On Big Data Applications

P. M. Pardalos[1], [2]

[1]Center for Applied Optimization
Department of Industrial and Systems Engineering
University of Florida

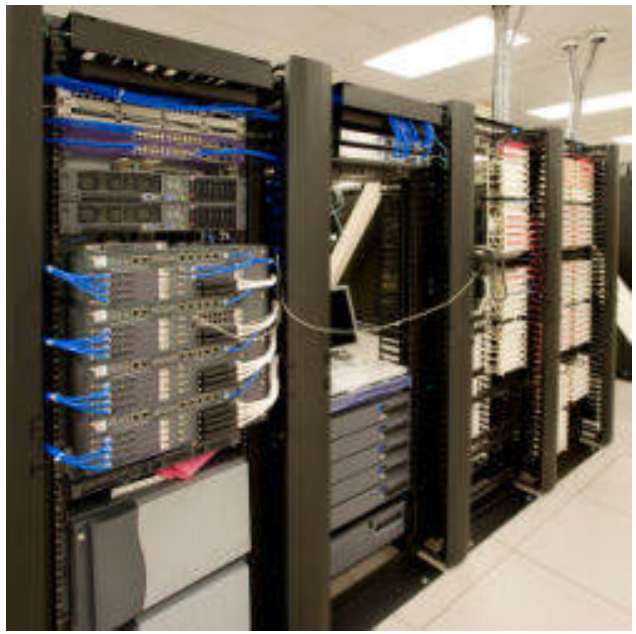
[2]Laboratory of Algorithms and Technologies for Networks Analysis (LATNA)
National Research University Higher School of Economics

ItForum, 2013

How big will the data be tomorrow?

Cisco: **Global 'Net Traffic to Surpass 1 Zettabyte in 2016**
http://cacm.acm.org/news/150041-cisco-global-net-traffic-to-surpass-1-zettabyte-in-2016/fulltext

SI decimal prefixes		Binary usage	IEC binary prefixes	
Name (Symbol)	Value		Name (Symbol)	Value
kilobyte (kB)	10^3	2^{10}	kibibyte (KiB)	2^{10}
megabyte (MB)	10^6	2^{20}	mebibyte (MiB)	2^{20}
gigabyte (GB)	10^9	2^{30}	gibibyte (GiB)	2^{30}
terabyte (TB)	10^{12}	2^{40}	tebibyte (TiB)	2^{40}
petabyte (PB)	10^{15}	2^{50}	pebibyte (PiB)	2^{50}
exabyte (EB)	10^{18}	2^{60}	exbibyte (EiB)	2^{60}
zettabyte (ZB)	10^{21}	2^{70}	zebibyte (ZiB)	2^{70}
yottabyte (YB)	10^{24}	2^{80}	yobibyte (YiB)	2^{80}



What is BigData?

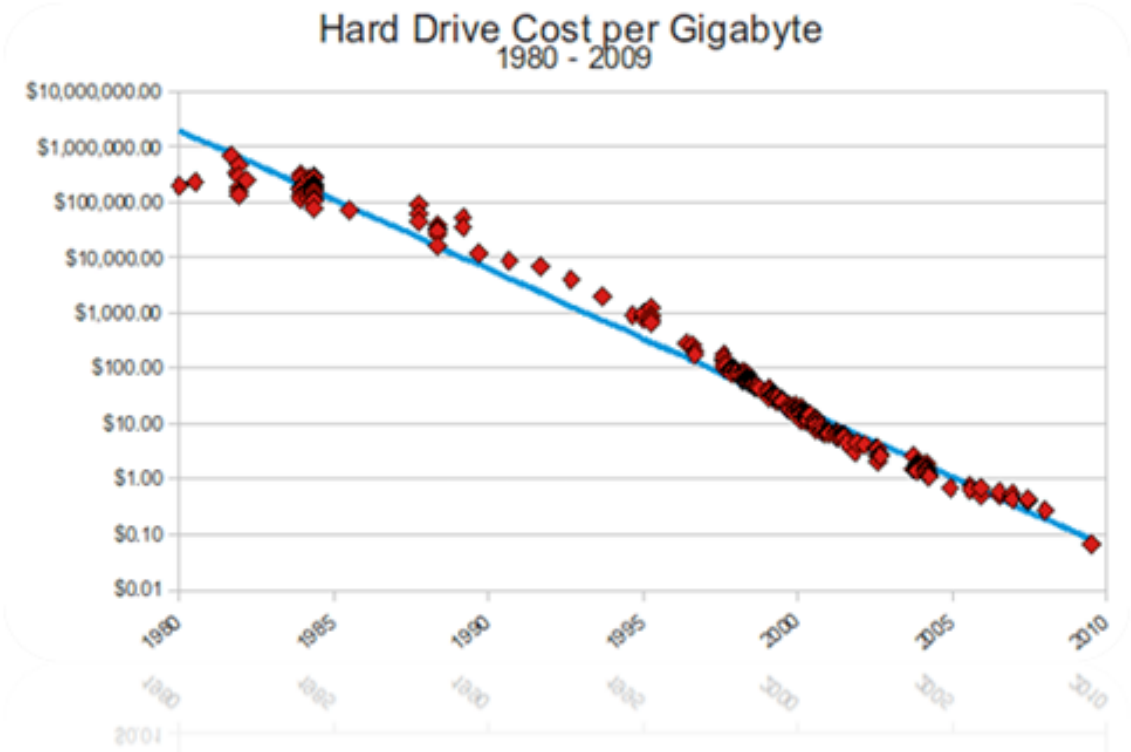
- The proliferation of massive datasets brings with it a series of special computational challenges.
- This data avalanche arises in a wide range of scientific and commercial applications.
- With advances in computer and information technologies, many of these challenges are beginning to be addressed.

HandBook of Massive Datasets

Handbook of Massive Data Sets, co-editors: J. Abello, P.M. Pardalos, and M. Resende, Kluwer Academic Publishers, (2002).

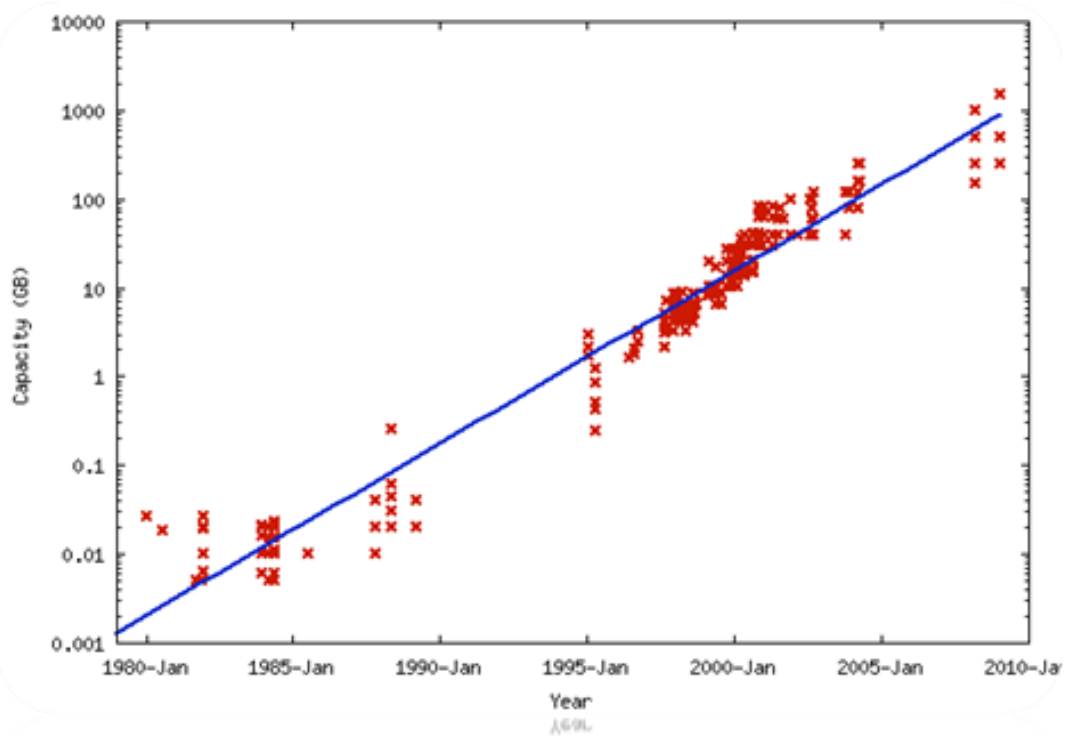
- Data Mining: the practice of searching through large amounts of computerized data to find useful patterns or trends.
- Optimization: An act, process, or methodology of making something (as a design, system, or decision) as fully perfect, functional, or effective as possible; specifically : the mathematical procedures (as finding the maximum of a function) involved in this.

Hard drive Cost



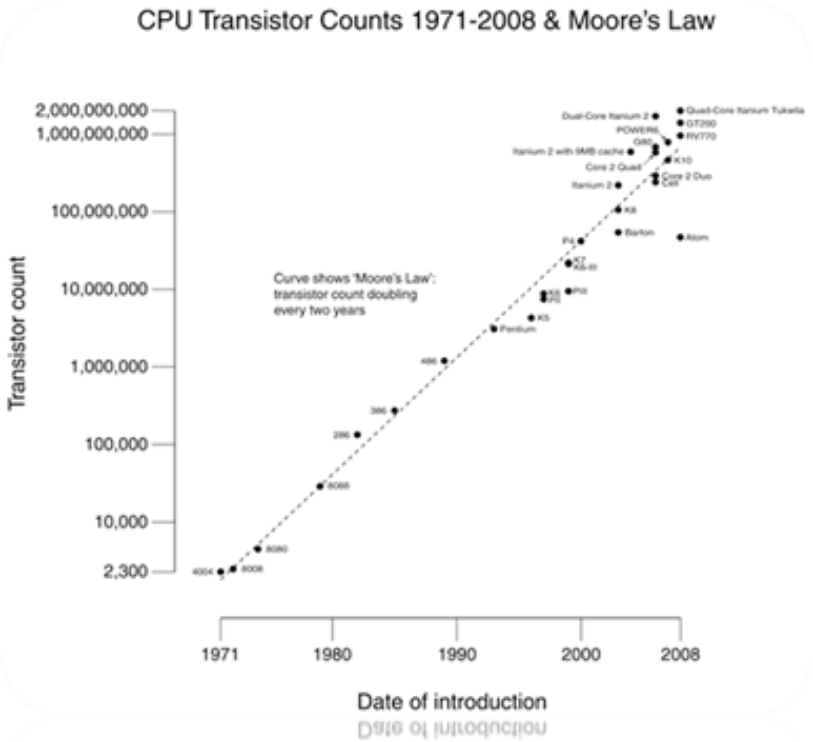
- Approximately 1/10 cheaper every 5 years

Hard drive Capacity



- Approximat ely 10 times more every 5 years

Processing power



- Number of transistors of a computer processor double every two years

Some Examples of Massive Datasets

- Internet Data
- Weather Data
- Telecommunications Data
- Medical Data
- Demographics Data
- Financial Data

Handling Massive Datasets

New computing technologies are needed to handle massive datasets

- Input-Output Parallel Systems
- External Memory Algorithms
- Quantum Computing

An example in biomedicine

- Acute Kidney Injury (AKI) is one of common complications in post-operative patients.
- The in-hospital mortality rate for patients with AKI may be as high as 60%.
- An elevated serum creatinine (sCr) level in blood is commonly recognized as an indicator of AKI.
- The degree to which patterns of sCr change are associated with in-hospital mortality is unknown.

Study Outline

Objectives:

- To develop a comprehensive model for assessing the mortality risk in post-operative patients
- To establish a quantitative relationship between sCr pattern and mortality risk

Results:

- A probabilistic model was designed to assess mortality risk in post-operative patients
- The model provided high discriminative capacity and accuracy (ROC = 0.92)
- A quantitative association between sCr time series and mortality risk was established
- Simple and informative sCr risk factors were derived

Data Set

- We performed a retrospective study involving patients admitted to Shands Hospital (Gainesville, FL) from 2000 through 2010.
- For each patient who underwent a surgery detailed clinical and outcome data were collected.
- The analysis included 60,074 patients who underwent in-patient surgery at Shands Hospital during a 10-years period.

Logistic Function

Probabilistic score [1]

$$\mathbb{P}(C=1|X=x)=\left(1+\exp\left(w_0+\sum_{i=1}^{m} w_i g_i(x_i)\right)\right)^{-1}; \quad (1)$$

$C \in \{0,1\}$ is an outcome;
$X=(X_1,\ldots,X_m)$ - the risk factors;
g_i - nonlinear function associated with the ith factor
$w_0=\ln \mathbb{P}(C=1)/\mathbb{P}(C=0)$ is the a priori log odds ratio.
$\{w_i\},\ i>0$ - weights indicating importance of the risk factors

[1] S. Saria, A. Rajani,J. Gould,D. Koller,A. Penn, Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants, Sci Transl Med, 2010

Nonlinear Risk Factors Estimation

From the probabilistic score equation we can derive

$$w_0 + \sum_{i=1}^{m} w_i \cdot g_i(x_i) = \log \frac{\mathbb{P}(X = x|C = 0)}{\mathbb{P}(X = x|C = 1)} + \log \frac{\mathbb{P}(C = 0)}{\mathbb{P}(C = 1)}$$

Under assumption of risk factors independence:

$$w_0 + \sum_{i=1}^{m} w_i \cdot g_i(x_i) = \sum_{i=1}^{m} log \frac{\mathbb{P}(X_i = x_i|C = 0)}{\mathbb{P}(X_i = x_i|C = 1)} + \log \frac{\mathbb{P}(C = 0)}{\mathbb{P}(C = 1)}.$$

For continuous factors $\mathbb{P}(X_i = x_i|C)$ implies
$\mathbb{P}(X_i \in [x_i - \varepsilon, x_i + \varepsilon]|C)$ where ε is sufficiently small

Model Parameters

$$\max_{w} \sum_{k\in\{0,1\}} \sum_{j:\ C^j=k} \ln \mathbb{P}(C^j = k | X_i = x_i^j, \ldots, X_i = x_m^j);$$

j denotes patients in the data set;
$C^j \in \{0, 1\}$ - outcome for the jth patient;
x_i^j - value of ith risk factor for the jth patient.

We solved this problem approximately with interior point method implemented in MATLAB

Nonlinear risk functions

- We get

$$g_i(x_i) = \ln \frac{\mathbb{P}(X_i = x_i|C = 0)}{\mathbb{P}(X_i = x_i|C = 1)};$$

$$w_0 = \ln \frac{\mathbb{P}(C = 1)}{\mathbb{P}(C = 0)};$$

$$w_i = 1, \ i = 1, \ldots, m.$$

- The probabilities $P(X_i = x_i|C), \ i = 1, \ldots, m$ were learned from the data using maximum likelihood principle.

Discrete risk factors estimation

For discrete valued factors

$$g_i(x) = \ln\left(\frac{\#\{j : C^j = 1,\ x_i^j = x\}}{\#\{j : C^j = 1\}}\,\frac{\#\{j : C^j = 0\}}{\#\{j : C^j = 1,\ x_i^j = x\}}\right).$$

Continuous Risk Factors Estimation

For continuous factors we defined

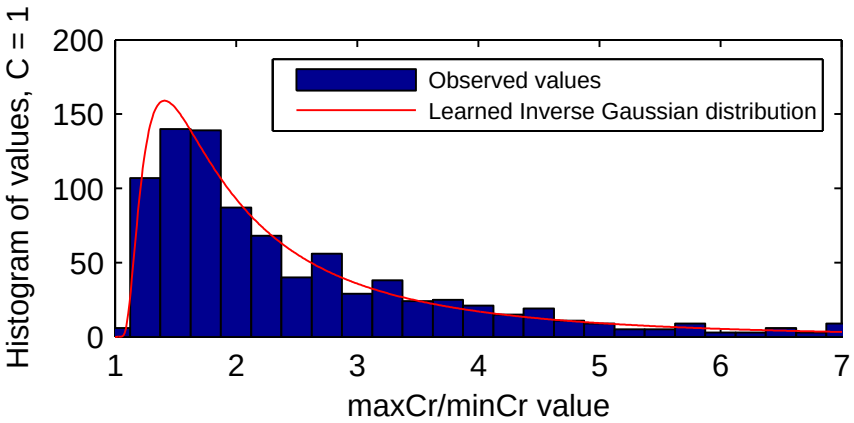
$$g_i(x) = \ln \frac{f_i^0(x)}{f_i^1(x)},$$

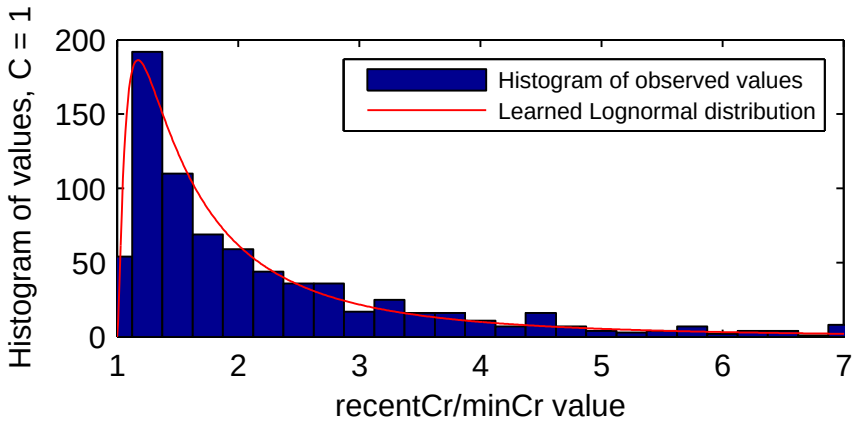
$$f_i^k(x) = \varphi^*(x - h^*, \theta^*),$$

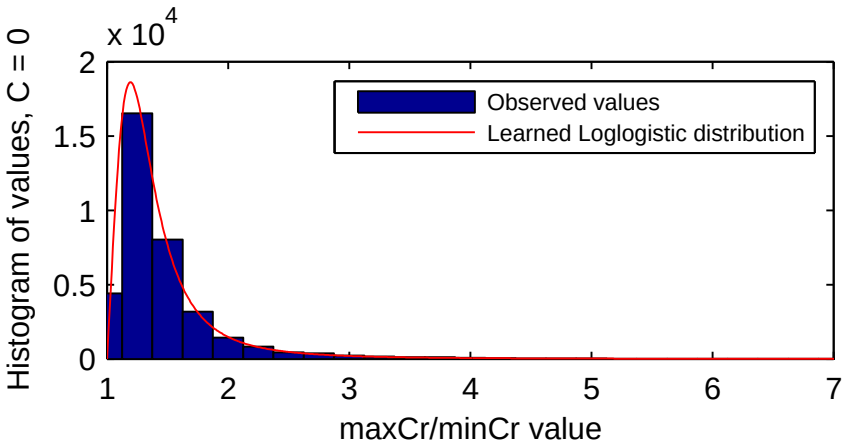
$$\{\varphi^*, h^*, \theta^*\} = \arg \max_{\varphi \in \Phi, \theta, h} L_i^k(\varphi, \theta, h),$$

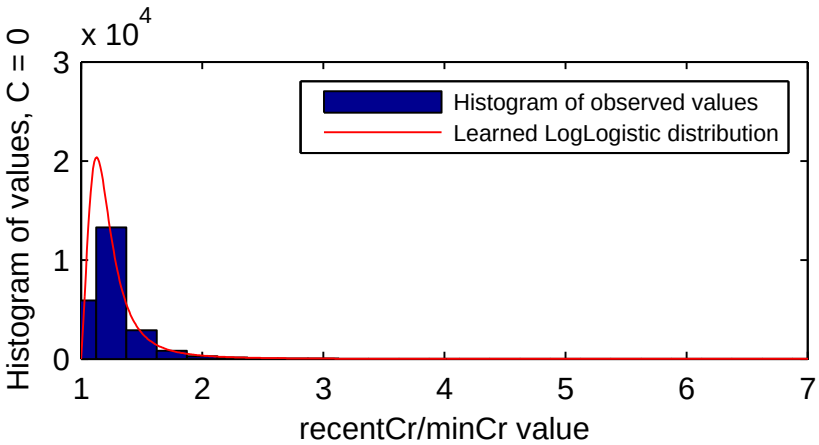
$$L_i^k(\varphi, \theta, h) = \sum_{j: C^j = k} \ln \frac{\varphi(x_i^j - h, \theta)}{F_\varphi(l_i^2, \theta) - F_\varphi(l_i^1, \theta)}.$$

Continuous Risk Factors Estimation

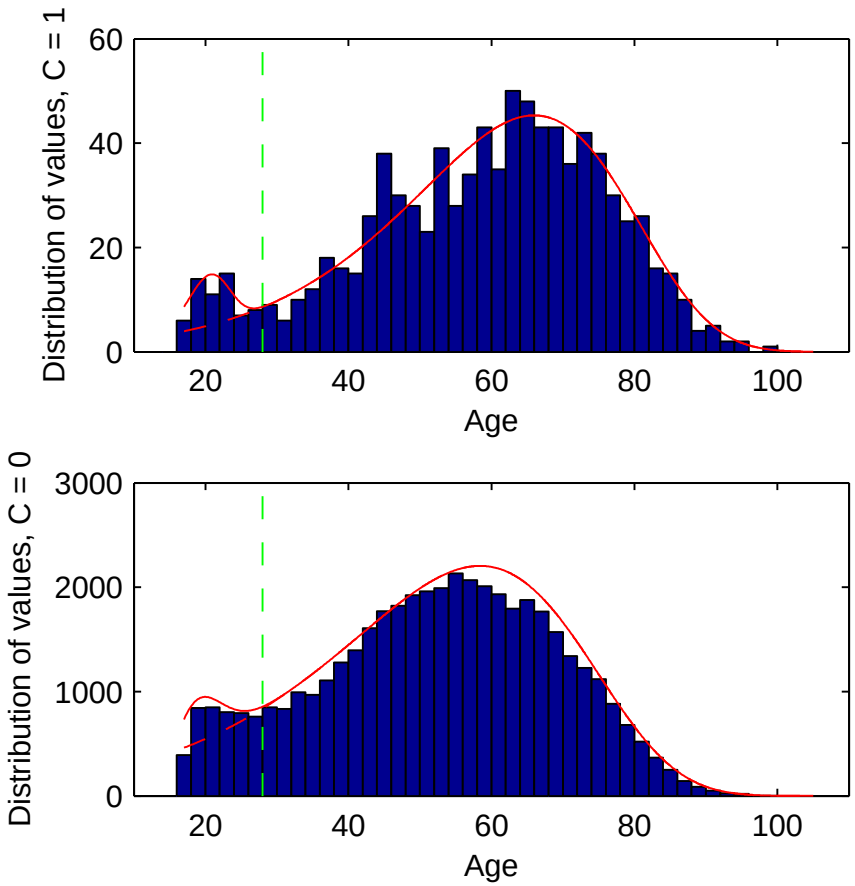








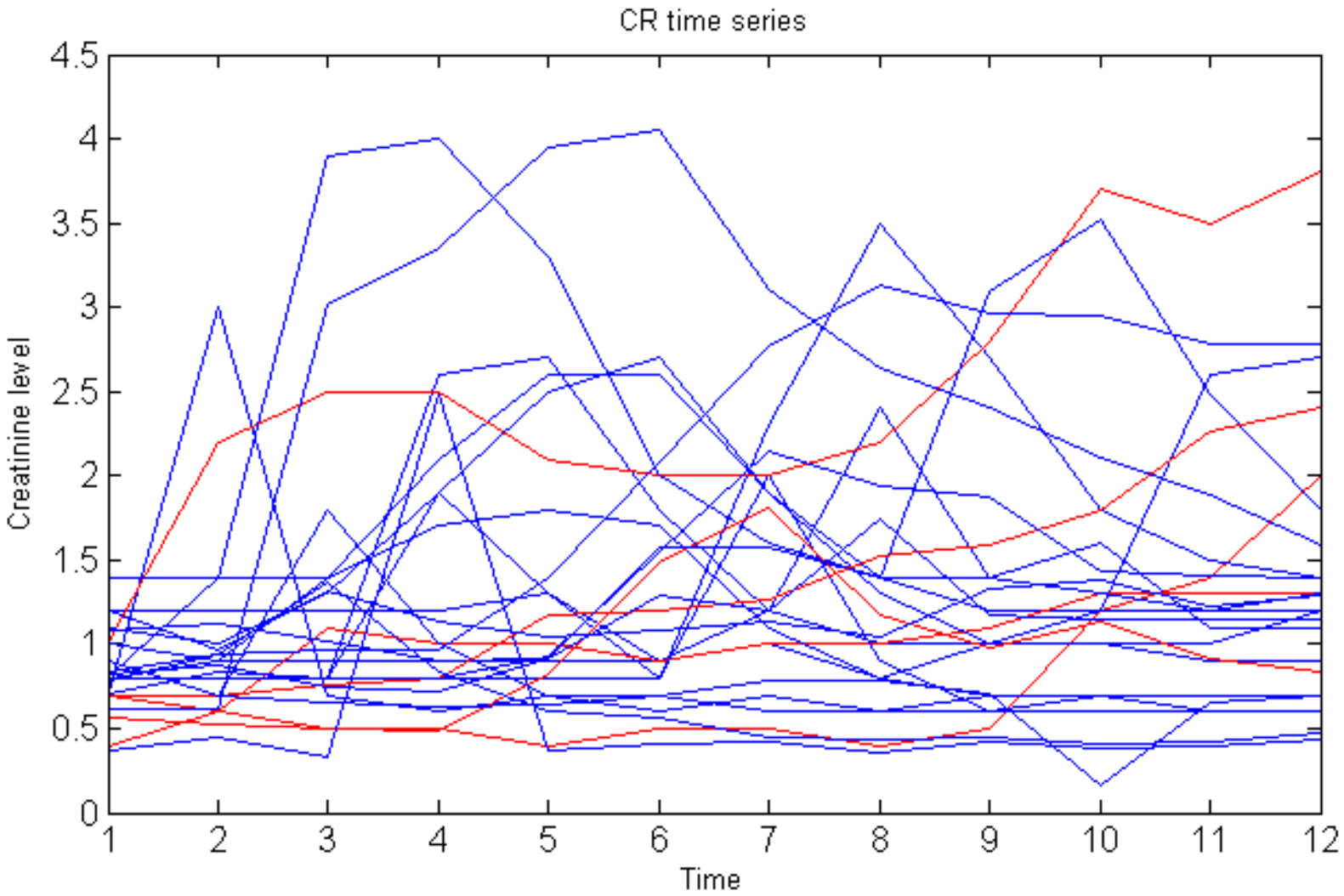
Age Risk Factor



Which creatinine (Cr) characteristics are ”good”?

- Absolute Cr value is not very informative
- (maximal Cr value)/(minimal Cr value) works good
- Recent Cr value is important

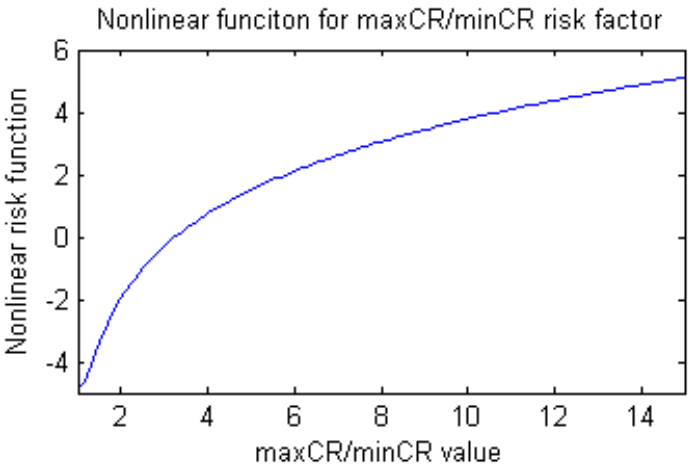
Cr plots in patients with elevated Cr level

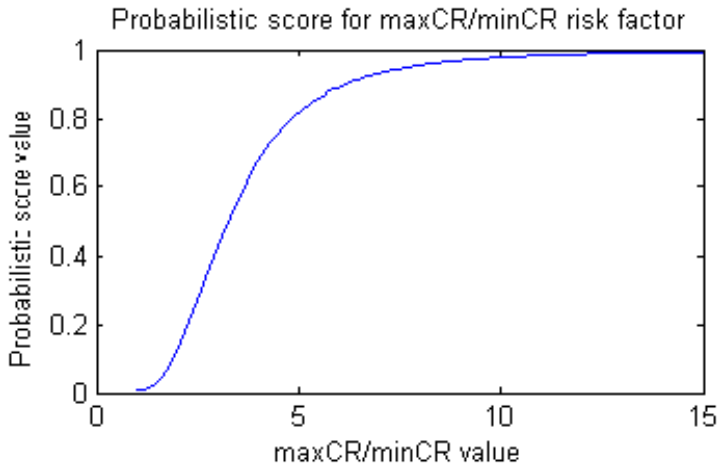


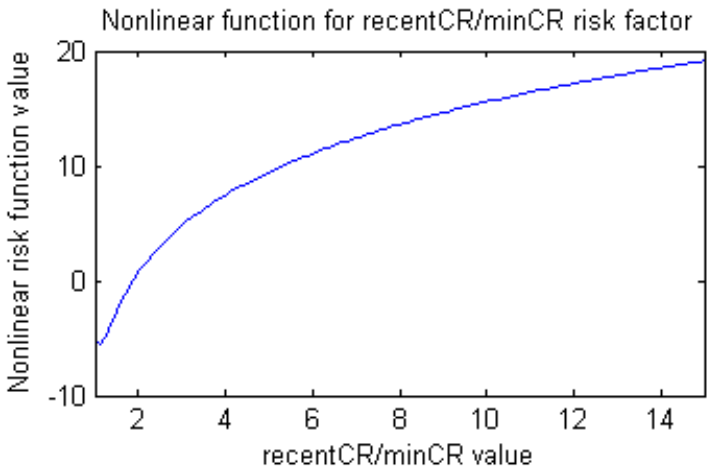
Our Speculations

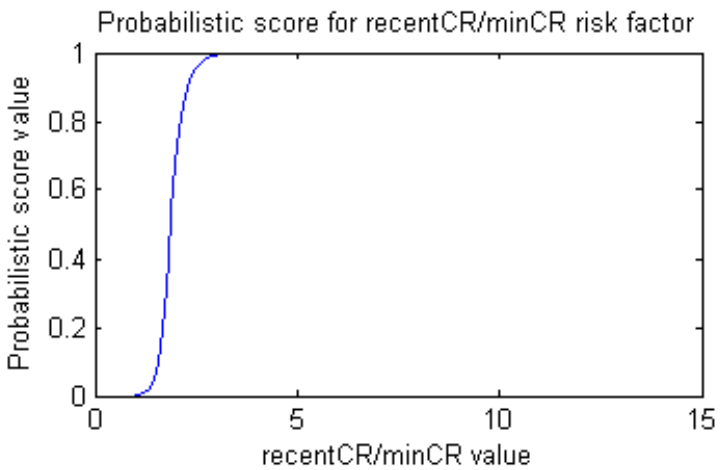
- The risk is defined by a function $R(RC, OD)$,
  RC stands for recent condition
  OD stands for overall damage suffered during AKI
- One should look for features describing these two factors

Resulting nonlinear risk functions





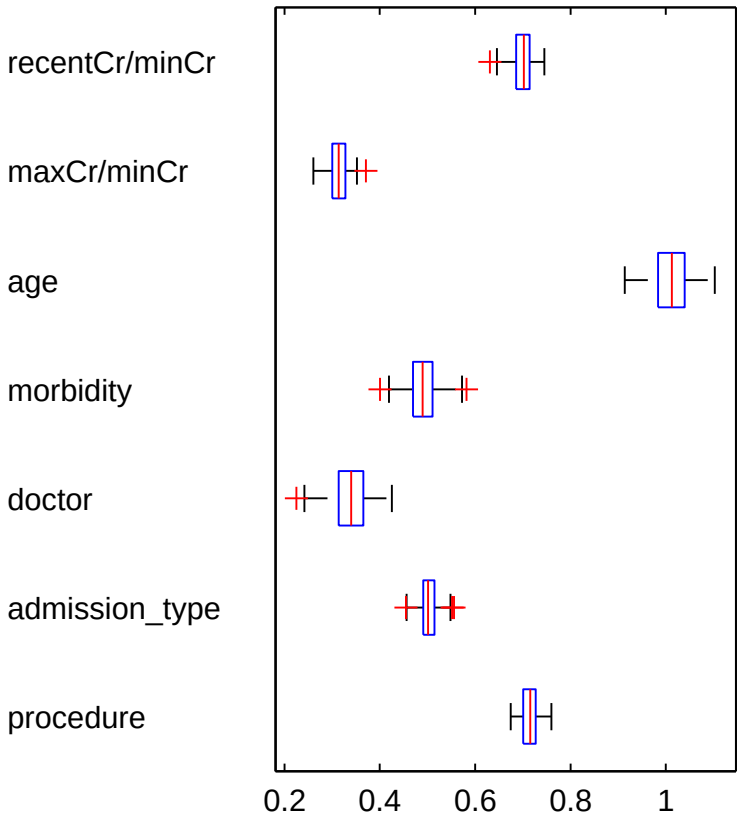


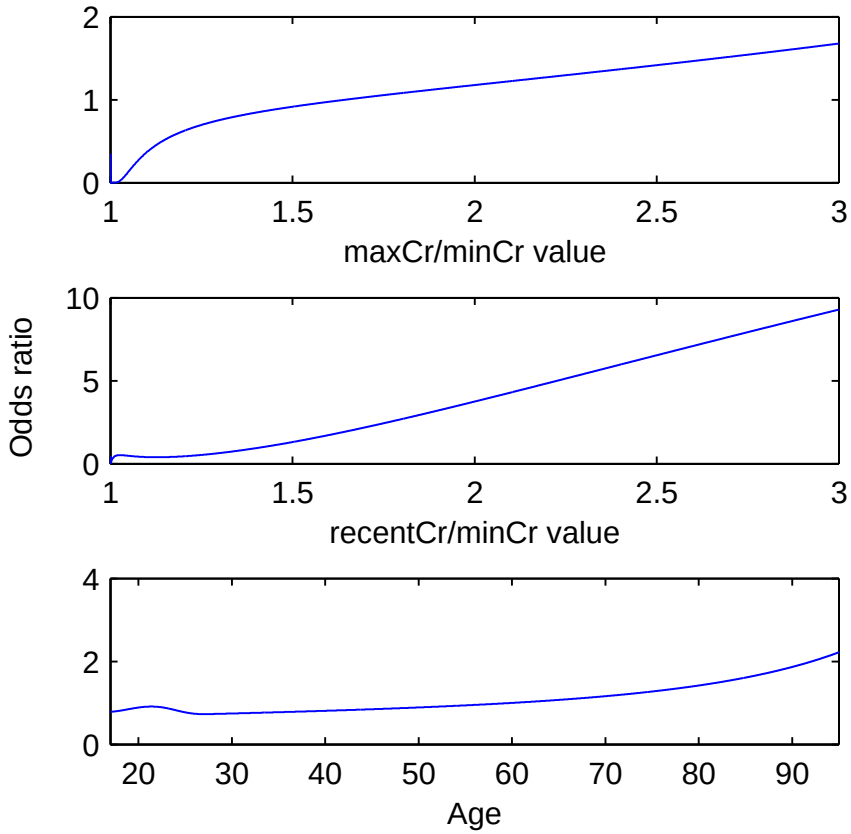


Classification accuracy evaluation

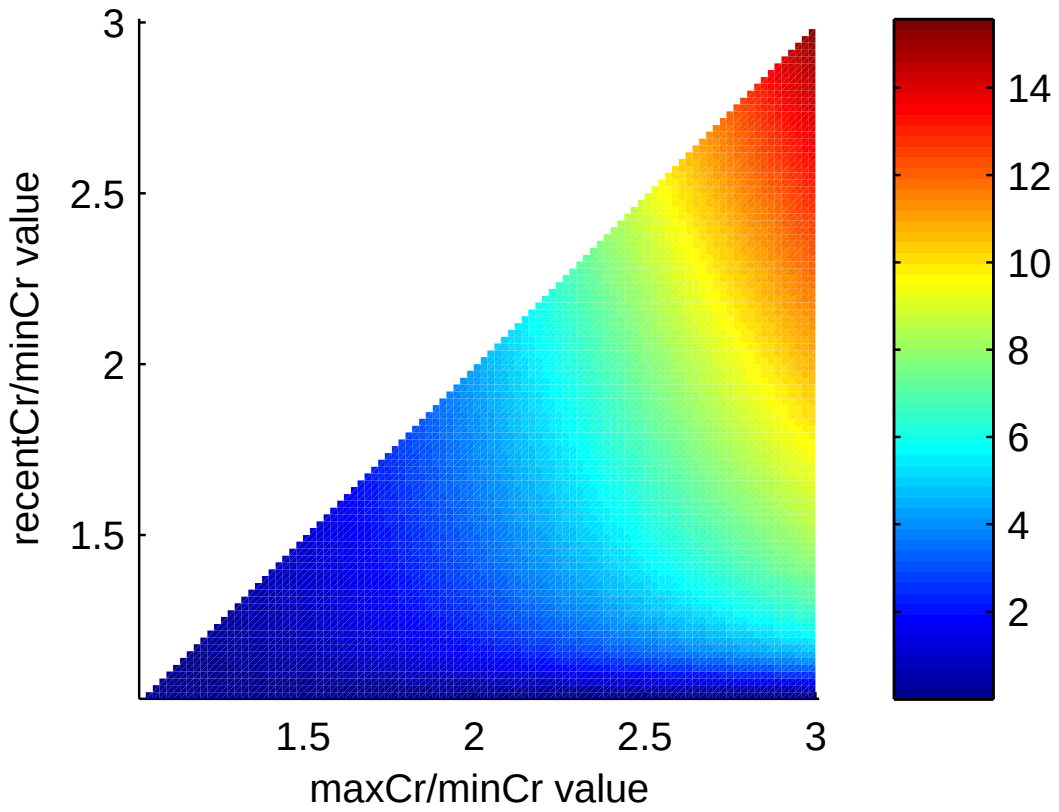
- 70/30 cross-validation analysis has been performed
- Random split into 70% training subset and 30% testing subset
- The average over 100 runs results are reported

Learned risk factors weights

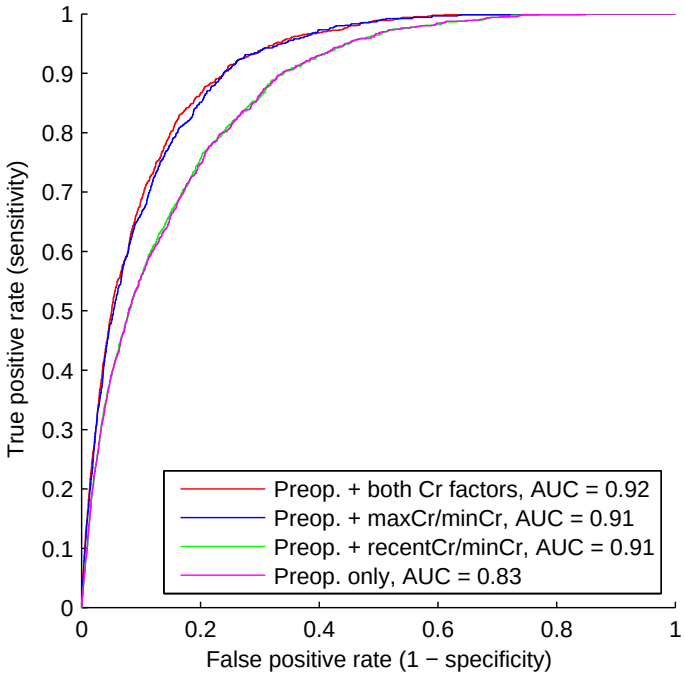


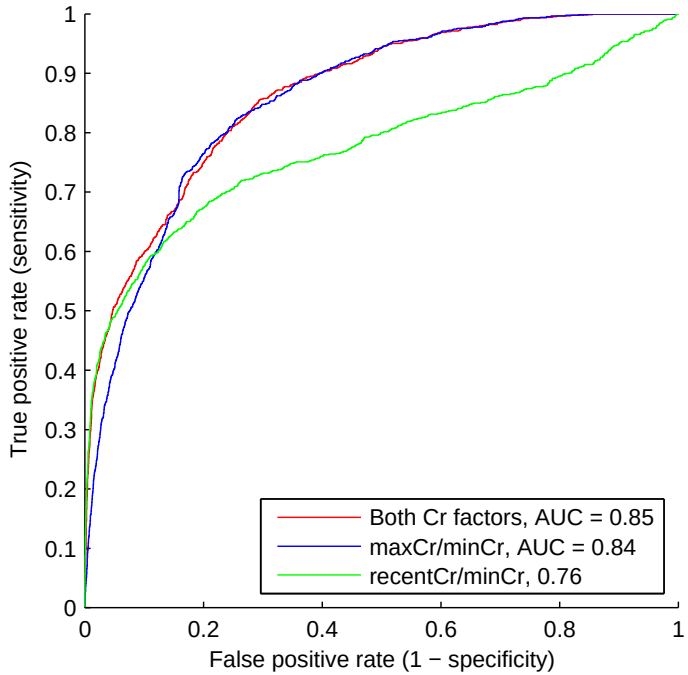


The odds ratio based on Cr factors

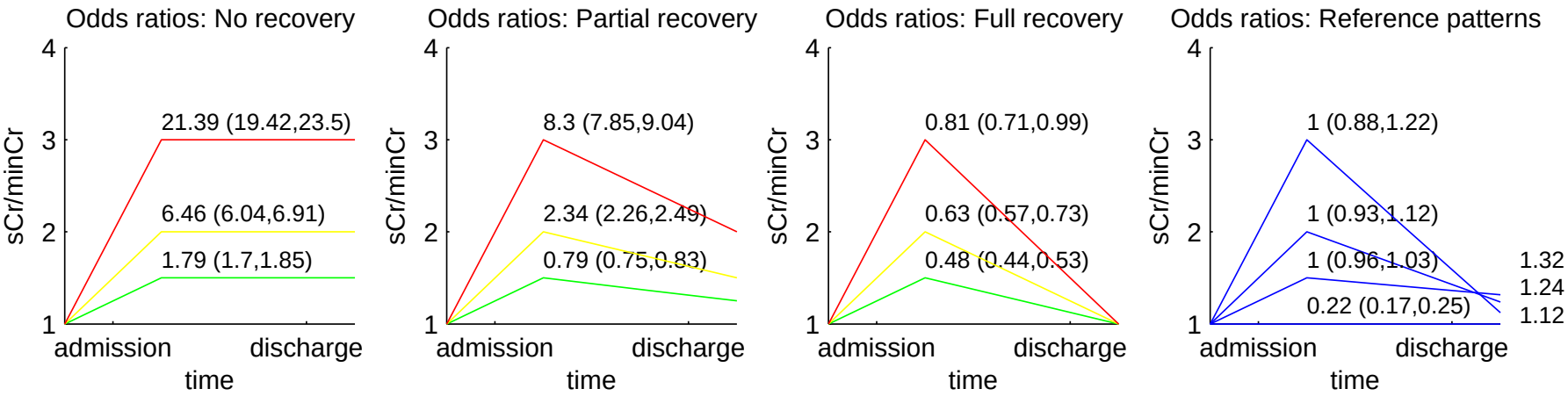


The Resulting ROC Curves

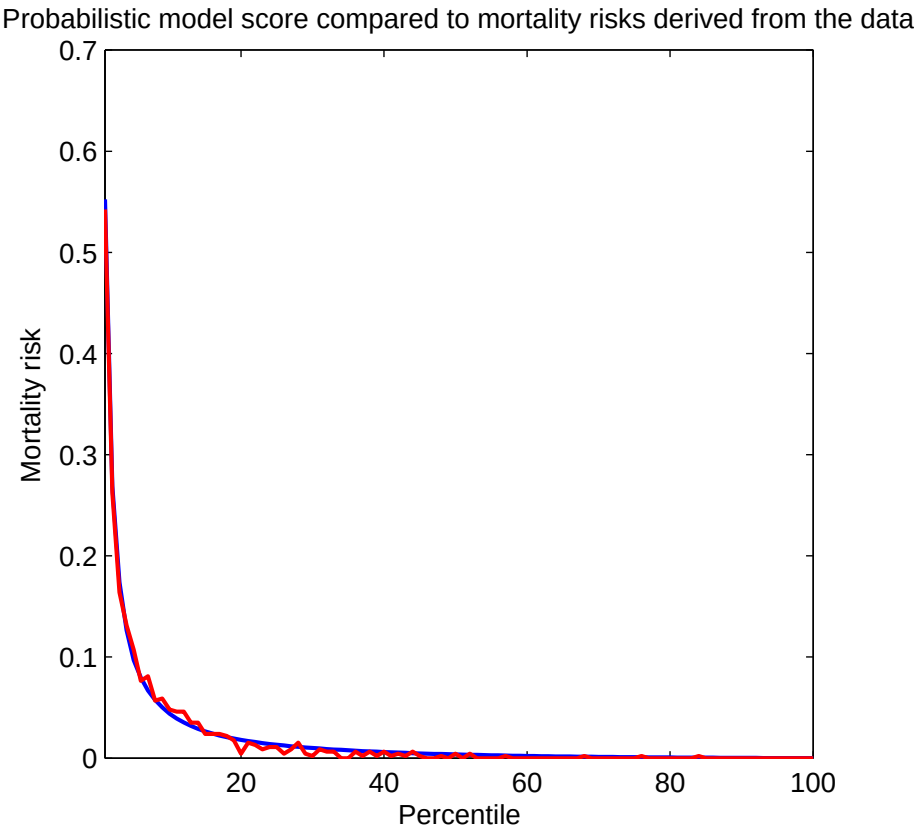




Cr time series patterns



Model Calibration



Risk factors included	Prob. model AUC (95% CI)
Preop.	0.835 (0.817-0.854)
Hosmer-Lemeshow statistics	s = 95.06; p = 0.57
Preop. + maxCr/minCr	0.900 (0.888-0.911)
Hosmer-Lemeshow statistics	s = 84.33; p = 0.84
Preop. + recentCr/minCr	0.910 (0.897-0.923)
Hosmer-Lemeshow statistics	s = 72.19; p = 0.98
Preop. + both sCr factors	0.917 (0.906-0.929)
Hosmer-Lemeshow statistics	s = 68.94; p = 0.99
maxCr/minCr only	0.834 (0.816-0.852)
Hosmer-Lemeshow statistics	s = 1,818; p = 0
recentCr/minCr only	0.791 (0.764-0.818)
Hosmer-Lemeshow statistics	s = 112.88; p = 0.14
both sCr factors	0.855 (0.837-0.873)
Hosmer-Lemeshow statistics	s = 137.31; p = 0.01

Conclusions

- Relatively high classification accuracy was obtained using probabilistic model
- Creatinine time series do provide complimentary information on patient's risk of death
- Presumably AKI effect on risk of death depends on two factors:
  - Kidney recent condition
  - Overall kidney damage suffered since surgery

The End

Thank you!