Screening

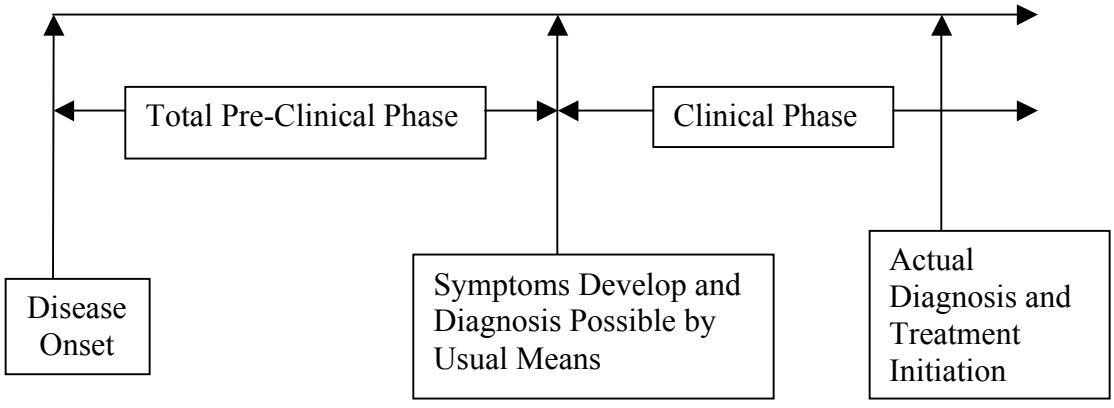
The goal of a screening program is to apply a **simple and inexpensive test** to a large number of persons in order to classify them as likely or unlikely to have a disease of interest. The ultimate goal is to reduce the morbidity and mortality from that disease in the persons that are screened by detecting disease at an earlier stage, where a treatment might have a more beneficial effect. The components of a beneficial screening program include a suitable

1. Disease
2. Test
3. Population

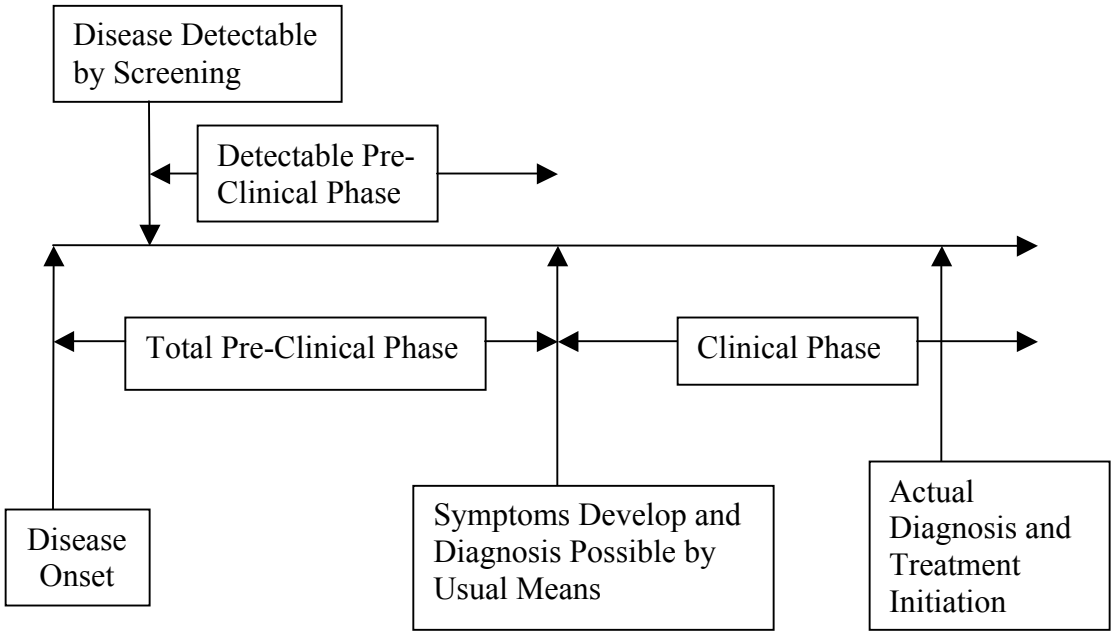
Time Periods

` A disease that is suitable for a screening program should have **serious consequences** (e.g. fatal or severe/prolonged morbidity) to merit the time and cost as the target of a screening program. It must **have a treatment** that, when applied to a case of disease that is detected by screening, is more effective than treatment applied after symptoms when the case is detected by usual means. Finally, the disease should have a **high prevalence in the Detectable Preclinical Phase** (**DPCP).**

Clinical disease begins with the development of signs or symptoms that would lead to a diagnosis of that disease through usual means. Pre-Clinical Disease refers to the existence of disease that has not advanced to a stage where it could be detected by usual means. The **Total Preclinical Phase** begins with the first development of disease (biologic onset) and ends with the development of signs/symptoms and the diagnosis of disease by usual means. These time periods are depicted in the following figure:



The **Detectable Preclinical Phase (DPCP)** refers to a portion of the TCPP and begins when the disease can be detected by the screening test. The length of the DPCP depends on the screening test's ability to detect the disease before signs and symptoms develop. A long DPCP enhances a screening program's chances to detect disease well before it would normally be diagnosed. The following figure shows the timing of the DCPC.



Screening Test

A suitable screening test is one that accurately detects the presence of absence of pre-clinical disease. This is usually measured by two performance measure

1. Sensitivity = P(Test + | Pre-Clinical Disease Exists)

2. Specificity = P(Test - | Pre-Clinical Disease Does Not Exist)

These measures can be estimated from the following 2x2 table

	Pre-Clinical Disease		
	Present	Absent	Total
Test +	A (True Positives)	B (False Positives)	A+B
Test -	C (False Negatives)	D (True Negatives)	C+D
Total	A+C	B +D	

Sensitivity = A/(A+C)

Specificity = D/(B+D)

In addition to having a high Sensitivity and Specificity, a suitable screening test should be low in cost, painless, and not cause morbidity or mortality.

As an example, shows the performance of a screening test of physical examination and mammography for the detection of breast cancer from the Health Insurance Plan (HIP) of Greater New York (Shapiro et al. *Lead Time in Beast Cancer Detection and Implications for Periodicity of Screening*. Am J Epidemiol 100:357-66. 1974 and Hennekens and Buring. *Epidemiology in Medicine* Little Brown 1987 **),**

	Pre-Clinical Disease		
	Present	Absent	Total
Test +	132	983	1115
Test -	45	63650	63695
Total	177	64633	64810

Sensitivity = 132/177 = 0.75
Specificity = 63650/64633 = 0.98

The Sensitivity and Specificity of a Test are characteristics of the screening test. Therefore, one might expect that these values would not change is a screening test were applied to different populations. However, it may be possible that the screening test's ability to detect pre-clinical disease may depend on the severity of the pre-clinical disease. For example, pre-clinical disease states that are about to become clinical disease may be more easily detected by the screening test than less advanced pre-clinical disease. Therefore, the value for the sensitivity of a test may vary of populations that differ in severity of the cases of pre-clinical disease.

Screening Program

A Screening Program involves using a particular screening test in a particular population of asymptomatic individuals. Process measures that reflect the suitability of a screening program include process measures of the screening test (Sensitivity and Specificity) as well as the following

1. Number of people examined
2. Detected prevalence of disease in DPCP
3. Cost
4. Follow-up treatment and outcome
5. Positive Predictive Value of the Test
6. Negative Predictive Values of the Test

The Positive (PV+) and Negative (PV-) Predictive Values of a test refer to the tests ability to predict the presence and absence of the disease. These measures are defined as

Positive Predictive Value = P(Pre-Clinical Disease Present| Test +)

Negative Predictive Value = P(Pre-Clinical Disease Absent| Test -)

The following table shows these values for the HIP data shown above

	Pre-Clinical Disease		
	Present	Absent	Total
Test +	132	983	1115
Test -	45	63650	63695
Total	177	64633	64810

PV+ = 132/1115 = 0.12
PV- = 63650/63695 = 0.999

The Positive and Negative Predictive Values are **posterior probabilities**. They are predictions of an outcome (pre-clinical disease) that take data (test results) into account. They also depend on the base prevalence of disease in the screened population (**prior probability**). Therefore they depend on characteristics of the screening tests (Sensitivity and Specificity) and also characteristics of the population being screened (prevalence of pre-clinical disease). These dependencies are demonstrated by the following expression of Bays Theorem.

(PV+)/(1-(PV+))= [P(D)/(1-P(D))][sensitivity/(1-specificity)]

(PV-)/(1-(PV-)) = [(1-P(D))/P(D)][specificity/(1-sensitivity)]

Posterior Odds = (Prior Odds)(Likelihood Ratio)

These formulas demonstrate that Positive Predictive of Test depends on prevalence of preclinical disease in a population. Screening in a high risk population will results in higher PV+ and enhanced the suitability of a screening program. One means to increase the prevalence of pre-clinical disease in DPCP of a population is to restrict enrollment of participants in the screening program to those with one of more risk factors for the disease. Another option is to screen a population at an optimum frequency. An initial screen of a population will remove prevalent cases of detectable pre-clinical

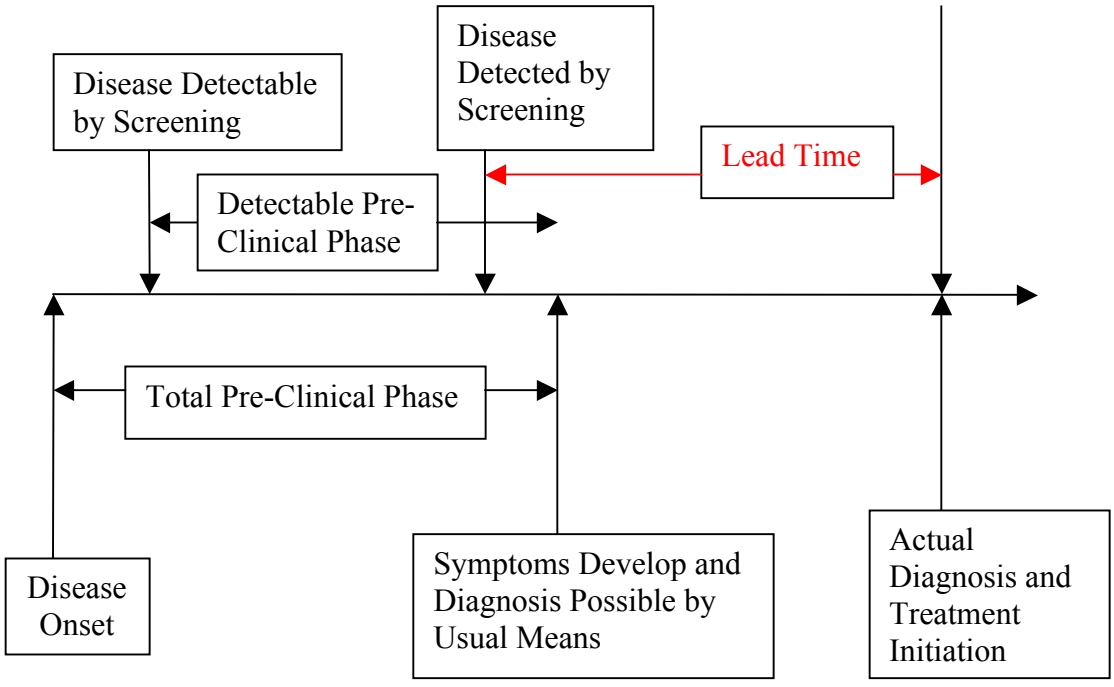
disease. Sufficient time should elapse for incident cases of pre-clinical disease to develop in that population before it re-screened.

Evaluation

The ultimate goal of a screening program is to reduce the morbidity and mortality from that disease in the persons that are screened by detecting disease at an earlier stage, where a treatment might have a more beneficial effect. Therefore, the effect of a screening program can be detected by comparing outcomes from participants in a screening program to similar other individuals who were not part of a screening program. This can be done using the study design options that were discussed in previous lecture notes. For example, with an experimental design, study participants can be randomized to participate in a screening program or to usual care, then mortality rates in each group can be compared from the point of randomization.

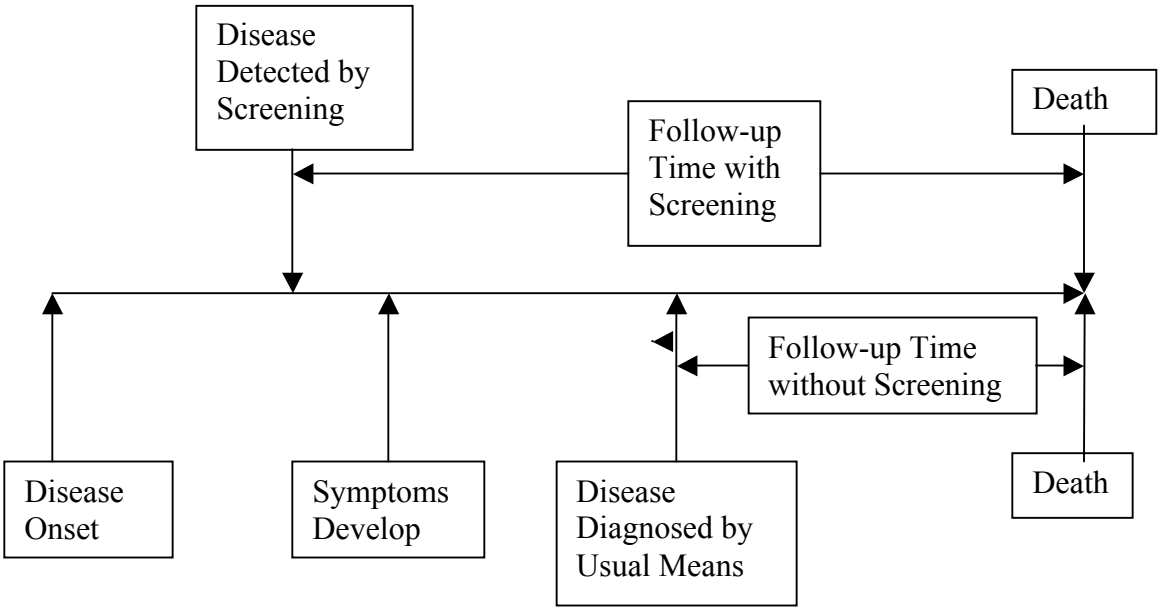
On the other hand, evaluation a screening program with a non-experimental Cohort Study has some potential problems. Individuals who agree to participate in a screening program may not be comparable to those who refuse to participate, providing a potential for confounding. In addition the potential for **Lead Time Bias** and **Length Bias** can be considered.

Lead Time is the additional time an individual lives with knowledge of disease because of earlier diagnosis from screening. A long Lead Time is desirable property for a screening program if early treatment results in better outcomes. The following figure depicts the lead time from a screening program.



Lead Time Bias occurs when the follow-up time for the screened participants does not account for the Lead Time, since individuals in the comparison group will not benefit from a lead time. For example, suppose follow-up for all individuals began at the time of diagnosis. For the screening group follow-up would begin at the time of screen-detected disease diagnosis. This group would show longer follow-up times (and lower incidence rates of death) even if treatment (at any time) had no effect, because of the benefit of Lead Time.

This is depicted in the following figure. The bottom part of the figure describes the life course of an individual without screening. The top part of the figure describes the life course with early detection of disease through screening. However, the figure assumes that early treatment has no benefit and that the individual dies at the same time under both scenarios.



Length Bias occurs when evaluating a screening program because of the expectation that screening may detect prevalent cases of disease in the Detectable Pre-Clinical Phase that have a favorable prognosis. Since prevalence is a function of incidence and duration, an initial screening in a population may detect cases of pre-clinical disease with a long duration. Some of these cases may never develop into clinical cases of disease. Cases of disease detected from a screening program may have better prognosis that those detected by usual means. Therefore, better outcomes in the screened groups may not reflect the effect of the screening program but rather the types of cases that are detected by screening. This potential bias lessens when evaluating the second application of a screening program, since many of the slow growing cases of disease may have been detected by the initial screen.

Ethical issues may also be important to consider when implementing a screening program. For example, screen-detected cases are often subjected to more invasive diagnostic tests or might be treated with therapies with potential serious side-effects. False positive cases receive no benefit from such additional testing or treatment. In addition, the impact on quality of life for true positive cases might not be worth the extra testing and treatment if the expected benefit from early treatment is minimal. For this reason, screening for disease in the very elderly might not be desirable.

Clinical Prediction Rules

Prediction is an integral part of clinical medicine. Prognostic models quantify the likelihood of developing (prognosis) or possessing (diagnosis) an outcome of interest. A **prediction rule** is an algorithm for estimating this likelihood based on values of selected predictors for this outcome. Thus, the main motivations for developing clinical prediction rules include:

1. Predicting a future state of health based on information that is currently available (**Prognosis**), and

2. Predicting a current state of health that is not easily observable based on other information that is available and more easily observable (**Diagnosis**).

Examples of Clinical Prediction Rules

Although a clinical prediction rule could be based on expert opinion, usually these rules are based on empirical relationships (associations in a data set) between other identified predictors and the outcome. Some Common types of clinical prediction rules include:

1. Stratification Patterns
2. Regression Models
3. Point Scoring Systems
4. Neural Networks
5. Other Data Mining Methods

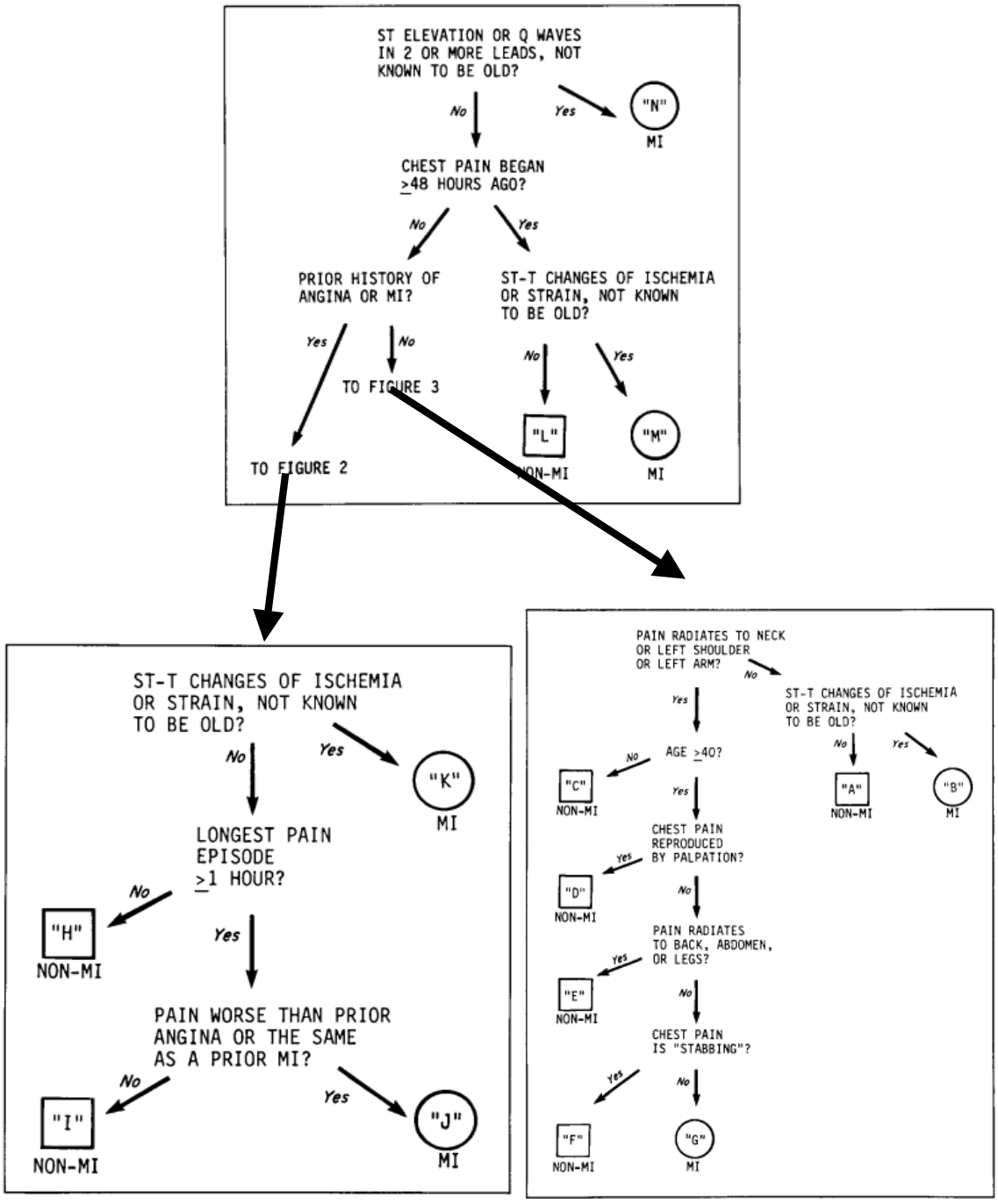
The choice of which type of prediction rule to adopt for a given situation depends on a number of factors, including the number and nature of candidate predictors, the size of the data set, the interrelationships among the predictors, and the potential acceptability of the rule by clinicians.

Stratification Patterns involves grouping the subjects into subgroups (strata) defined by combinations of categories of categorical predictors. The stratum-specific cumulative incidence or the prevalence of the outcome is used as the estimated risk for all patients within that stratum. When the number of categorical predictors is large, then cross-stratification by all predictors usually results in too many subgroups with insufficient number of subjects within most subgroups to provide accurate risk estimates.

Recursive Partitioning (**Classification Trees**) creates an asymmetric stratification pattern by applying a step-wise stratification algorithm to sequential divide subgroups of subjects into smaller subgroups with different estimated outcome risks. This method initially divides the data into two subsets by the single "best" predictor. Each subset in then divided into two additional subsets by the best predictor for that subgroup. The process continues in a recursive manner until there is no evidence that further division of each existing subset would improve prediction. The resulting stratification is

asymmetric, not only in that some portions of the data set undergo more levels of stratification than others, but also that different predictors can be used in different parts of the stratification patterns to determine the subsets. An example of a prediction rule based on a clinical prediction rule in presented in the following figure.

Figure. Recursive Partitioning algorithm for prediction the incidence of myocardial infarction for patients who present to an emergency department with a chief complaint of chest pain. (Goldman L, et al. *A compute protocol to predict myocardial infarction in emergency department patients with chest pain*. *N Engl J Med* 1988;318:797-803)



The classification tree presented in this figure involves 11 different binary predictors. Cross-stratification by all 11 binary factors would have resulted in 2,048 strata, rather than the 14 that are displayed in that figure. Therefore, Recursive Partitioning retains much of the simplicity of stratification while avoiding the problems of having too many strata from cross-stratification by all predictors.

The performance of the partitioning tree displayed in the following table from reference (Goldman L, et al. *A compute protocol to predict myocardial infarction in emergency department patients with chest pain*. ***N Engl J Med*** 1988;318:797-803).

Table 1. Myocardial Infarctions as Predicted Retrospectively (1379 Patients) and Prospectively (4770 Patients) by the Computer Protocol.

Patient Subgroup*	Retrospective Group	Prospective Groups						
		University Hospitals		Community Hospitals				Total
		1	2	1	2	3	4	
		number of infarctions/total number of patients (percent)						
A	1/259 (0.4)	5/290	4/288	11/380	2/145	0/66	2/49	24/1218 (2)
B†	2/22 (9)	9/27	7/43	10/45	9/18	3/10	1/7	39/150 (26)
C	0/40	0/31	1/35	2/29	0/16	0/9	0/4	3/124 (2)
D	0/25	0/32	1/25	0/13	0/3	0/2	0/4	1/79 (1)
E	0/17	0/11	2/19	1/14	1/10	1/5	0/6	5/65 (8)
F	1/20 (5)	0/14	0/15	1/21	0/2	0/4	0/1	1/57 (2)
G†	30/91 (33)	12/69	4/55	24/104	3/19	4/24	4/23	51/294 (17)
H	0/96	5/100	3/99	5/132	1/30	1/26	1/17	16/404 (4)
I	3/57 (5)	0/39	0/49	0/68	2/20	0/11	0/7	2/194 (1)
J†	17/118 (14)	14/86	5/63	7/80	3/27	2/35	3/13	34/304 (11)
K†	38/152 (25)	27/119	7/61	27/96	10/20	11/25	7/22	89/343 (26)
L	1/234 (0.4)	2/338	0/327	8/201	4/101	3/45	0/27	17/1039 (2)
M†	8/50 (16)	12/50	3/22	8/40	4/12	4/21	1/4	32/149 (21)
N†	158/198 (80)	60/82	20/36	102/132	34/38	32/46	14/16	262/350 (75)
	259/1379 (19)	146/1288 (11)	57/1137 (5)	206/1355 (15)	73/461 (16)	61/329 (19)	33/200 (17)	576/4770 (12)

Each row of this table refers to a subgroup in the partitioning tree in the previous figure. The first column of table reports the cumulative incidence of MI (estimated risk) in each of the subgroups in the data set that created the partitioning tree (**Training Set** or **Derivation Set**). The last column reports the cumulative incidence for each subgroup using a different data set (**Testing Set** or **Validation Set**) that were not used in the construction of the partition tree.

In general, there is high agreement between risk estimates from the two data sets. However, the reported risk from the training set for a low- group (e.g. subgroup A: 0.4%) underestimates what is reported for that subgroup in the testing set (2%). Similarly, the reported risk in the training set for a high-risk subgroup (e.g. subgroup N: 80%) overestimates what is reported for that subgroup in the testing sets (75%). This reflects the potential problem (**overtraining**) of using a single data set to create a prediction rule and then using that same data to describe the performance of the rule in future patients.

The estimates in the first column of the previous table 1 are **optimistic** in that they underestimate the risk for low-risk subgroups and overestimate the risk for high-risk subgroups. When the training set is used to find the "**best"** fitting prediction rule for that training set data set, then chances are that this rule will not perform as well in a testing set. Therefore is it important to have a more valid way to estimate the performance of a rule on future patients, and using an independent testing set if often the desired approach.

When a testing set is not available, then re-sampling methods such as **bootstrapping** and **cross-validation** (described later in these notes) provide alternative methods for assessing the optimism of a prediction rule.

The most common method for developing clinical prediction is a **Regression Model**, describing the risk (P) of being in the outcome category of interest as a function of the predictors. For example, the Framingham Risk Model is a prediction rule to estimate the 10-Year Risk for developing Coronary Heart Disease based on six risk factors: age, sex, diabetes, smoking, cholesterol and blood pressure. The model was based on a Cox Regression Model that was fit to 2489 men and 2856 women from the original Framingham Cohort Study and from the Framingham Offspring Study. The model incorporated categories of total cholesterol, LDL cholesterol and HDL cholesterol as defined by the National Cholesterol Education Program (NCEP). Blood pressure was also represented by categories by the Joint National Committee (JNC-V). (Wilson PWF, et al. *Prediction of coronary heart disease using risk factor categories.* Circulation 1998:97:1837-47). The estimated risk of a subject is determined by substituting that subject's values for the risk factors into the regression model.

An example of a prediction rule based on a logistic regression model is the Mortality Prediction Model (MPM). This rule predicts the risk of in-hospital death for patients admitted to an intensive care unit (Lemeshow S, et al. *Predicting the Outcome of Intensive Care Unit Patients.* J Am Stat Assoc 1988;83(402):348-356.). Panel A of the following table displays the seven predictors that are used in this rule. As is typically the case, these predictors were chosen from a larger pool of potential predictors. Panel B of this table displays the formula for the Logistic Regression Model that defines this prediction rule. Panel C demonstrates the calculation of the estimated risk of death from this model for a subject with specified values for the predictors in the model.

Table Mortality Prediction Model (MPM) for Predicting In-Hospital Mortality among Patients admitted to an Intensive Care Unit.

A. Predictors

CONS	Level of Consciousness (1 if coma or deep stupor, 0 otherwise)
TYPE	Type of Admission (1 if emergent, 0 if elective)
CANCER	Cancer as Part of Present Problem (1 if yes, 0 if no)
CPR	Prior CPR (1 if yes, 0 if no)
INFECT	Infection (1 if probable, 0 otherwise)
AGE	Age in Years
SBP	Systolic Blood Pressure
SBP2	SBP squared.

B. Prediction Rule

$$\log(P/(1-P)) = -1.370 + 2.44(CONS) + 1.81(TYPE) + 1.49(CANCER) + .974(CPR) + .965(INFECT) + .0368(AGE) -.0606(SBP) + .000175(SBP2)$$

C. Sample Estimated Risk Calculation

CONS	= 1 (patient has coma or deep stupor)
TYPE	= 1 (emergent admission)
CANCER	= 0 (cancer part of present problem
CPR	= 0 (no prior CPR)
INFECT	= 0 (no probable infection)
AGE	= 50 (50 years of age)
SBP	= 150 (systolic blood pressure = 150)
SBP2	= 22500 (SBP squared)

$$\text{Log}(P/(1-P)) = -.1370 + 2.44 + 1.81 + .0368(50) - .0606(150) + .000175(22500) = .8005$$

$$P/(1-P) = \exp(.8005) = 2.227$$

$$P = 2.227/3.227 = 0.69$$

Panel C demonstrates the amount of calculations needed to obtain risk estimates from a logistic regression model. Although easy to program on a computer, the amount of calculations may limit the use of such a model in actual practice. However, when the predictors are binary in scale and the regression coefficients are proportional in scale, the predictive information in a regression model can be approximated by a simpler **Point Scoring System.** A scoring system assigns a weight (number of points) to each predictor. The weights are proportional to the regression coefficient of the predictors. Ranges of the sum of weights define risk categories.

As an example of a scoring system, Panel A of the following table displays the predictors that were chosen for a logistic regression model to predict the probability of bacteremia in blood samples of 1007 hospitalized patients (Bates DW, et al. *Predicting bacteremia in hospitalized patients*. Ann Intern Med 1990;113:495-500). Panel B displays the logistic regression model that defines the prediction rule. The regression coefficients in the logistic regression model show a pattern of similarity. For example, the coefficients for TEMP, CHILLS, POSEXAM, and COMORB are similar in value. Dividing the coefficients in the model by .32 and rounding these values to the nearest integer results in the scoring system displayed in Panel C of the table.

Table Scoring System to Predict the Risk of Bacteremia among Blood Tests.

A. Predictors

TEMP	Indicator variable for having a maximum temperature $\geq$ 38.3 C (1 = yes, 0 = no)
DCLASS2	Indicator variable specifying that a subject's diagnosis falls in a predefined category of rapidly fatal diseases (1 = yes, 0 = no)
DCLASS3	Indicator variable specifying that a subject's diagnosis falls in a predefined category of ultimately fatal diseases (1 = yes, 0 = no)
CHILLS	Indicator variable for the presence of chills (1 = yes, 0 = no)
DRUG	Indicator variable for intravenous drug abuse (1 = yes, 0 = no)
POSEXAM	Indicator variable for a positive focal abdomen examination (1 = yes, 0 = no)
COMORB	Indicator variable for a having one of a specified set of comorbid conditions (1 = yes, 0 = no)

B. Prediction Rule

$$\log(P/(1-P)) = -4.14 + .91*TEMP + 1.40*DCLASS1 + .65*DCLASS2 + .96*CHILLS + .287*DRUG + 1.03*POSEXAM + .96*COMORB$$

C. Scoring System

$$\text{Number of Points} = 3*TEMP + 4*DCLASS1 + 2*DCLASS2 + 3*CHILLS + 4*DRUG + 3*POSEXAM + 3*COMORB$$

The intercept term in the original regression model (Panel B of the previous table) is not included in the scoring system, since this would only add a constant to the total number of points calculated for any subject. The scoring system was used to develop a classification rule (shown the following table) based on ranges of points.

Table: Classification rule based on an Integer-Based Scoring System for predicting the presence of bacteremia among hospitalized patients.

	Risk Score			
	0-2 Points	3 Points	4-5 Points	$\geq$ 6 Points
Training Set (n=1007)	4/303= 0.01	11/236 = 0.05	18/204 = 0.09	41/264=0.16
Testing Set (n=509)	3/155=0.02	8/121=0.07	9/88=0.10	21/145=0.14

An examination of the results for the Training Set in this table shows that the risk of bacteremia ranges from 1% (4/303) in patients with 0-2 points to 16% (41/264) in patients with ≥ 6 points. However, as mentioned previously the performance of the prediction rule in these patients may be optimistic and over-estimate the performance in future patients. This is seen measuring the performance of the rule in a Testing Set of 509 different patients. The estimated risk of patients with 0-2 points is low (3/155 = 2%), but not as low as in the Training Set. Similarly, the estimated risk of patients with ≥ points is high (21/145 = 14%), but not as high as in the Training Set.

Although the integer-based scoring system in Panel C of the previous table may be easier to use than the original logistic regression model in Panel B, the gain in simplicity has its price. First, the regression coefficients from Panel B undergo some degree of rounding to create the weights of the scoring system, thereby losing some of the predictive information incorporated in the original model. In addition, because the weights in Panel C only reflect the relative importance of the predictors, they can no longer be used to generate estimated risks for the outcome. Categories defined by low number of points have lower risk of developing the outcome. However, the estimated risk for any category must be based on the cumulative incidence of the outcome among the subjects in that category .

Evaluation of a Clinical Prediction Rule

Performance Measures

The validity of a prediction rule is often quantified by various performance measured by how well its estimates agree with the actual outcomes of subjects. Measures of **discrimination** refer to the ability of the prediction rule to separate subjects with different outcomes into categories according to their values for the prediction rule. Measures of **calibration** refer to the ability of the model's estimated risk to agree with actual outcomes within groups of subjects.

Measures of Discrimination

One common method for assessing a prediction rule's ability to discriminate between the categories of a binary outcome is to determine the distribution of the outcome categories when subjects are ranked by their estimated risks. Ideally, cases of the outcome will tend to have high ranks of estimated risk, while non-cases will ten to have low ranks. The most common way to display the separation of outcome positive subjects from outcome negative subjects in this ranking is to calculate the ROC curve (Receiver Operating Characteristic). This curve is created by defining a classification rule based on a threshold of estimated risk. Subjects above that threshold are classified as

"high risk" subjects and subjects below that threshold are classified as "low risk" subjects. A **classification table** can then be displayed comparing the categories of the classification rule to those of the actual outcome. The following table displays the format of a confusion matrix along with the formulas for the **sensitivity** and **specificity** of the classification.

Table: 2x2 Table Displaying the Validity of a Binary Classification Rule.

	Outcome +	Outcome -
High-Risk	A	B
Low-Risk	C	D
Total	A+C	B+D

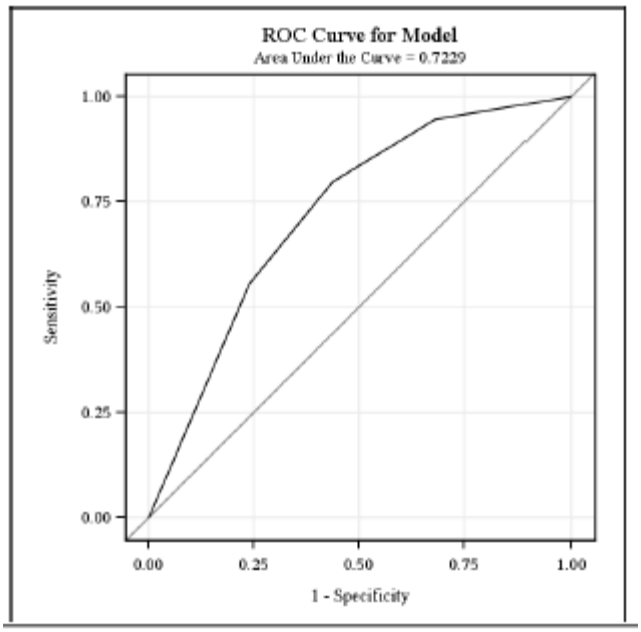
Sensitivity = A/(A+C)
Specificity = D/(B+D)

Varying the threshold for defining high risk generates a series of tables like that displayed in the previous table, with corresponding values for sensitivity and specificity. The ROC curve depicts the overall relationship between the prediction rule and the outcome by graphing the value for the sensitivity from each classification table on the vertical axis and for (1-specificity) on the horizontal axis. An ideal curve is one with points in the top left-hand region of the graph, reflecting a classification rule with high sensitivity and high specificity. Since an axis of the ROC curve ranges from 0.0 to 1.0, the total area in a box bounded by these axes is 1.0. Therefore, the ideal curve is one whose area under the curve (AUC) is close to 1.0. The following table shows the classification tables for the scoring system presented in a previous table for the bacteremia data according to three different thresholds of high risk.

Table: Classification tables showing the risk of bacteremia according to different definitions of high risk for the data

	Bacteremia		Sensitivity	Specificity
	+	-		
High Risk (3 + points)	70	634		
Low Risk (0-2 points)	4	299		
Total	74	933	70/74=0.95	299/933=0.32
High Risk (4 + points)	59	409		
Low Risk (0-3 points)	15	524		
Total	74	933	59/74=0.80	524/933=0.56
High Risk (6 + points)	41	264		
Low Risk (0-5 points)	33	669		
Total	74	933	41/74=0.55	669/933=0.72

The following figure shows the ROC Curve and it corresponding area for the 1009 subjects in the training set for the bacteremia data set.



The area under the ROC curve (AUC) is 0.72. However, the ROC curve for the testing set in the bacteremia data has an AUC 0.69. The difference in area (0.72 – 0.69 = 0.03) estimates the amount of optimism that is obtained in the value from the AUC based on the model's performance in the training data set.

Measures of Calibration

Calibration refers to the degree of agreement between a subject's estimated outcome from a prediction rule and the subject's actual outcome. Measures of the degree of calibration commonly take on the form of "observed versus expected" comparisons of the outcome. A common approach is to assign subjects to risk categories according to sub-ranges of predicted risk. Within each risk category, the average predicted risk is compared to the observed cumulative incidence of the outcome. Alternatively, the sum of the predicted risks in a category provides an estimate of the expected number of outcomes for that category (E), which can be compared to the actual number of outcomes for that category (O).

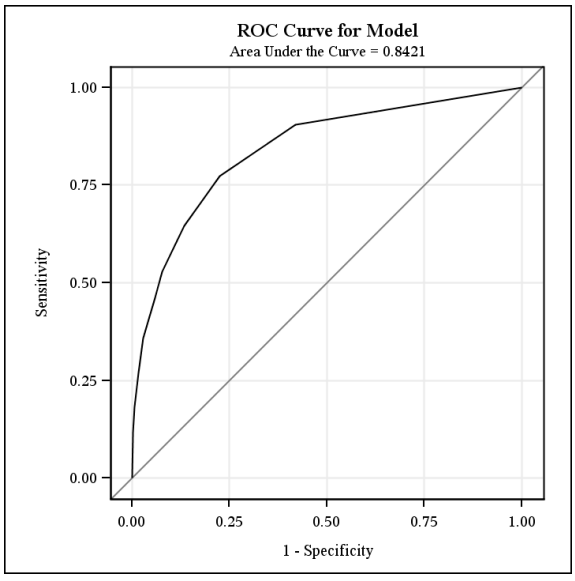
A common summary measure of calibration for binary outcomes is the Hosmer and Lemeshow goodness-of-fit statistic (Hosmer DW, Lemeshow S. *Applied Logistic Regression Second Edition.* John Wiley & Sons; New York: 2000). An example of a calibration table is given in the following table, which shows the performance of the MPM prediction algorithm described in a previous table. This algorithm was developed on an initial data set of 775 admissions to an intensive care unit. The results listed in the following tables describe the performance of this algorithm in another set of 1997 admissions to the intensive care unit.

Table :. Calibration of the Mortality Prediction Model (MPM) in 1997 patients admitted to an intensive care unit.

Range of MPM P(Dying)	Number of Subjects (N)	Deaths Observed	Deaths Expected
.00 - .09	967	38	41.9
.10 - .19	365	52	52.2
.20 - .29	194	50	48.2
.30 - .39	139	47	48.8
.40 - .49	88	41	39.0
.50 - .59	56	26	30.0
.60 - .69	56	35	35.9
.70 - .79	48	35	36.0
.80 - .89	35	26	29.5
.90 - 1.0	49	46	46.7
Total	1997	396	408.2

$X^2_{HL} = 4.94$, $p = .90$

In general, the previous table shows good agreement between expected and observed outcomes for each risk category, suggesting good calibration of risk estimates. Furthermore, the following ROC curve and its corresponding AUC show good discriminating ability of the risk estimates.



Determining the Predictors to Include in a Prediction Rule

Often the pool of potential predictors is large and the choice is to include all or a representative subject of the predictors in a model. The optimal number of predictors to include in a rule should be guided by the amount of information that is contained in the data set. For a binary outcome, a very rough rule of thumb for model stability is to require a minimum number of subjects in each outcome category for every predictor considered for the analysis. The suggested minimum number of subjects has ranged from 5 to 10 (Wasson JH, et al. *Clinical prediction rule: application and methodologic standards.* N Engl J Med 1985;313:793-799.). For continuous outcomes, the suggested rule of thumb is to require 10 subjects for every candidate predictor (Harrell FE, et al. *Regression modeling strategies for improved prognostic prediction*. ***Stat in Med*** 1984).

When the number of potential predictors exceeds the limit suggested by these guidelines then a model based on all predictors is not only unstable and complex but is also likely to be optimistic and not generalize to other data sets. One solution to this problem is to select only a subset of the predictors for the model based on associations found in the data.

Parsimony

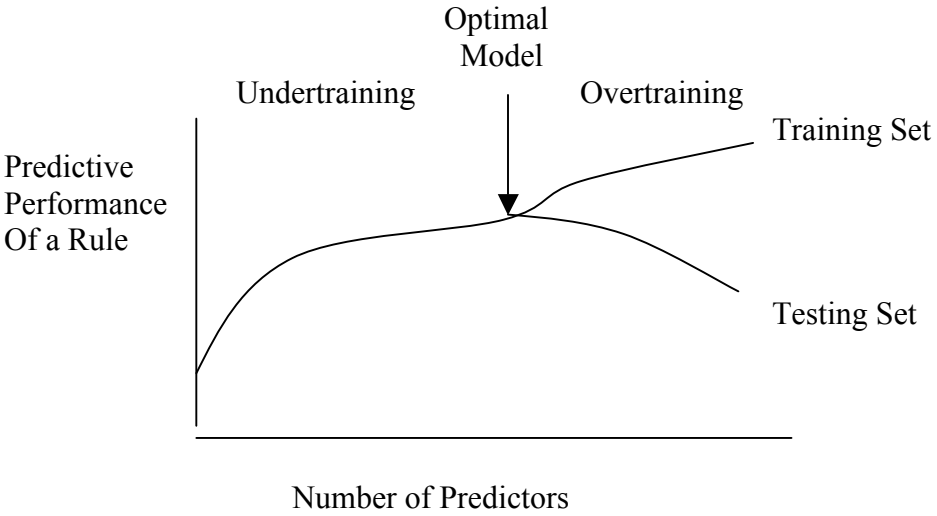
Parsimony pertains to the number of predictors to include in a prediction rule. **Occam's Razor (**William of Occam) states that "*Pluralitas non est ponenda sine neccesitate*" (plurality should not be posited without necessity). **Albert Einstein** stated a similar theme when saying “Everything should be made as simple as possible, but no simpler”.

For model building, parsimony implies including only those factors that are true predictors, not only in the data on which the model is created (training set) but also on other data sets on which it is to be applied (testing sets). This has direct relevance to models that are created by variable selection algorithms. A **forward selection algorithm** builds a model in steps, each time adding a factor that adds the more statistically significant, incremental predictive information to the model. **A backwards elimination algorithm** builds a model by starting with a large model that contains all of the factors and then reduces the model in steps, each time eliminating the least statistically significant factor.

Although commonly used for developing prediction rules, variable selection algorithms possess serious potential problems. Since they involve a large number of tests of significance, these methods increase the likelihood for overtraining through multiple testing. Although a single non-predictor has a 5% chance of demonstrating a statistically significant relationship to the outcome in a dataset, 10 independent non-predictors have a 40% chance of at least one of them reaching statistical significance. In the extreme, 100 independent non-predictors have a 99% chance of one reaching statistical significance.

Parsimony suggests that the model building process should cease at the point when selected factors would not be true predictors in other data sets. This is demonstrated by following figure showing the expected pattern of performance of series models created by a forward selection algorithm in training data set and evaluated in a testing data set. Models created in the early stage of the selection process tend to be based on predictors whose strong associations tend to generalize to other data sets. Therefore, the performance of these models in the training set also reflects the expected performance in other data sets. However, because of the search to identify factors related to the outcome in the training set, a point is often reached where factors are selected based on associations that are particular to the training set and are not repeated in a testing set. Including such factors in the model results in worse performance in a testing set. This process is often labeled as **overtraining** or **overfitting.** The challenge when using variable selection algorithms is to determine the optimal stopping point to avoid both problems.

Figure Expected change in predictive accuracy as the number of predictors in a prediction algorithm increases.



Validation

Validation of a clinical prediction rule involves obtaining "honest" estimates of the rule's performance in actual practice. Perhaps the simplest means for assessing the validity of a prediction rule is by examining its performance in an independent testing set. This is often implemented by randomly splitting a data set into a training data set and a testing data set. The prediction rule is developed in the training set and validated in the testing data set.

Alternatively, **cross-validation** divides the data set into a series of (usually disjoint but exhaustive) testing sets. For example, **10-fold cross-validation** divides the

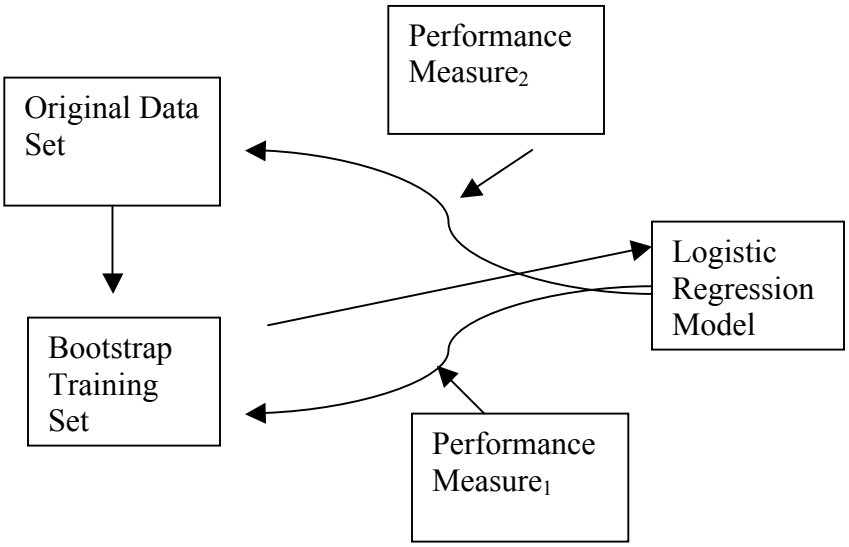
data set into 10 mutually exclusive testing sets with each subject in only one of these sets. A prediction rule is fit in the complement of each of these testing sets (90% of the data) and then evaluated in the corresponding testing set. Finally, the performances of the 10 prediction rules in the 10 testing sets are averaged and used as an estimate of the performance of the single prediction rule built on the entire data set. The following figure presents a graphical display of this method.

Figure: Ten-Fold Cross Validation.

Subset #1	**Subset #2**	**Subset #3**	**Subset #4**	**Subset #5**	**Subset #6**	**Subset #7**	**Subset #8**	**Subset #9**	**Subset #10**
Test Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1	Train Set # 1
Train Set # 2	**Test Set # 2**	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2	Train Set # 2
Train Set # 3	Train Set # 3	**Test Set # 3**	Train Set # 3	Train Set # 3	Train Set # 3	Train Set # 3	Train Set # 3	Train Set # 3	Train Set # 3
Train Set # 4	Train Set # 4	Train Set # 4	**Test Set # 4**	Train Set # 4	Train Set # 4	Train Set # 4	Train Set # 4	Train Set # 4	Train Set # 4
Train Set # 5	Train Set # 5	Train Set # 5	Train Set # 5	**Test Set # 5**	Train Set # 5	Train Set # 5	Train Set # 5	Train Set # 5	Train Set # 5
Train Set # 6	Train Set # 6	Train Set # 6	Train Set # 6	Train Set # 6	**Test Set # 6**	Train Set # 6	Train Set # 6	Train Set # 6	Train Set # 6
Train Set # 7	Train Set # 7	Train Set # 7	Train Set # 7	Train Set # 7	Train Set # 7	**Test Set # 7**	Train Set # 7	Train Set # 7	Train Set # 7
Train Set # 8	Train Set # 8	Train Set # 8	Train Set # 8	Train Set # 8	Train Set # 8	Train Set # 8	**Test Set # 8**	Train Set # 8	Train Set # 8
Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	Train Set # 9	**Test Set # 9**	Train Set # 9
Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	Train Set # 10	**Test Set # 10**

Bootstrapping is another method for estimating the validity of a prediction rule in the absence of an independent testing set. It involves re-sampling the original data set with replacement in order to obtain a new "bootstrap" training set of the same size as the original data set. A prediction rule is then developed on the bootstrap training set and a measure of its performance is calculated both on that data set and on the original data set, using the original data set as a testing set for the prediction rule developing on the bootstrap training set. The difference a performance measure on the two data set provides an estimate of the **optimism** of the performance of the prediction rule on the bootstrap training set. This process is then repeated multiple times and the average of the optimism values is used as an estimate of the optimism of a single prediction rule that is developed on the full data set. The following figure presents a graphical display of this method.

Figure Bootstrap estimate of the optimism of the Area of the ROC Curve (AUC) from a logistic regression model.



Optimism = $(\text{Performance Measure})_1 - (\text{Performance Measure})_2$