modify rule for new setting

Score	~S	Probability (%)
-1	-4.9636	1
0	-3.9774	2
1	-3.0074	5
2	-2.0492	11
3+	-1.071	25



McIsaac Score

Feature	Score
Temperature >38°C	1
Absence of cough	1
Swollen and tender anterior cervical nodes	1
Tonsillar swelling or exudates	1
Age	
3-14	1
15-44	0
45+	-1



Mclsaac Scoring System

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



General Principles for Generating Risk Scoring Systems †

- Calculating $\sum \beta_i X_i$ 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) $\sum \beta_i X_i$
- 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 VIII blood pressure categories

† Sullivan LM, Massaro JM, D'Agostino Sr RB. Tutorial in biostatistics: Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Statist. Med.* 2004;23:1631-60



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 indeterminate disease status



Discrimination

- Ability to assign 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 disease
 - Discrimination is a property of scores
 - Given that predicted probabilities can be interpreted as scores, it applies to probabilities as well
- Measures of discrimination
 - Sensitivity and specificity
 - ROC analysis
 - ROC area



Interpretation of ROC Area

- ROC areas can range between 0.5 (area under 45° line of no information) and 1.0 (area under ROC curve of a dichotomous test that has 100% sensitivity and specificity)
 - Area of 0.5 represents no ability to discriminate risk
 - Test assigns a similar distribution of scores to those in whom disease is present and those in whom disease is absent
 - Area of 1.0 represents perfect discrimination
 - No overlap in distribution of scores assigned to those in whom disease is present and those in whom disease is absent



Interpretation of ROC Area (2)

- Although curves with ROC areas of 0.5 and 1.0 are clearly distinguishable, there is little systematic information available about benefit of small increases in area under ROC curve (e.g., an increase from 0.75 to 0.77)
 - But, tests with larger areas under their ROC curve in general are more discriminating than are tests with smaller areas



Interpretation of ROC Area (3)

- Technically, ROC area equals probability that rule will correctly rank any randomly selected pair of persons, one in whom outcome of interest is present and one in whom it is absent
 - Nonparametric area represents p-value we derive from a Wilcoxon rank sum test
 - How often do pairs of patients walk into a provider's office; declare that one has disease while other does not, and then ask "which of us has a higher test score?"



Interpretation of ROC Area (4)

- ROC area is used as a measure of discrimination in many applications other than diagnostic test evaluation
 - C-statistic that is routinely reported by SAS as an index of discriminating ability of fitted logistic regressions models equals nonparametric area under logistic regression's ROC curve
 - Similarly, lroc command in STATA that can be run after logistic regression reports same area



McGinn et al ROC Curve

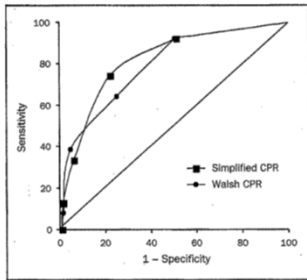


Figure 1. Receiver operating characteristic curves for Walsh and simplified clinical prediction rules (CPRs).



Mclsaac* Sensitivity and Specificity

Sample	Sensitivity (%)	Specificity (%)
All	100	93.2 (90.8 - 95.1)
<18	100	90.3 (86.4 - 93.4)
18+	100	96.5 (93.5 - 98.4)

* 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

Obs #	Truth	Pred Rule 1	Pred Rule 2
1	0	1.0	0.0
2	0	1.0	0.0
3	0	1.0	0.0
4	0	0.0	0.0
5	0	0.0	0.0
6	0	0.0	0.0
7	0	0.0	0.0
8	1	0.0	1.0
9	1	0.0	1.0
10	1	0.0	1.0
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?

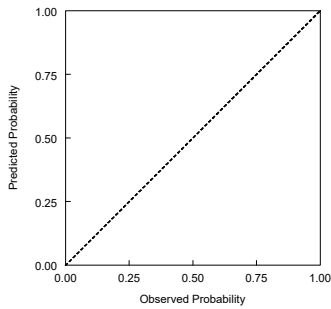


Why is calibration in the small a property of subsets of sample rather than of individual observations in sample?

Because a group of like patients may each have a 30% chance of disease (0.3), but in truth, 3 will have disease (1) and 7 won't (0)



Calibration Curve: Calibration in the Small



- 2 axes (Unaware of consensus)
- 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)



Step 3: Divide Rank Ordering Into Groups

- Hosmer and Lemeshow indicate goal of division is creation of equal sized groups, not, for example, to use deciles or ventiles of risk
 - e.g., lowest 5% of distribution, 5 to 10% of the distribution, etc. NOT observations with probabilities less than 5%, probabilities between 5% and 10%, etc.
- Noise can be added to calibration test if lots of tied predicted probabilities and (to keep group size equal) tied cases fall within more than 1 group (for example with Stata xtile command)
 - As far as possible, observed probability for ties should be equal within each group in which they are included....



Ties

Subgroup	Pred Prob (%)	Truth (Good)	Truth (Bad)	Truth (Bad)

	30	0	0	0
Subgroup 4	33.3	0	0	0
	33.3	0	1	0
	33.3	1	1	0
	33.3	0	0	0
	33.3	0	0	1
Subgroup 5	33.3	1	0	1
	35	0	0	0



Calibration in the Small

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

Roc area: 0.72

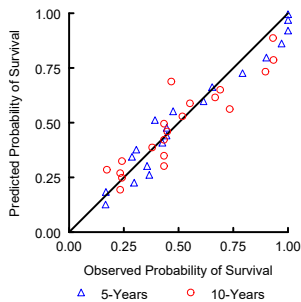


Step 6: Plotting Observed and Predicted

- Expectation that there will be dispersion around the 45° degree line
 - Points with larger vertical (or horizontal) distances from 45° degree line worse than points with smaller distances
- Most problematic when dispersion has a pattern
 - E.g., generally below 45° degree line for lower predicted probabilities and above 45° degree line for higher predicted probabilities OR “happy face” OR “sad face” OR sigmoidal



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 [analog 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 [analog of calibration in the large]”

Morelle R. Optical lattice atomic clock could 'redefine the second,' 07/09/13, <http://www.bbc.co.uk/news/science-environment-23231206>

- Microwave/caesium lose 1 second / 100m years; laser/strontium lose 1 second / 300m years
- (2016) optical single-ion ytterbium clock: accuracy 100 times better than cesium clocks



Calibration Statistics

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



Example Data

Not Cured				Cured			
Obs#	Cure	Inftype	Severe	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


	Not Cured	Cured
Infection type=1	30%	70%
Severity	3 (1.63)	1.4 (1.35)



logistic cure inftype severity

Logit estimates Number of obs = 20
 LR χ^2 (2) = 10.92
 Prob > χ^2 = 0.0042
 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

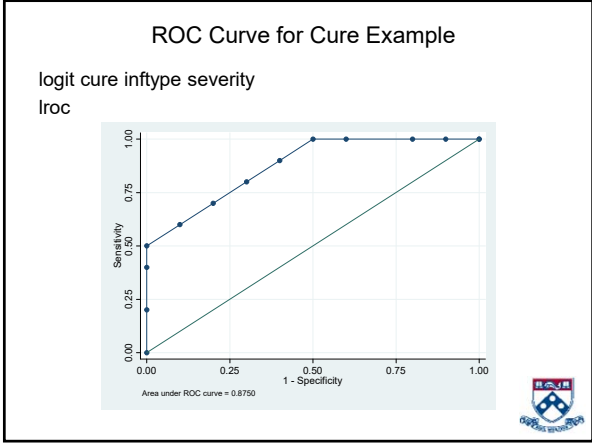


lroc,nograph

Logistic model for cure

number of observations = 20
 area under ROC curve = 0.8750





Hosmer and Lemeshow Statistic

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

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.1936	0	0.4	5	4.6	5
2	0.4595	3	2.1	3	3.9	6
3	0.7914	2	2.8	2	1.2	4
4	0.9830	5	4.7	0	0.3	5

Number of observations = 20
Number of groups = 4
Hosmer-Lemeshow chi2(2) = 1.95
Prob > chi2 = 0.3768

- p>0.05: no significant evidence of lack of calibration



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%?

$$0.2 = (0.05 * 0.833) + (0.95 * 0.167)$$



Calibration and Discrimination

- Rules can be well discriminating (or poorly discriminating) when calibration in the small is either good or bad
- Rules can be well calibrated (or poorly calibrated) when discrimination is small or large
- There is nothing in the ROC curve (discrimination) that tells us anything about calibration
- But is there information in the calibration curve that tells us something about discrimination?



What does calibration curve look like for a highly discriminating and well calibrated prediction rule?

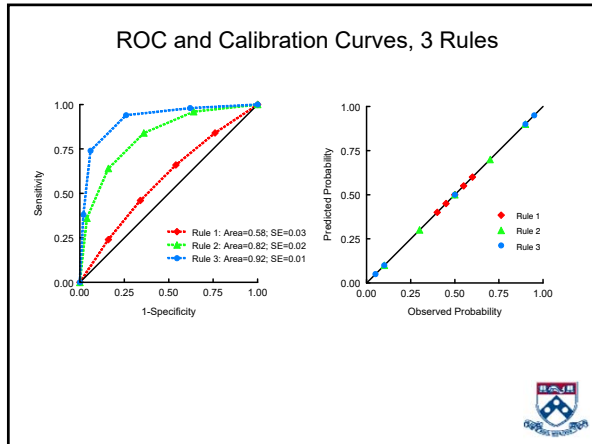


Example: Discrimination for Well Calibrated Rule

		Strata					
		1	2	3	4	5	Total
Rule 1	D+	40	45	50	55	60	250
	D-	60	55	50	45	40	250
	Pred P	40%	45%	50%	55%	60%	50%
Rule 2	D+	10	30	50	70	90	250
	D-	90	70	50	30	10	250
	Pred P	10%	30%	50%	70%	90%	50%
Rule 3	D+	5	10	50	90	95	250
	D-	95	90	50	10	5	250
	Pred P	5%	10%	50%	90%	95%	50%

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





Conclusions: Discrimination for Well Calibrated Rule

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

Must a well-discriminating rule be well calibrated?

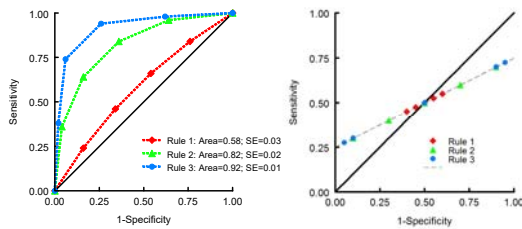
Example: Discrimination for Poorly-Calibrated Rule

		Strata					
		1	2	3	4	5	Total
Rule 1	D+	40	45	50	55	60	250
	D-	60	55	50	45	40	250
	Pred P	45%	47.5%	50%	52.5%	55%	50%
Rule 2	D+	10	30	50	70	90	250
	D-	90	70	50	30	10	250
	Pred P	30%	40%	50%	60%	70%	50%
Rule 3	D+	5	10	50	90	95	250
	D-	95	90	50	10	5	250
	Pred P	27.5%	30%	50%	70%	72.5%	50%

- All 3 prediction rules have poor calibration in the small, although all 3 rules are calibrated in the large



ROC and Calibration Curves, 3 Rules



Conclusion: Discrimination for Poorly-Calibrated Rule

- Discrimination has to do with points on calibration curve falling close to 0 and 1 on observed axis, not with points on calibration curve falling near the 45 line
- Independent of whether points on calibration curve are near 45 degree line, discrimination cannot be very good when points are clustered together on calibration curve
- Independent of whether points on calibration curve are near 45 degree line, discrimination will be good when points are pushed towards both ends of calibration curve



When points on calibration curve are clustered (e.g., between 5% and 25%, between 40% and 60%, and between 75% and 95%), does location of cluster affect discrimination?

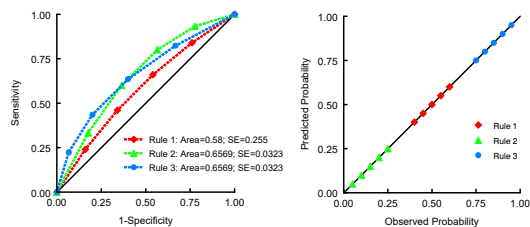


Example: Discrimination and Location of Cluster

		Strata				
		1	2	3	4	5
Rule 1	D+	40	45	50	55	60
	D-	60	55	50	45	40
	Pred P	40%	45%	50%	55%	60%
Rule 2	D+	5	10	15	20	25
	D-	95	90	85	80	75
	Pred P	5%	10%	15%	20%	25%
Rule 3	D+	75	80%	85%	90%	95
	D-	25	20%	15%	10%	5
	Pred P	75%	80%	85%	90%	95%



ROC and Calibration Curves, 3 Rules



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 centered at 50%
 - ROC areas are equal as clusters symmetrically approach two ends of probability distribution, and are increasing
 - Whether or not a rule is well calibrated, 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??

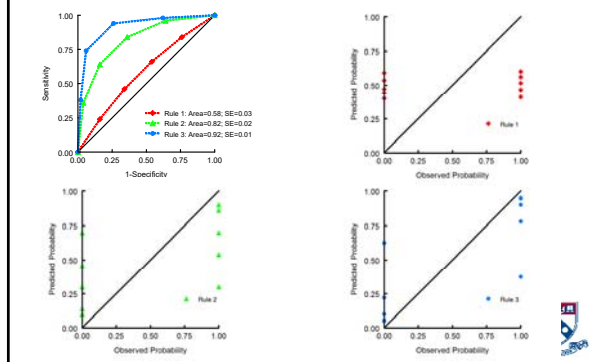


Diseased and Nondiseased Individuals

- 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 disease as 95%, 5 of 100 will not have disease
 - Implication: Even 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%



Suppose We Did Order Lexicographically By Disease Status and Observed Probability???



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”

* Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, et al. Assessing the performance of prediction models: a framework for traditional and novel Measures. *Epidemiology*. 2010; 21:126-38.



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

* Material in sections VIII, X, and XI drawn in part from Toll DB, Janssen KJM, Vegouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: A review. J Clin Epi. 2008;61:1085-94.



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

- Tompkins' Decision Rule (Ann Int Med, 1977):
 - 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 ≤ 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 ≥ 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+ (>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 (eg, 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

* Reilly & Evans. Ann Int Med. 2006;144:201-9



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 *

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

EBM Working Group cited in Reilly & Evans



Standards of Evidence for Prediction Rules *

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

EBM Working Group cited in Reilly & Evans



TRIPOD Reporting GUIDELINES

Colins GS, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. Ann Intern Med. 2015; 162: 55-63. doi:10.7326/M14-0697

- TRIPOD reporting checklist (distributed on CANVAS)

Moon KGM, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015; 162: W1-W73. doi:10.7326/M14-0698

- TRIPOD explanation and elaboration