Conditional Anomaly Detection

Michal Valko, Milos Hauskrecht (CS), G. Cooper, S. Visweswaran, M. Saul (DBMI), A. Seybert (Pharm), J. Harrison, A. Post (PHS, Virginia)

Motivation

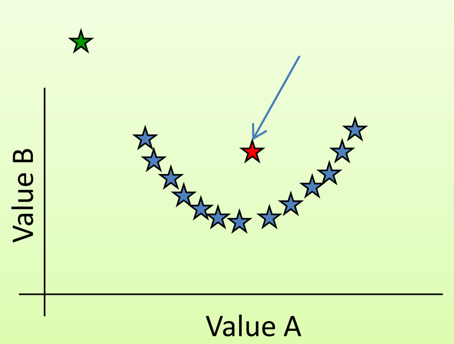
Fact: Medical errors account for 200 000 **preventable** deaths a year.
(Wall Street Journal on July 27, 2004)

Main goal: Detect **anomalies** in clinical **decisions.**

- **Patient records** today have: demographics, conditions, labs, medications administered, procedures performed,…
- **Errors** in decisions are costly and may be life threatening
- **Knowledge-based alerting systems** exist, but are expensive to build and maintain

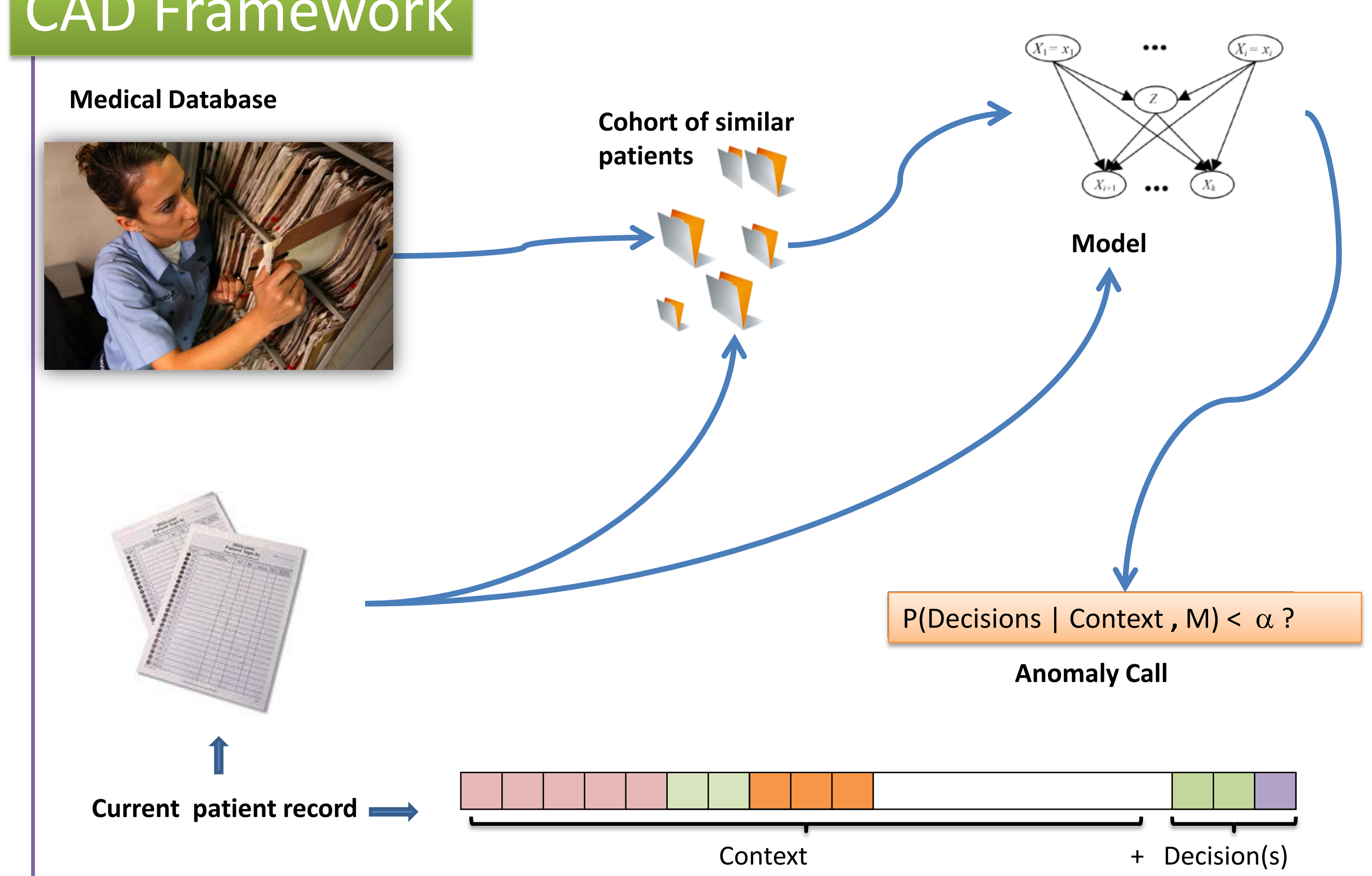
Solution: Evidenced based methods requiring minimal expert knowledge and relying on the historical data.

Conditional Anomaly



In the medical setting: the identification of unusual patient management decisions with respect to the past patients who suffer from the same or similar condition

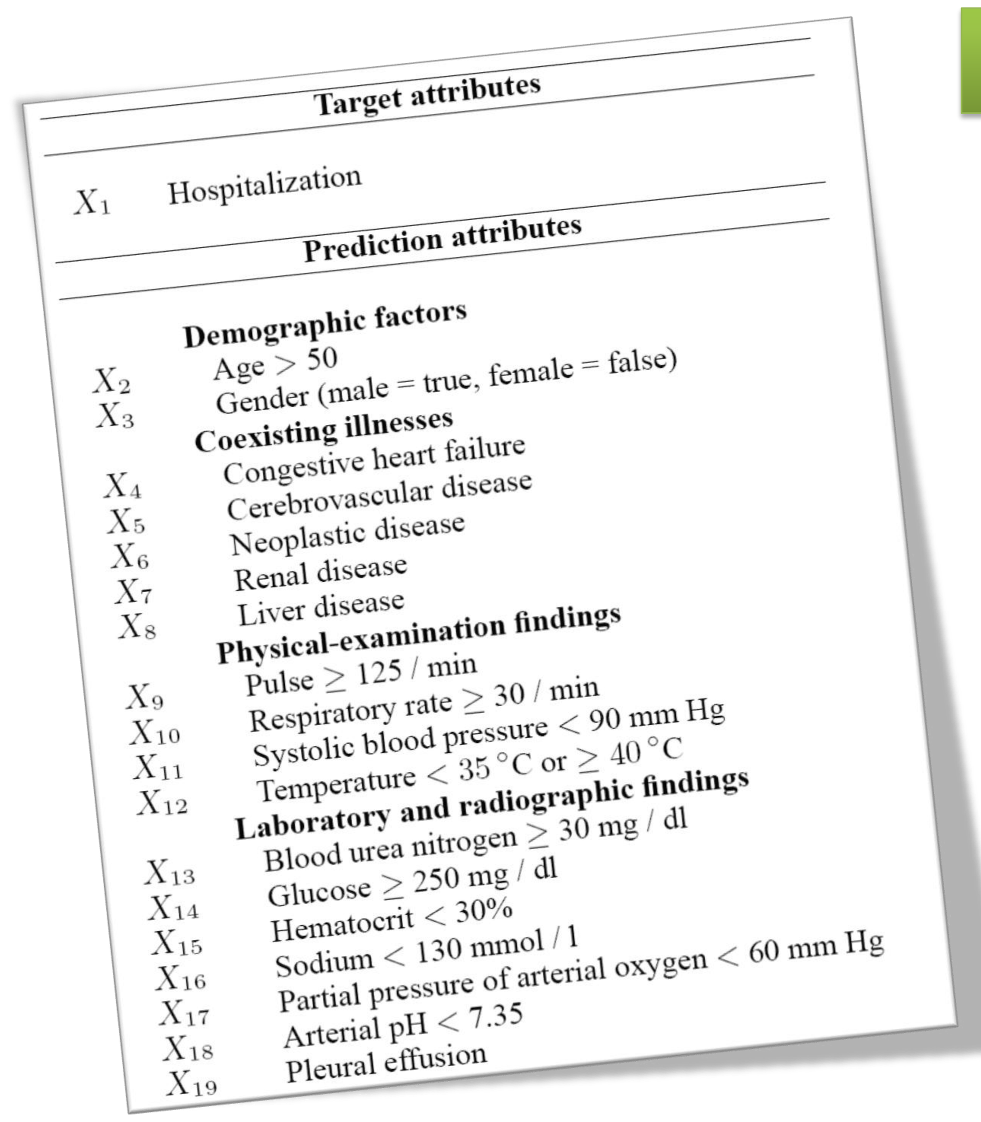
CAD Framework



Main question: Given the values of context variables for the current patient are the values of the decision variables for that patient unusual?

Experiments

- PORT dataset (Kapoor 1996)
- Patients diagnosed with the community acquired **pneumonia**
- 2287 patient cases
- 19 discrete attributes
- no missing values
- 100 evaluated by the panel of three physicians
- 23 anomalies
- Goal: Detect whether the decision of hospitalization is *anomalous*



	Target attributes
X_1	Hospitalization
	Prediction attributes
	Demographic factors
X_2	Age > 50
X_3	Gender (male = true, female = false)
	Coexisting illnesses
X_4	Congestive heart failure
X_5	Cerebrovascular disease
X_6	Neoplastic disease
X_7	Renal disease
X_8	Liver disease
	Physical-examination findings
X_9	Pulse ≥ 125 / min
X_{10}	Respiratory rate ≥ 30 / min
X_{11}	Systolic blood pressure < 90 mm Hg
X_{12}	Temperature $< 35\,^\circ$C or $\geq 40\,^\circ$C
	Laboratory and radiographic findings
X_{13}	Blood urea nitrogen ≥ 30 mg / dl
X_{14}	Glucose ≥ 250 mg / dl
X_{15}	Hematocrit $< 30\%$
X_{16}	Sodium < 130 mmol / l
X_{17}	Partial pressure of arterial oxygen < 60 mm Hg
X_{18}	Arterial pH < 7.35
X_{19}	Pleural effusion

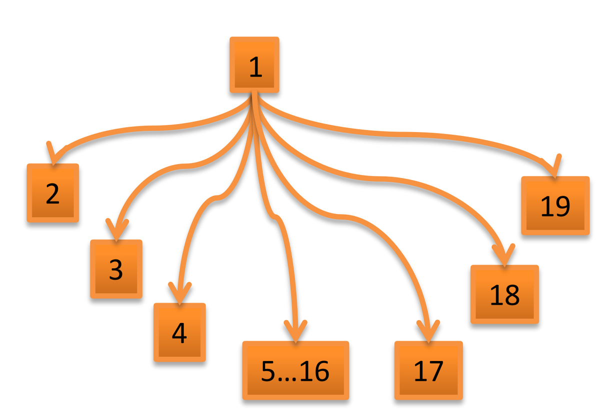
Methods

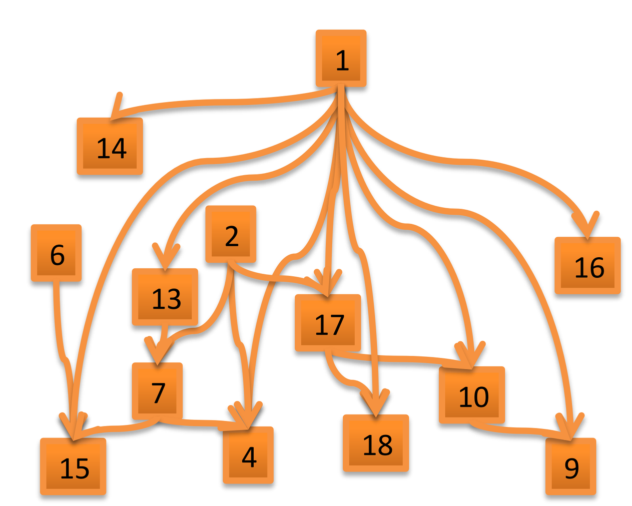
Metric:
- Standard Euclidean metric $\sqrt{\sum_i (p_i - q_i)^2}$
- Learn linear projection with Neighborhood Component Analysis (Goldberger et al . 2005) using decision as the class label

$$p_{ij} = \frac{\exp(-||Ax_i - Ax_j||^2)}{\sum_{k \neq i} \exp((-||Ax_k - Ax_j||^2)} \qquad \arg\max_A g(A) = \arg\max_A \sum_i \sum_{j \in C_i} p_{ij}$$

Patient Selection Methods:
- All patients
- *k*-closest patients with respect to the chosen metric

Probabilistic Models:
- Fixed Naïve Bayes Structure
- SoftMax model induced by the metric
- Instance Specific model: Bayesian Network from the data using Approximate Edge Marginals with MCMC (Eaton & Murphy 2007)

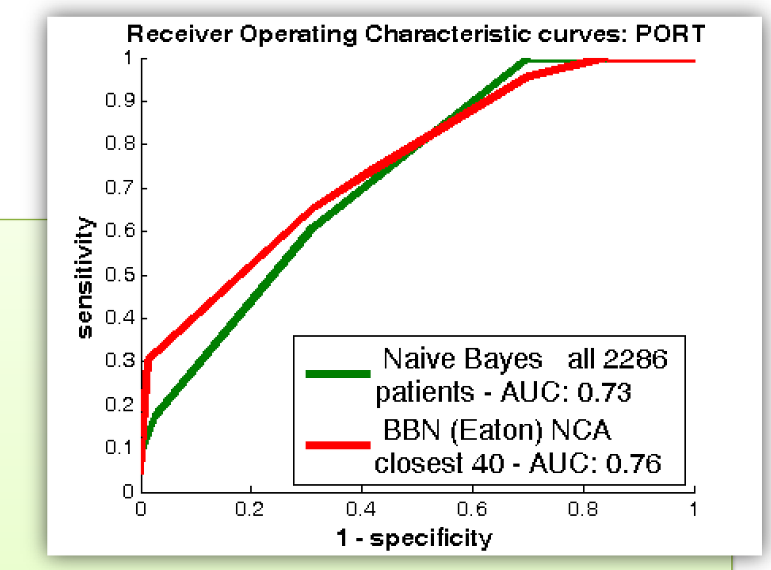




Results

STRUCTURE	METRIC	SELECTION	AUC ROC	SP > 95%
Naïve Bayes	any	ALL	72.7%	11.6%
	Euclidean	CLOSEST 40	74.6%	16.4%
	NCA	CLOSEST 40	70.0%	16.8%
SoftMax	Euclidean	ALL	76.2%	8.0%
	Euclidean	CLOSEST 40	76.2%	8.0%
	NCA	ALL	77.9%	18.0%
	NCA	CLOSEST 40	76.9%	20.2%
BBN Eaton	any	ALL	79.0%	13.8%
	Euclidean	CLOSEST 40	72.2%	17.8%
	NCA	CLOSEST 40	75.5%	26.4%

SP > 95%: AUC for ROC in acceptable range (with specificity >95%)



Discussion

- SP>95% - statistic of the interest: Hospitals will not use system with a high **false alarm rate**
- using only closer patients works better in this important ROC range
- Low number of variables opened way for exact models
- **Structure** learning improved the performance: ~50% increase
- **Instance-specific models:**
  - 1) Models can be simpler (require less examples)
  - 2) Models can be tuned to the individual patients
- **Metric** learning alleviates the effect of redundant and noisy features

Current/Future work:
- How to select the appropriate number of closest patients?
- Would learning multiple models from the different populations help?
- HIT dataset with **thousands** of records per patient
- Anomaly detection in time

- Milos Hauskrecht, Michal Valko, Branislav Kveton, Shyam Visweswaram, Gregory Cooper: **Evidence-based Anomaly Detection in Clinical Domains** in *Annual American Medical Informatics Association conference* (AMIA 2007)
- Michal Valko, Milos Hauskrecht: **Distance metric learning for conditional anomaly detection**, *Twenty-First International Florida AI Research Society Conference* (FLAIRS 2008, to appear)

University of Pittsburgh

Research funded by NLM grant R21-LM009102