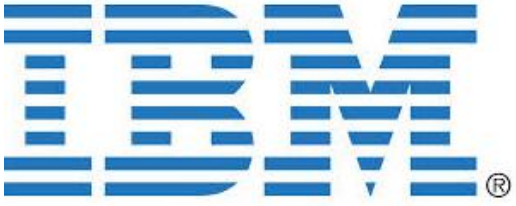


Big Data Analytics for Healthcare

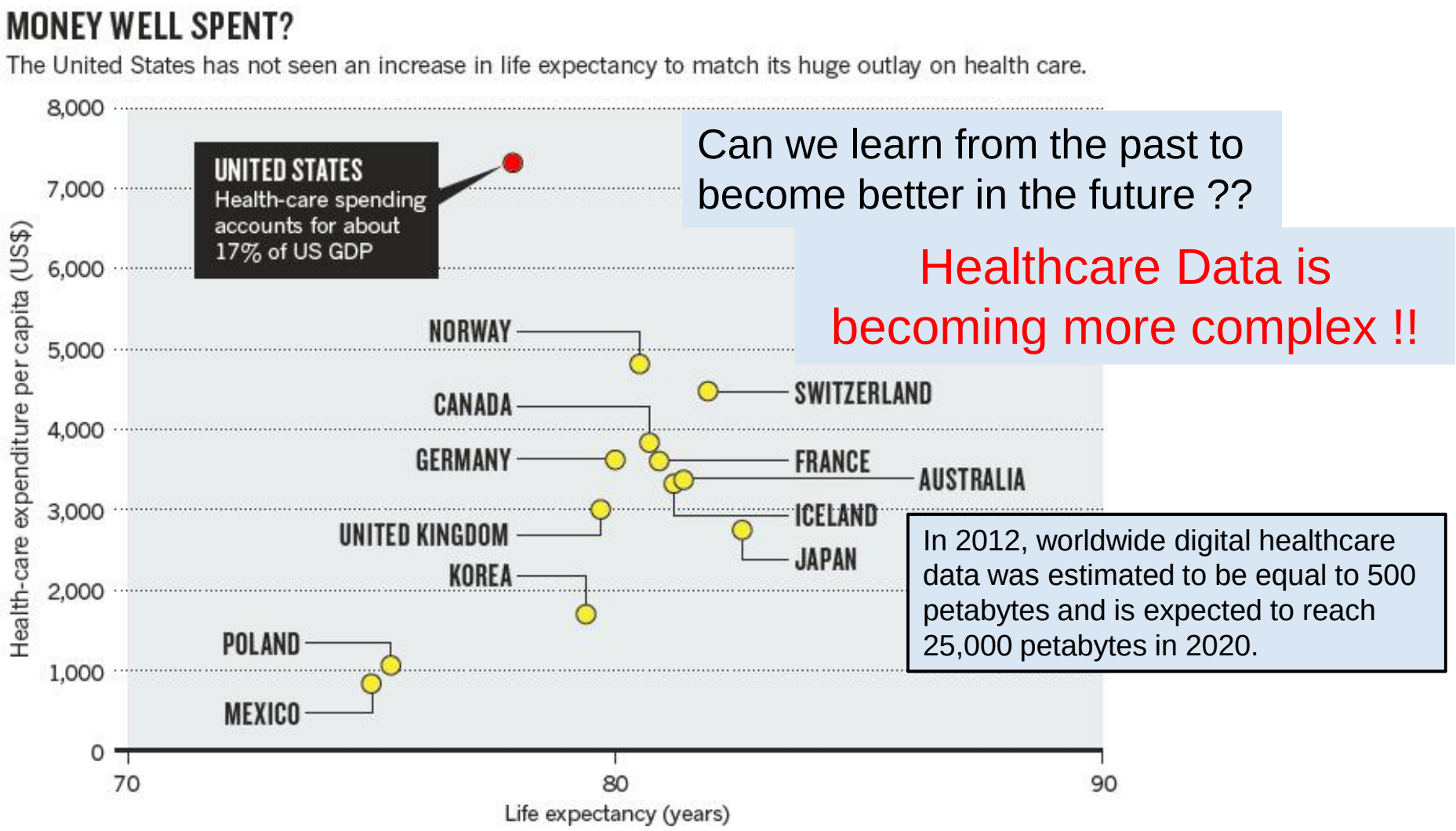
Jimeng Sun
Healthcare Analytics Department
IBM TJ Watson Research Center

Chandan K. Reddy
Department of Computer Science
Wayne State University

Tutorial presentation at SIGKDD 2013.

The updated tutorial slides are available at http://dmkd.cs.wayne.edu/TUTORIAL/Healthcare/

Motivation



Can we learn from the past to become better in the future ??

Healthcare Data is becoming more complex !!

In 2012, worldwide digital healthcare data was estimated to be equal to 500 petabytes and is expected to reach 25,000 petabytes in 2020.

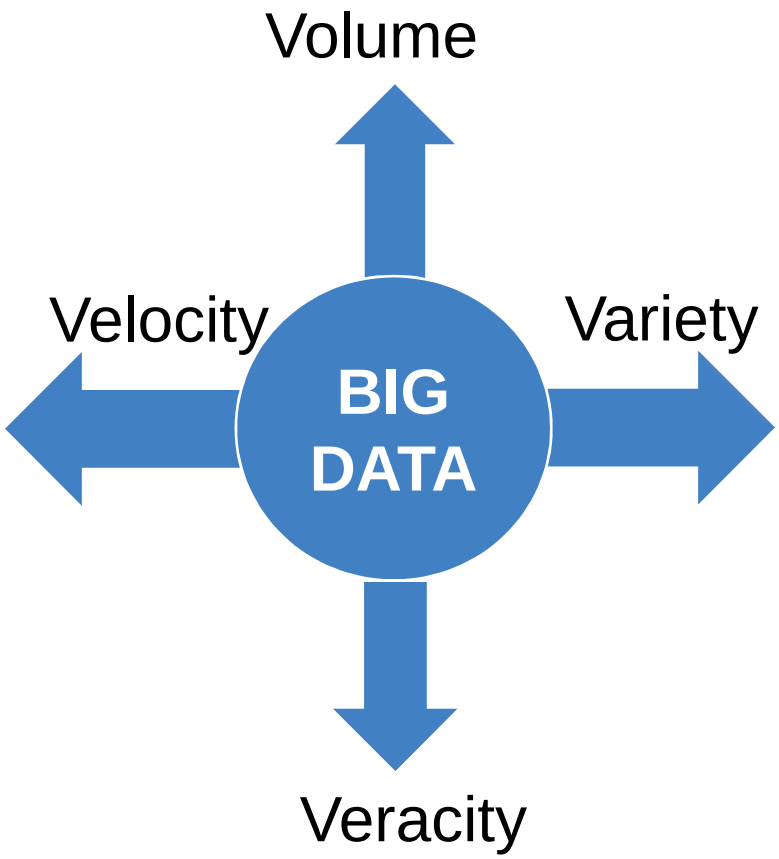
Hersh, W., Jacko, J. A., Greenes, R., Tan, J., Janies, D., Embi, P. J., & Payne, P. R. (2011). Health-care hit or miss? *Nature*, *470*(7334), 327.

Organization of this Tutorial

- **Introduction**
- **Clinical Predictive Modeling**
  - **Case Study: Readmission Prediction**
- **Scalable Healthcare Analytics Platform**
- **Genetic Data Analysis**
- **Conclusion**

What is Big Data

- Large and complex data sets which are difficult to process using traditional database technology.



The four dimensions (V's) of Big Data

Big data is not just about size.

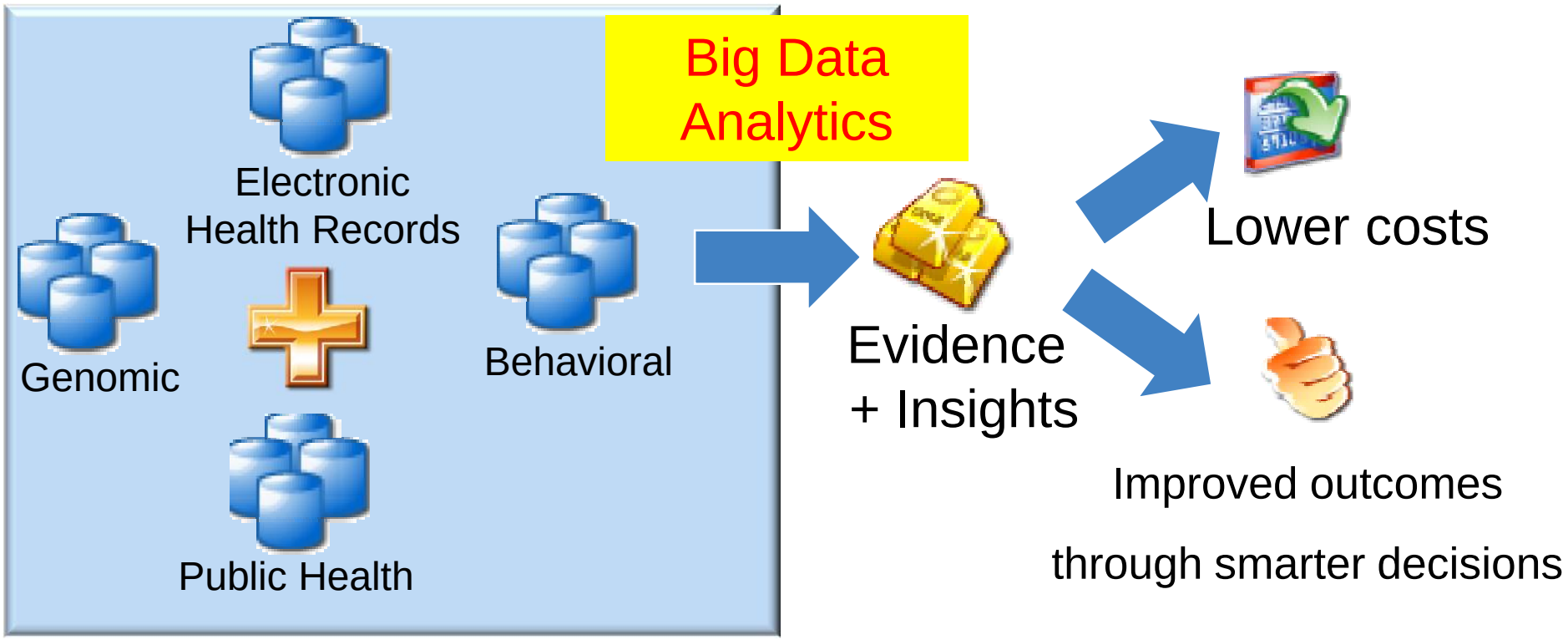
- Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data.
- It aims to answer questions that were previously unanswered.

Healthcare Analytics in the Electronic Era

- Old way: **Data are expensive and small**
  - Input data are from clinical trials, which is small and costly
  - Modeling effort is small since the data is limited

- EHR era: **Data are cheap and large**
  - Broader patient population
  - Noisy data
  - Heterogeneous data
  - Diverse scale
  - Longitudinal records

Overall Goals of Big Data Analytics in Healthcare



GOAL: Provide Personalized care through right intervention to the right patient at the right time.

Examples for Big Data Analytics in Healthcare

Government Initiatives

- **Medicare Penalties:** Medicare penalizes hospitals that have high rates of readmissions among patients with Heart failure, Heart attack, Pneumonia.

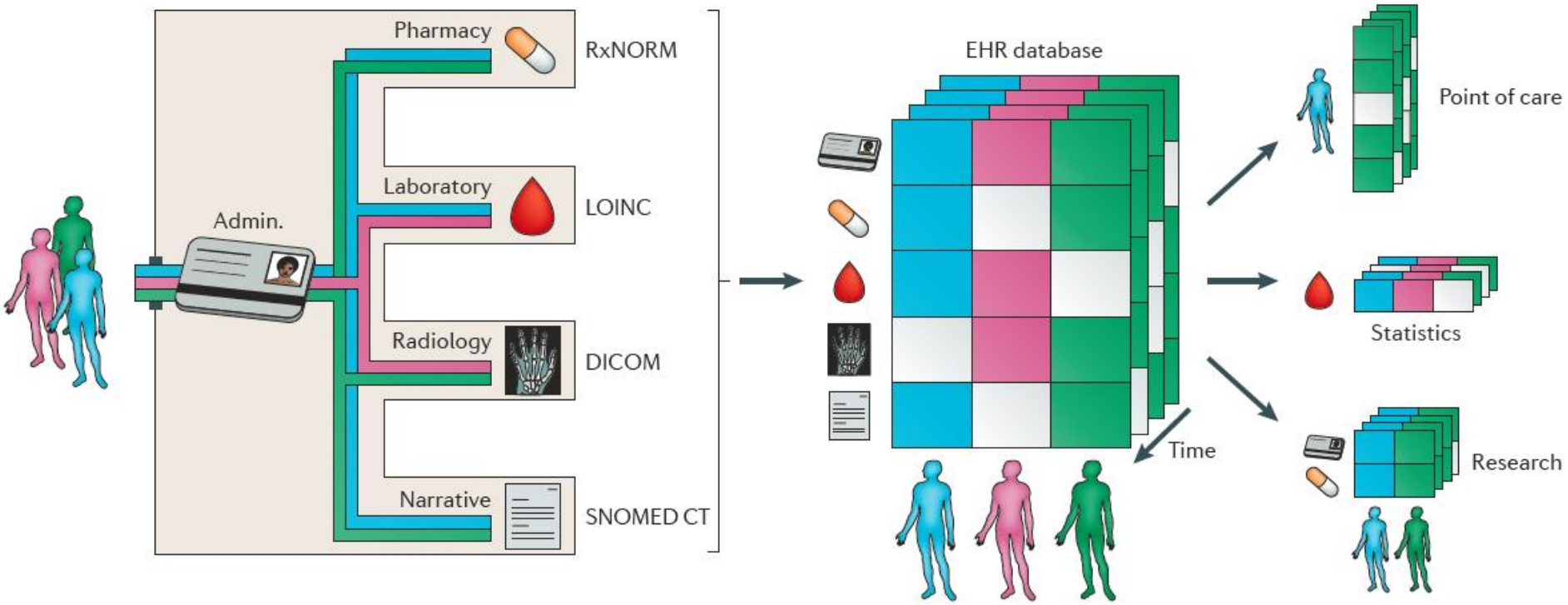
- **BRAIN Initiative:** Find new ways to treat, cure, and even prevent brain disorders, such as Alzheimer's disease, epilepsy, and traumatic brain injury. A new bold $100 million research initiative designed to revolutionize our understanding of the human brain.

Industry Initiatives

- **Heritage Health Prize:** Develop algorithms to predict the number of days a patient will spend in a hospital in the next year. http://www.heritagehealthprize.com

- **GE Head Health Challenge:** Methods for Diagnosis and Prognosis of Mild Traumatic Brain Injuries. Develop Algorithms and Analytical Tools, and Biomarkers and other technologies. A total of $60M in awards.

Data Collection and Analysis



Effectively integrating and efficiently analyzing various forms of healthcare data over a period of time can answer many of the impending healthcare problems.

Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." *Nature Reviews Genetics* (2012).

PREDICTION MODELS FOR CLINICAL DATA ANALYSIS

Different Kinds of Outcomes

- Economic Cost → **Continuous Outcomes**
- Diagnostic
  - Presence of a disease → **Binary Outcomes**
  - What disease → **Categorical Outcomes**
  - How serious → **Ordinal Outcomes**
- Survival analysis → **Survival Outcomes**
- ⋮

Examples for Different Outcomes in Healthcare

- **Binary Outcomes**
  - Death: yes/no
  - Adverse event: yes/no
- **Continuous Outcomes**
  - Days of Hospital stay
  - Visual analogue score
- **Ordinal Outcomes**
  - Quality of life scale
  - Grade of tumour progression
  - Count data (Number of heart attacks)
- **Survival Outcomes**
  - Cancer survival
  - Clinical Trials

Continuous Outcomes

- **Linear Regression**

The Outcome y is assumed to be the linear combination of the x_k variables with the estimated regression coefficients β_k.

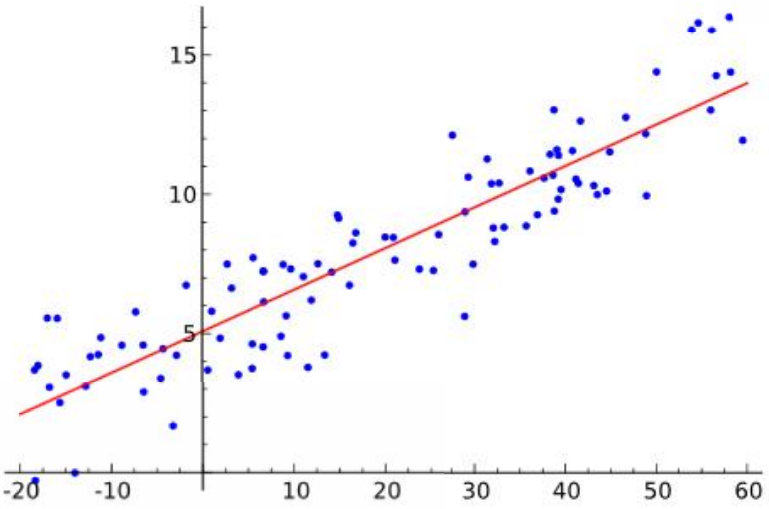
$$y = \beta_0 + \sum_{k=1}^{l} \beta_k x_k$$

the coefficients β_k usually estimated by minimize the *RSS* ("residual sum of squares"). Various penalized *RSS* methods are used to shrink the β_k, and achieve a more stabilized model.

$$RSS = (Y - X\vec{B})^T (Y - X\vec{B})$$

The minimum of the sum of squares is found by setting the gradient to zero.

$$\vec{B} = (X^T X)^{-1} (X^T Y)$$



- **Generalized Additive Model (GAM)**

$$y = b_0 + f_k(x_k) + error$$

Where b_0 refers to the intercept, f_k refers to functions for each predictor which is more flexible.

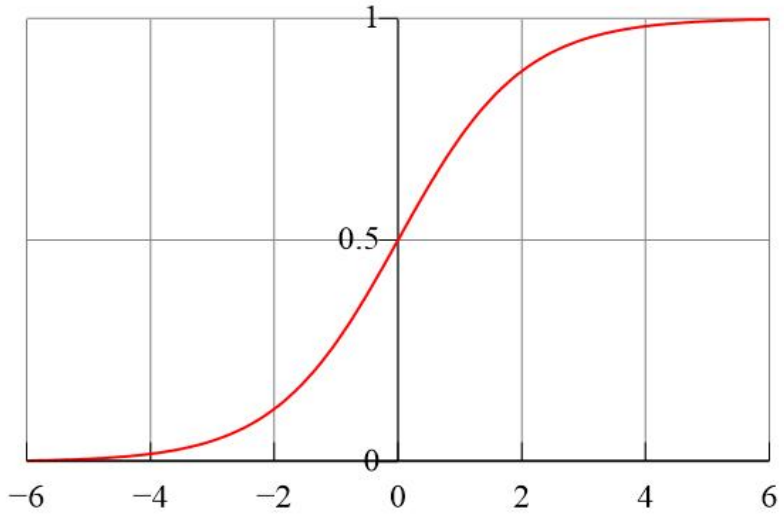
Binary Outcomes

- **Logistic Regression**

The model is stated in terms of the probability that y=1, rather than the outcomes *Y* directly, it can be viewed as a linear function in the logistic transformation:

$$\Pr(y = 1|x) = \frac{1}{1 + e^{-z}}$$

where $z = \sum_{k=0}^{l} \beta_k x_k$, the coefficients β_k usually estimated by maximum likelihood in a standard logistic regression approach.



Binary Outcomes

- **Bayes Modeling**

The predict is given based on the Bayes theorem:

$$P(D|X) = \frac{P(X|D) \cdot P(D)}{P(X)}$$

P(X) is always constant, and the prior probability P(D) can be easily calculated from the training set. Two ways to estimate the class-conditional probabilities $P(X|D)$: naïve Bayes and Bayesian belief network.

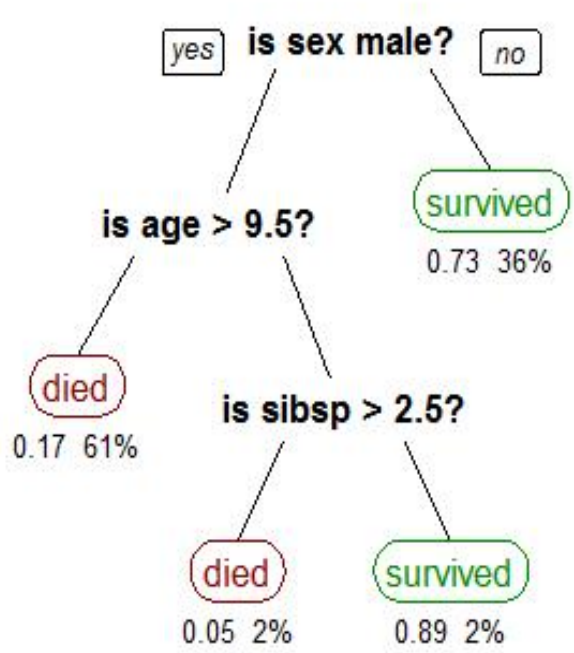
naïve Bayes: assuming the attributes are conditionally independent, so

$$P(X|D) = \prod_{i=1}^{d} P(x_i|D)$$

Binary Outcomes

- **Classification and Regression Trees**

Recursively split the patients into smaller subgroups. Splits are based on cut-off levels of the predictors, which can maximize the difference between two subgroups, and minimize the variability within these subgroups.



Some other commonly used binary outcomes models:
Multivariate additive regression splines (MARS) models
Support Vector Machine (SVM)
Neural Nets …

Categorical Outcomes

- **Polytomous Logistic Regression**

The model for j outcomes categories can be written as:

$$Logodds(y = j\ vs.\ y = reference) = \alpha_j + \sum_{k=1}^{l} \beta_{k,j} x_{k,j}$$

One vs all approach. Similar to multi-class prediction.

j-1 models are fitted and combined to the prediction.

Ordinal Outcomes

- **Proportional Odds Logistic Regression**

A common set of regression coefficients is assumed across all levels of the outcome, and intercepts are estimated for each level.

$$logit(y_j) = \alpha_j + \sum_{k=1}^{l} \beta_k x_k$$

Survival Outcomes

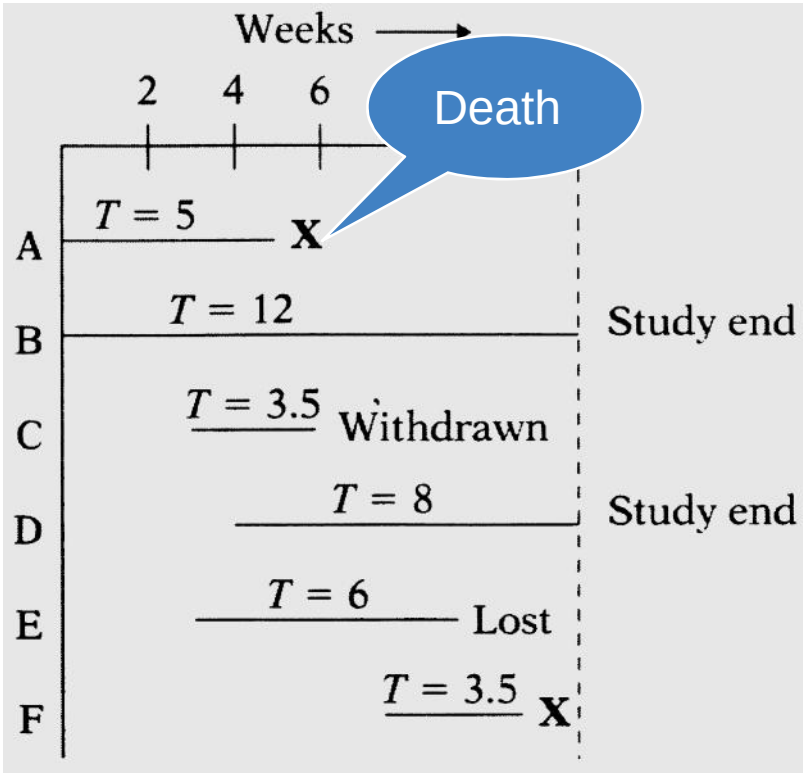
- **Survival Analysis**

Survival Analysis typically focuses on time to event data. Typically, survival data are not fully observed, but rather are censored. In survival analysis, the primary focus is on the survival function.

$$S(t) = \Pr(T \geq t)$$

where T is the Failure time or the time that a event happens. So the $S(t)$ means the probability that the instance can survive for longer than a certain time t. The censoring variable is the time of withdrawn, lost, or study end time.

The hazard function: the probability the "event" of interest occurs in the next instant, given survival to time t.



Right censored

Survival Outcomes

- Non-Parametric Approaches
  - Kaplan-Meier Analysis
  - Clinical Life Tables
  - Estimates the Survival Function

- Semi-Parametric Approaches
  - Cox Proportional Hazards Model
  - Estimates the Hazard Function

Survival Outcomes

- **Kaplan-Meier Analysis**

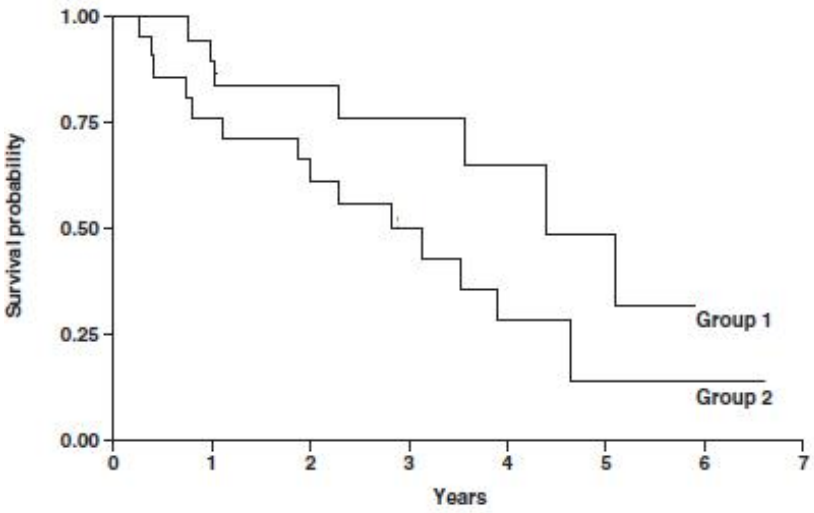
Kaplan-Meier analysis is a nonparametric approach to survival outcomes. The survival function is:

$$\hat{S}(t) = \prod_{j:T_j<t} (1 - \frac{d_j}{r_j})$$

where

- $T_i \ldots T_K$ is a set of distinct death times observed in the sample.
- d_j is the number of deaths at T_j.
- c_j is the number of censored observations between T_j and T_{j+1}.
- r_j is the number of individuals "at risk" right before the j^{th} death.

$$r_j = r_{j-1} - d_{j-1} - c_{j-1}$$



Efron, Bradley. "Logistic regression, survival analysis, and the Kaplan-Meier curve." Journal of the American Statistical Association 83.402 (1988): 414-425.

Survival Outcomes

Example

Patient	Days	Status	Patient	Days	Status	Patient	Days	Status
1	21	1	15	256	2	29	398	1
2	39	1	16	260	1	30	414	1
3	77	1	17	261	1	31	420	1
4	133	1	18	266	1	32	468	2
5	141	2	19	269	1	33	483	1
6	152	1	20	287	3	34	489	1
7	153	1	21	295	1	35	505	1
8	161	1	22	308	1	36	539	1
9	179	1	23	311	1	37	565	3
10	184	1	24	321	2	38	618	1
11	197	1	25	326	1	39	793	1
12	199	1	26	355	1	40	794	1
13	214	1	27	361	1			
14	228	1	28	374	1			

Status

1: Death

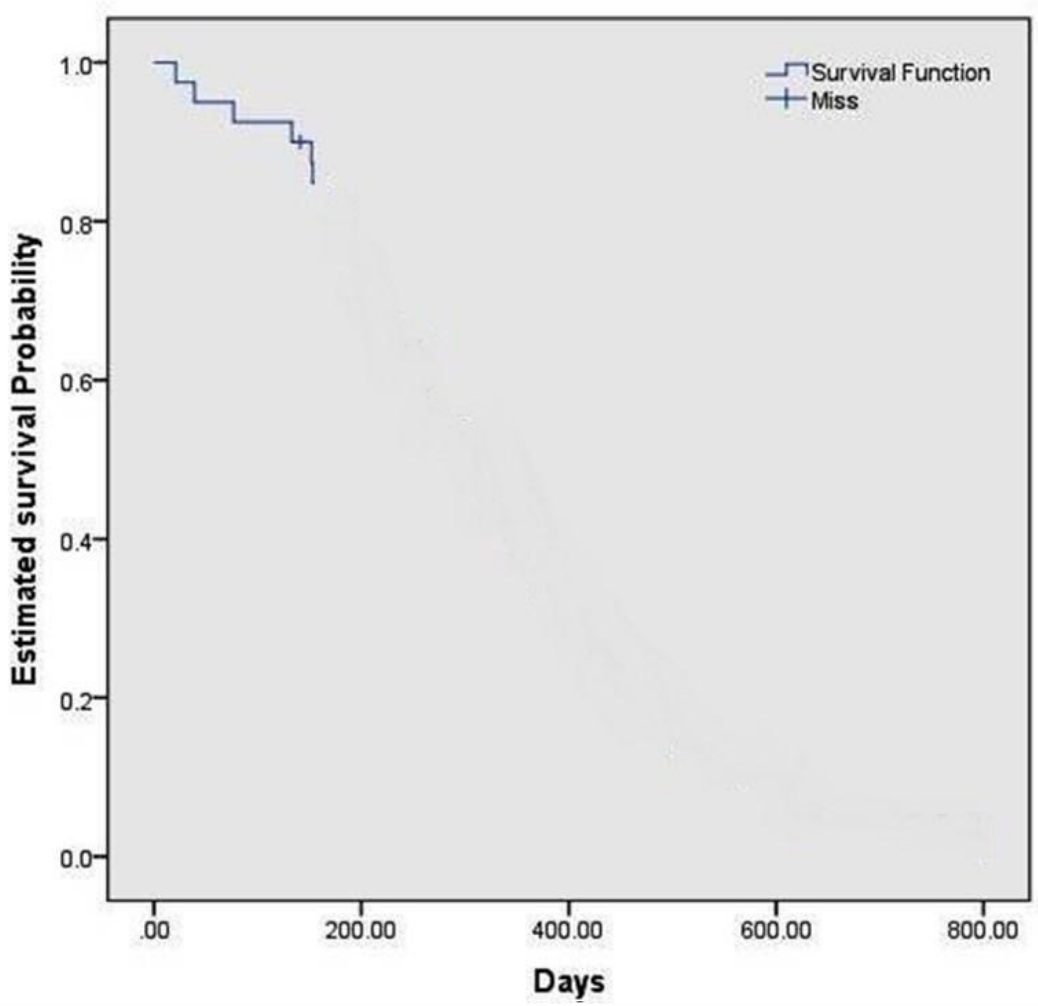
2: Lost to follow up

3: Withdrawn Alive

Kaplan-Meier Analysis

Kaplan-Meier Analysis

i	Time	Status	d_i	c_i	r_i	$S(t)$
1	21	1	1	0	40	0.975
2	39	1	1	0	39	0.95
3	77	1	1	0	38	0.925
4	133	1	1	0	37	0.9
5	141	2	0	1	36	.
6	152	1	1	0	35	0.874
7	153	1	1	0	34	0.849



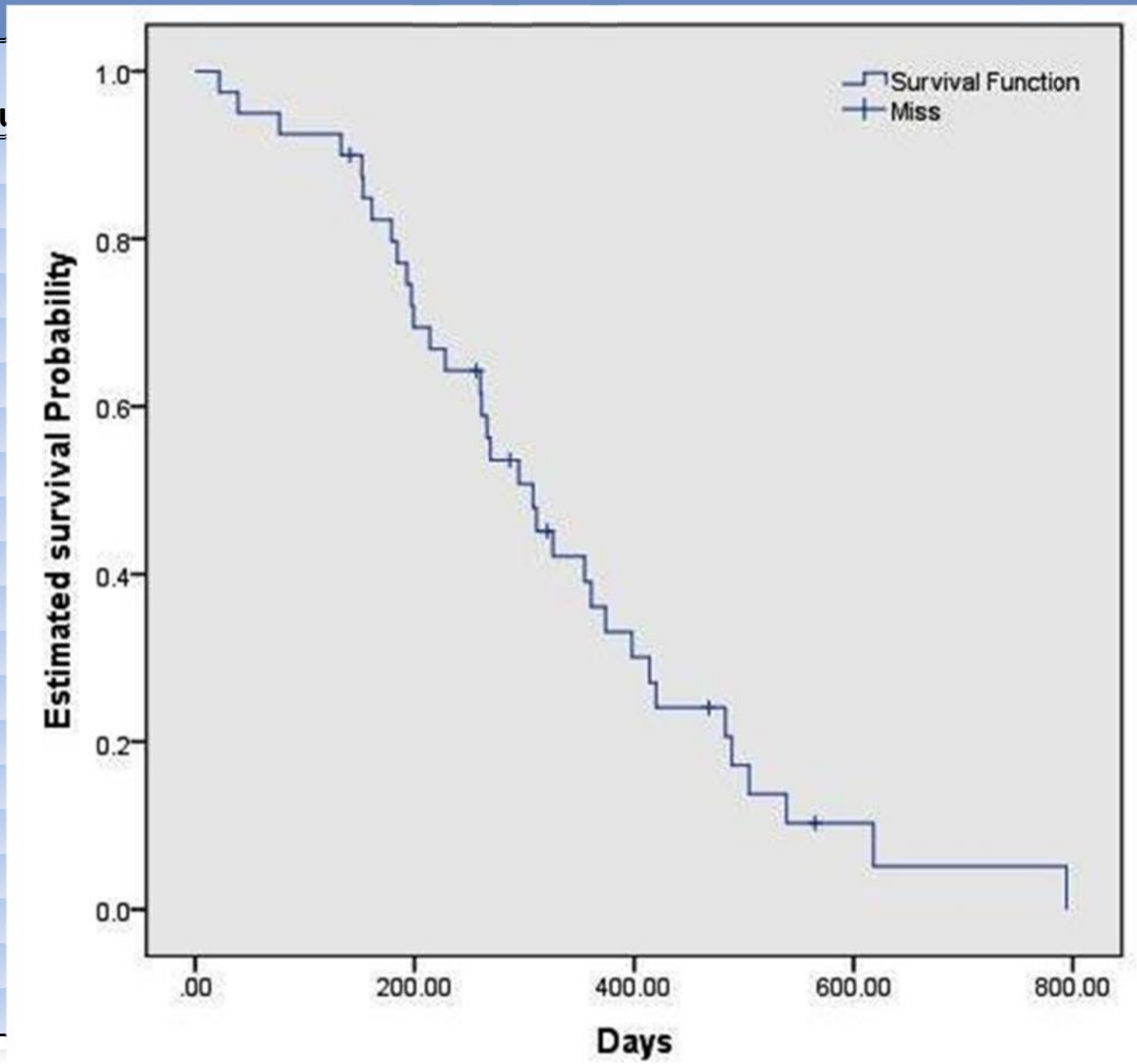
K_M Estimator:

$$\hat{S}(t) = \prod_{j:T_j<t} (1 - \frac{d_j}{r_j})$$

Kaplan-Meier Analysis

K_M Estimator:

i	Time	Statu	rror	$\sum d_i$	r_i
1	21	1		18	20
2	39	1	81	19	19
3	77	1	81	20	18
4	133	1	81	21	17
5	141	2		21	16
6	152	1	81	22	15
7	153	1	81	23	14
8	161	1	8	24	13
9	179	1	79	25	12
10	184	1	77	26	11
11	193	1	75	27	10
12	197	1	72	28	9
13	199	1		28	8
14	214	1	7	29	7
15	228	1	66	30	6
16	256	2	61	31	5
17	260	1	55	32	4
18	261	1		32	3
19	266	1	46	33	2
20	269	1		34	1



Clinical Life Tables

- Clinical life tables applies to grouped survival data from studies in patients with specific diseases, it focuses more on the conditional probability of dying within the interval.

the j^{th} time interval is $[t_{j-1}, t_j)$ **VS.**
$T_i \ldots T_M$ is a set of distinct death

The survival function is:

$$\hat{S}(t_j) = \prod_{\iota<j} (1 - \frac{d_\iota}{r_\iota'})$$

nonparametric

Assumption:

- at the beginning of each interval: $r_j' = r_j - c_j$
- at the end of each interval: $r_j' = r_j$
- on average halfway through the interval: $r_j' = r_j - c_j/2$

K_M analysis suits small data set with a more accurate analysis,
Clinical life table suit for large data set with a relatively approximate result.

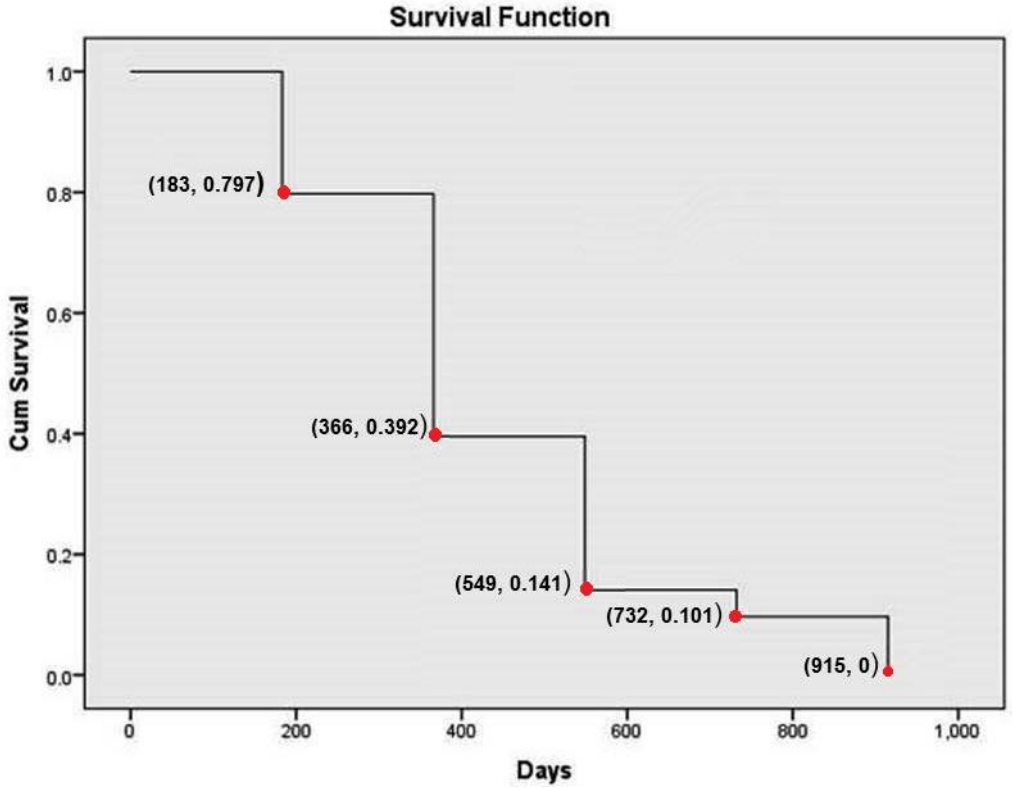
Cox, David R. "Regression models and life-tables", Journal of the Royal Statistical Society. Series B (Methodological) (1972): 187-220.

Clinical Life Tables

NOTE :
The length of interval is half year(183 days)

Clinical Life Table

Interval	Interval Start Time	Interval End Time	r_i	c_i	r_i'	d_i	$S(t)$	Std. Error of $S(t)$
1	0	182	40	1	39.5	8	0.797	0.06
2	183	365	31	3	29.5	15	0.392	0.08
3	366	548	13	1	12.5	8	0.141	0.06
4	549	731	4	1	3.5	1	0.101	0.05
5	732	915	2	0	2	2	0	0



Clinical Life Table :

$$\hat{S}(t_j) = \prod_{i<j} (1 - \frac{d_i}{r_i'})$$

On average halfway through the interval: $r_j' = r_j - c_j/2$

Survival Outcomes

- **Cox Proportional Hazards model**

This semi-parametric model is the most common model used for survival data, it can be written as:

$$\lambda_i(t|Z) = \lambda_0(t)e^{\beta Z} \quad \Rightarrow \quad \log\left(\frac{\lambda_i(t|Z)}{\lambda_0(t)}\right) = \sum_{k=1}^{p} \beta_k z_{k,i}$$

Where $\lambda_0(t)$ is the baseline hazard function which must be positive. The function shows that Cox Proportional Hazards model is a linear model for the log of the hazard ratio. Maximum Likelihood methods are used to estimate the perimeter, Cox (1972) uses the idea of a partial likelihood, to generalize it for censoring.

$$L_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta Z_i}}{\sum_{j \in R(X_i)} e^{\beta Z_j}}\right]^{\delta_i}$$

β are estimated without any assumptions about the form of $\lambda_0(t)$. This is what makes the model semi-parametric.

Cox, David R. "Regression models and life-tables", Journal of the Royal Statistical Society. Series B (Methodological) (1972): 187-220.

Survival Outcomes

COX (cont.)

$$L_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta Z_i}}{\sum_{j \in R(X_i)} e^{\beta Z_j}} \right]^{\delta_i} = \prod_{i=1}^{K} \frac{e^{\beta Z_i}}{\sum_{j \in R(T_i)} e^{\beta Z_j}}$$

$R(T_i)$ as the risk set at the ith failure time

log-partial likelihood:

$$l(\beta) = \log \left[L_p(\beta) \right] = \log \left[\prod_{i=1}^{K} \frac{e^{\beta Z_i}}{\sum_{j \in R(T_i)} e^{\beta Z_j}} \right] = \sum_{i=1}^{K} \left[\beta Z_i - \log \left[\sum_{j \in R(T_i)} e^{\beta Z_j} \right] \right]$$

The partial likelihood score equations are:

$$U(\beta) = \frac{\partial}{\partial \beta} l(\beta) = \sum_{i=1}^{n} \delta_i \left[Z_i - \frac{\sum_{j \in R(T_i)} Z_j e^{\beta Z_j}}{\sum_{j \in R(T_i)} e^{\beta Z_j}} \right]$$

The maximum partial likelihood estimators can be found by solving $U(\beta) = 0$.

Regularized Cox Regression

- Regularization functions for cox regression

$$\hat{\beta} = argmin\ l(\beta) + \mu * P(\beta)$$

$P(\beta)$ is a sparsity inducing norm such and μ is the regularization parameter.

Method	Penalty Term Formulation
LASSO	$\sum_{k=1}^{p} \lvert\beta_k\rvert$
Ridge	$\sum_{k=1}^{p} {\beta_k}^2$
Elastic Net (EN)	$\mu \sum_{k=1}^{p} \lvert\beta_k\rvert + (1-\mu) \sum_{k=1}^{p} {\beta_k}^2$
Adaptive LASSO (AL)	$\sum_{k=1}^{p} w_k \lvert\beta_k\rvert$
Adaptive Elastic Net (AEN)	$\mu \sum_{k=1}^{p} w_k \lvert\beta_k\rvert + (1-\mu) \sum_{k=1}^{p} {\beta_k}^2$

Promotes Sparsity

Handles Correlation

Sparsity + Correlation

Adaptive Variants are slightly more effective

Survival Outcomes

- **Random Survival Forests**

Steps:

1. Draw B booststrap samples from the original data (63% in the bag data, 37% Out of bag data(OOB)).
2. Grow a survival tree for each bootstrap sample. randomly select p candidate variables, and splits the node using the candidate variable that maximizes survival difference between daughter nodes.
3. Grow the tree to full size, each terminal node should have no less than $d_0 > 0$ unique deaths.
4. Calculate a Cumulative Hazard Function (CHF) for each tree. Average to obtain the ensemble CHF.
5. Using OOB data, calculate prediction error for the ensemble CHF.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S. (2008). Random Survival Forests. Annals of Applied Statistics 2, 841–860.

Survival Outcomes

- **Survival Tree**

Survival trees is similar to decision tree which is built by recursive splitting of tree nodes. A node of a survival tree is considered “pure” if the patients all survive for an identical span of time.

A good split for a node maximizes the survival difference between daughters. The logrank is most commonly used dissimilarity measure that estimates the survival difference between two groups. For each node, examine every allowable split on each predictor variable, then select and execute the best of these splits.

- **Logrank Test**

The logrank test is obtained by constructing a (2 X 2) table at each distinct death time, and comparing the death rates between the two groups, conditional on the number at risk in the groups.

LeBlanc, M. and Crowley, J. (1993). Survival Trees by Goodness of Split. Journal of the American Statistical Association 88, 457–467.

Survival Outcomes

Let $t_1, \ldots, t_K$ represent the K ordered, distinct death times. At the j-th death time, we have the following table:

	Die/Fail		
Group	Yes	No	Total
0	d_{0j}	$r_{0j} - d_{0j}$	r_{0j}
1	d_{1j}	$r_{1j} - d_{1j}$	r_{1j}
Total	d_j	$r_j - d_j$	r_j

$$X^2_{logrank} = \frac{\left[\sum_{j=1}^{K}\left(d_{0j} - r_{0j} \times d_j / r_j\right)\right]^2}{\sum_{j=1}^{K} \frac{r_{1j} r_{0j} d_j (r_j - d_j)}{r_j^2 (r_j - 1)}}$$

the numerator is the squared sum of deviations between the observed and expected values. The denominator is the variance of the d_{0j} (Patnaik ,1948).

The test statistic, $X^2_{logrank}$, gets bigger as the differences between the observed and expected values get larger, or as the variance gets smaller.

Segal, Mark Robert. "Regression trees for censored data." Biometrics (1988): 35-47.

Evaluation of Performance

- **Brier Score (Mean Squared Error)**

Brier Score is used to evaluate the binary outcomes, which is simply defined as $(Y - \hat{Y})^2$, where the squared difference between actual outcomes Y and predictions $\hat{Y}$ are calculated.

- **R^2 (R-squared or coefficient of determination**

The amount of R^2 is an overall measure to quantify the amount of information in a model for a given data set. It gives the percent of total variation that is described by the variation in X.

The total variation is defined as: $TotalVar(y) = \sum_{i=1}^{N}(y_i - \bar{y})^2$

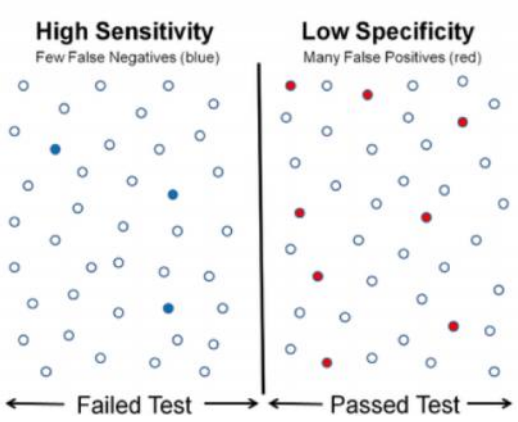
$\bar{y}$ is the mean value of the y, and N is the total number of individuals.

$$R^2 = 1 - \frac{MSE}{TotalVar(y)}$$

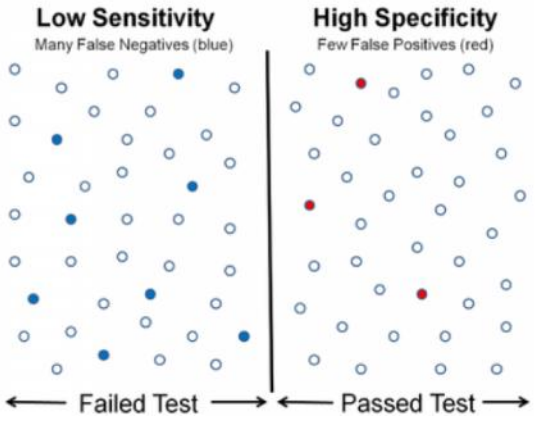
Sensitivity and Specificity

- **Sensitivity:** Proportion of actual positives which are correctly identified.
- **Specificity:** Proportion of actual negatives which are correctly identified.

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	**True Positive**	**False Positive** (Type I error)	Positive predictive value = Σ True Positive / Σ Test Outcome Positive
	Test Outcome Negative	**False Negative** (Type II error)	**True Negative**	Negative predictive value = Σ True Negative / Σ Test Outcome Negative
		Sensitivity = Σ True Positive / Σ Condition Positive	**Specificity** = Σ True Negative / Σ Condition Negative	

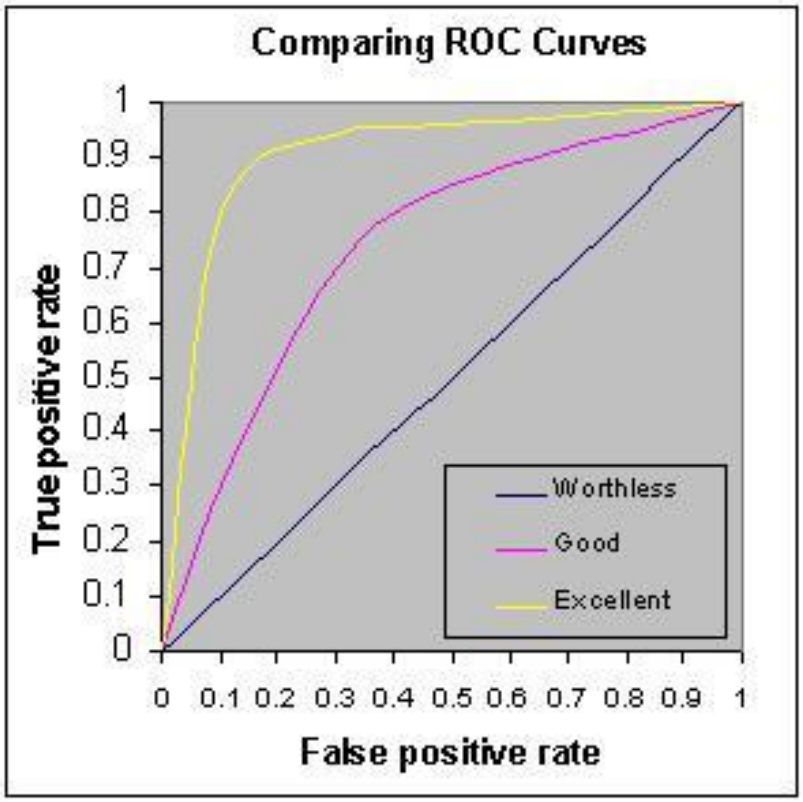


Source: Wikipedia Page



ROC Curve

A ROC Curve is been used to quantify the diagnostic value of a test over its whole range of possible cutoffs for classifying patients as positive vs. negative. In each possible cutoff the true positive rate and false positive rate will be calculated as the X and Y coordinates in the ROC Curve.



C-Statistic

- The C-statistic is a rank order statistic for predictions against true outcomes. Commonly used concordance measure is

- The c-statistic also known as the concordance probability is defined as the ratio of the concordant to discordant pairs.
- A (i,j) pair is called a **concordant pair** if $t_i > t_j$ and $S(t_i) > S(t_j)$. t_i and t_j are the survival times and $S(t_i)$ and $S(t_j)$ are the predicted survival times.
- A **discordant pair** is one where if $t_i > t_j$ and $S(t_i) < S(t_j)$.
- For a binary outcome c is identical to the area under the ROC curve.

- In c-index calculation all the pairs where the instance with shorter survival time is censored are not considered as comparable pairs.

Uno, Hajime, et al. "On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data." Statistics in medicine 30.10 (2011): 1105-1117.

CASE STUDY: HEART FAILURE READMISSION PREDICTION

Penalties for Poor Care - 30-Day Readmissions

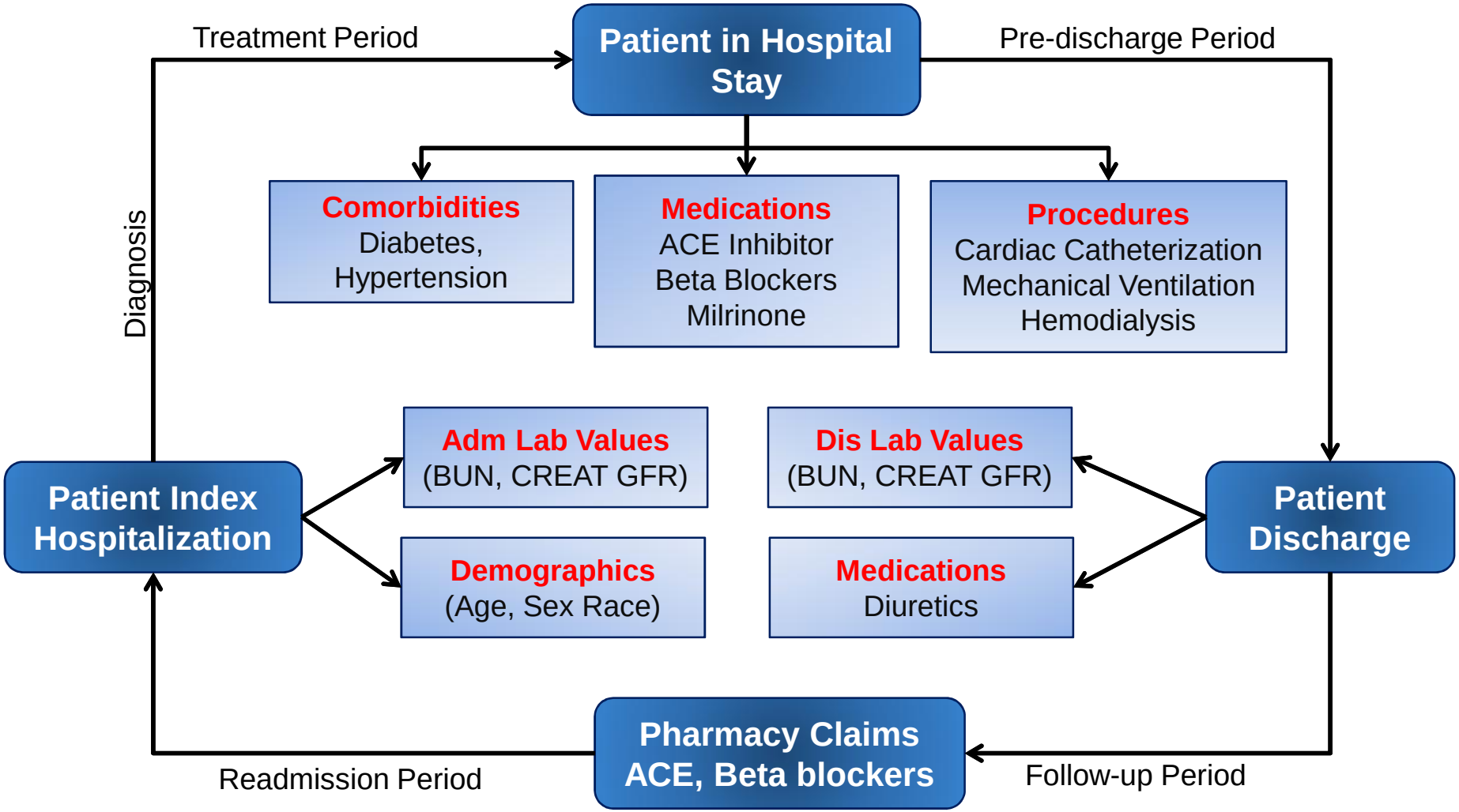
- Hospitalizations account for more than 30% of the 2 trillion annual cost of healthcare in the United States. Around 20% of all hospital admissions occur within 30 days of a previous discharge.
  - not only expensive but are also potentially harmful, and most importantly, they are often preventable.

- Medicare penalties from FY12 - heart failure, heart attack, and pneumonia.

- Identifying patients at risk of readmission can guide efficient resource utilization and can potentially save millions of healthcare dollars each year.

- Effectively making predictions from such complex hospitalization data will require the development of novel advanced analytical models.

Patient Readmission Cycle



Challenges with Clinical Data Analysis

- Integrating Multi-source Data – Data maintained over different sources such as procedures, medications, labs, demographics etc.
- Data Cleaning – Many labs and procedures have missing values.
- Non-uniformity in Columns – Certain lab values are measured in meq/DL and mmol/DL.
  - Solution: Apply the logarithmic transformation to standardize and normalize the data.
- Multiple instances for unique patients – Several records from different data sources are available for the same hospitalization of a patient.
  - Solution: Obtain summary statistics such as minimum, maximum and average for aggregating multiple instances into a single one.

Snapshot of EHR Data

DEMOGRAPHICS

ptID	Age	Sex	Race
1665	57	M	African American

LABS DATA

ptID	LABName	LABval	Low	High	Time	Date
1665	BUN	24	10	25	12:00	07/01/2007
1665	BUN	28	10	25	12:30	07/01/2007
1665	BUN	32	10	25	09:00	07/02/2007
1665	BUN	26	10	25	18:00	07/02/2007
1665	BUN	30	10	25	09:00	07/03/2007

CLINICAL FEATURE TRANSFORMATION

ptID	Log(BMAX)	Log(BMIN)	Log(BAVG)	BUNCOUNT	ALR
1665	1.5	1.38	1.44	5	0.8

ALR – Abnormal Labs Ratio

BMAX/BMIN/BAVG- BUN maximum/minimum/average

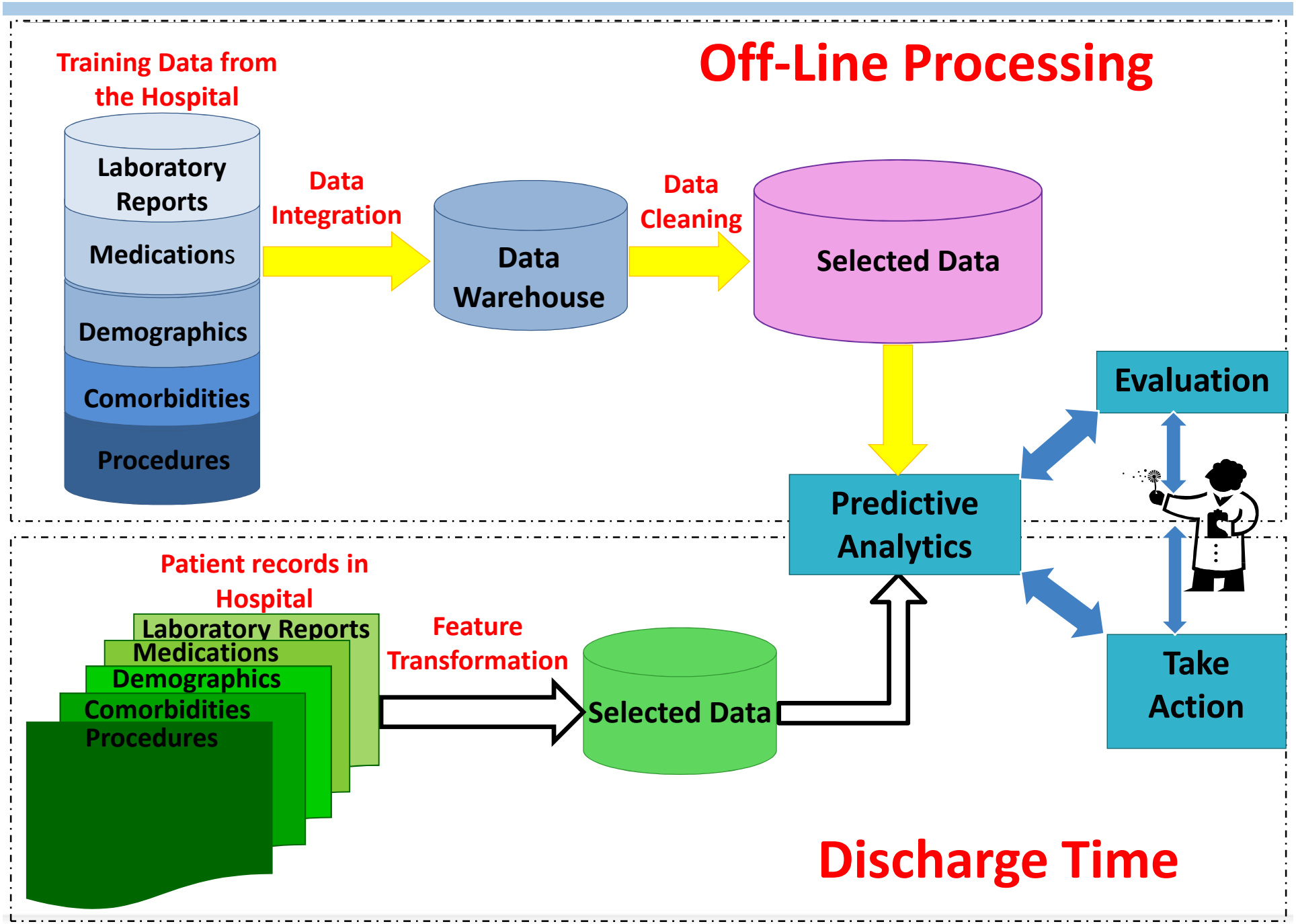
Snapshot of EHR Data

PROCEDURES AND MEDICATIONS

ptID	Procedures	Medications	Time	Date
1665	Cardiac cath	ACE inhibitor	12:00	07/01/2007
1665	Mech vent	ACE inhibitor	12:30	07/01/2007
1665	Cardiac cath	Beta blocker	09:00	07/02/2007
1665	Cardiac cath	ACE inhibitor	18:00	07/02/2007
1665	Mech vent	Beta blocker	09:00	07/03/2007

CLINICAL FEATURE TRANSFORMATION

ptID	ACE inhib	Beta block	Cardiac cath	Mech vent
1665	3	2	3	2


Off-Line Processing
Training Data from the Hospital
Laboratory Reports
Medications
Demographics
Comorbidities
Procedures
Data Integration
Data Warehouse
Data Cleaning
Selected Data
Evaluation
Predictive Analytics
Take Action
Patient records in Hospital
Laboratory Reports
Medications
Demographics
Comorbidities
Procedures
Feature Transformation
Selected Data
Discharge Time

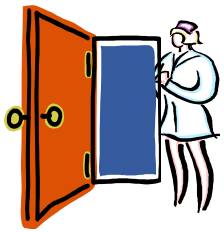
Interventions – Depending on the Risk scores at the Discharge

HOSPITAL RESOURCE UTILIZATION

Predictive Analytics

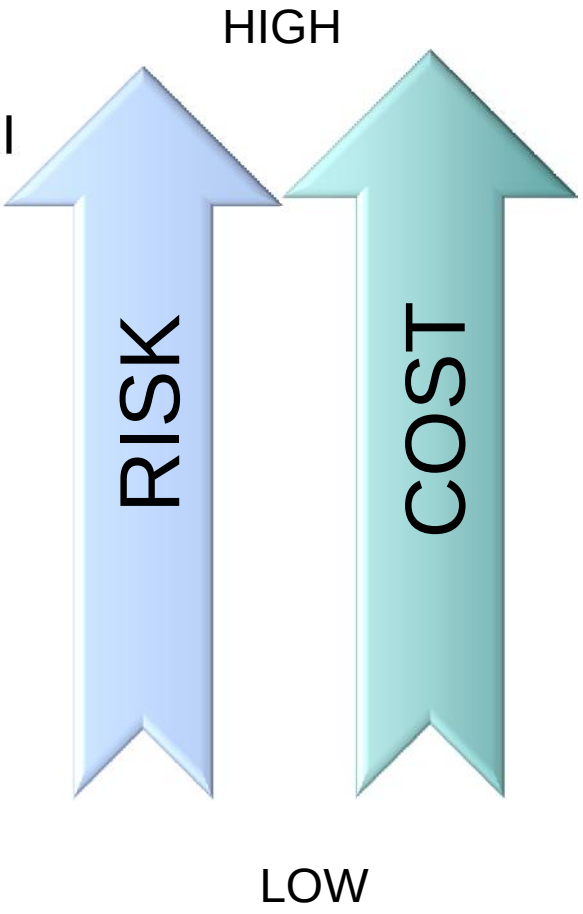
- Staying back in hospital

- Home nurse visitation



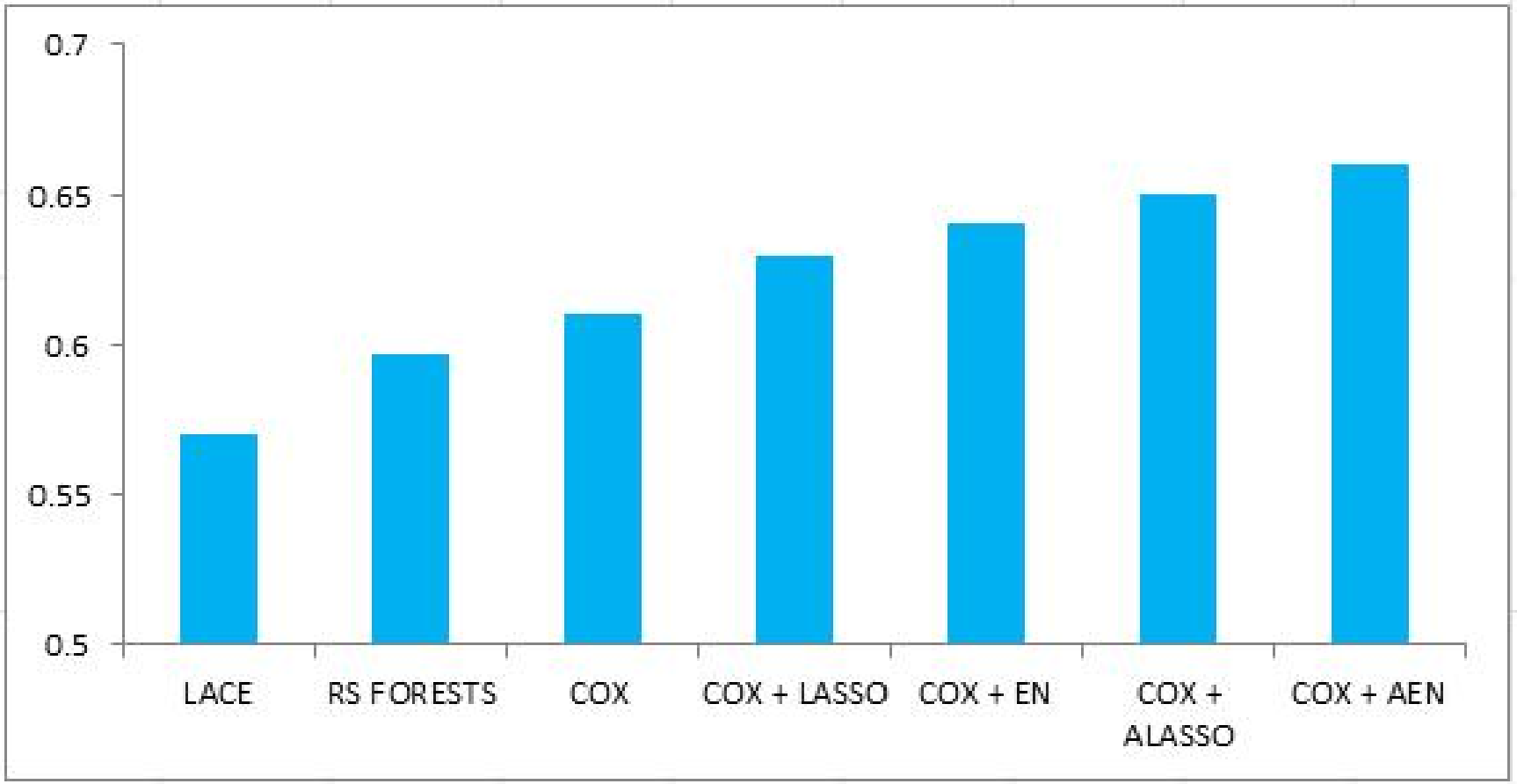
- Telephone support

- Automated reminder



THE COST SPENT FOR PATIENT CARE IS DETERMINED BY THE RISK

AUC Performance Comparisons



Heart Failure Patient data collected at Henry Ford Hospital between 2000-2009.

Biomarker Identification

Variables	Association	LE	COX	COX +LASSO	COX +ELASTIC
HGB	0.81	✗	✓	✓	✓
ckd	0.75	✗	✓	✓	✓
diabetes	0.71	✗	✗	✓	✓
hypertension	0.70	✗	✓	✓	✓
BUN/CREAT	0.66	✓	✓	✓	✓
age	0.66	✓	✗	✗	✗
cad	0.61	✗	✗	✓	✓
Heart failure	0.60	✓	✗	✓	✓
afib	0.60	✗	✗	✓	✓
HAP	0.57	✗	✗	✓	✓
pvd	0.56	✗	✗	✓	✓

Ross, Joseph S., et al. "Statistical models and patient predictors of readmission for heart failure: a systematic review." *Archives of internal medicine* 168.13 (2008): 1371.

Top Biomarkers obtained from Regularized COX

Biomarkers	Association	Values
HGB	Negative	HGB < 12.3 mg/DL
BUN	High Positive	BUN > 40 mg/DL
CREAT	High Positive	CRET > 1.21 mg/DL
GFR	High Negative	GFR < 48.61 mg/DL
PHOS	Positive	PHOS > 3.9 meq/DL
K	Positive	K > 3.7 meq/DL
MG	Positive	MG > 1.2 meq/DL
LDL	Negative	LDL < 0.6 mg/DL

Already known and well-studied Biomarkers

Newly found Biomarkers- Clinicians showing some interest

Conclusions and Future of our Study

- The performance of regularized Cox algorithms is better than that of simple Cox regression and other standard predictive algorithms.
- Biomarkers selected through this method are clinically relevant.
- One advantage of Cox models is that there is no re-training needed if we change the time of interest (from 30 days to 90 days).
- Currently working on temporal modeling using survival analysis.
- Adding claims data for a partial set of patients.