

Data Driven Analytics in Healthcare: Problems, Challenges and Future Directions

Fei Wang

Research Staff Member
IBM T. J. Watson Research Center
feiwang03@gmail.com

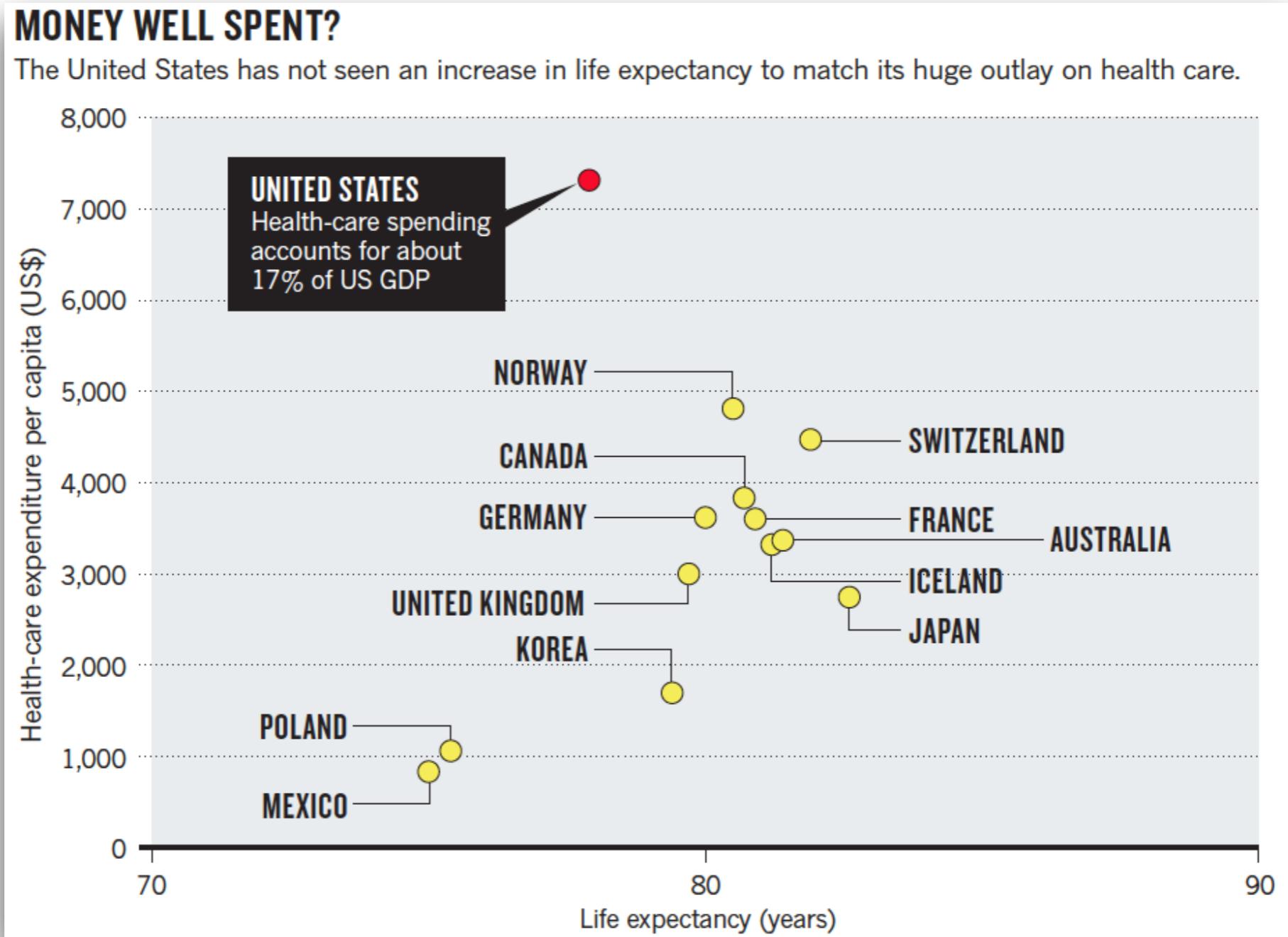
Roadmap

- Background
- Healthcare Data
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

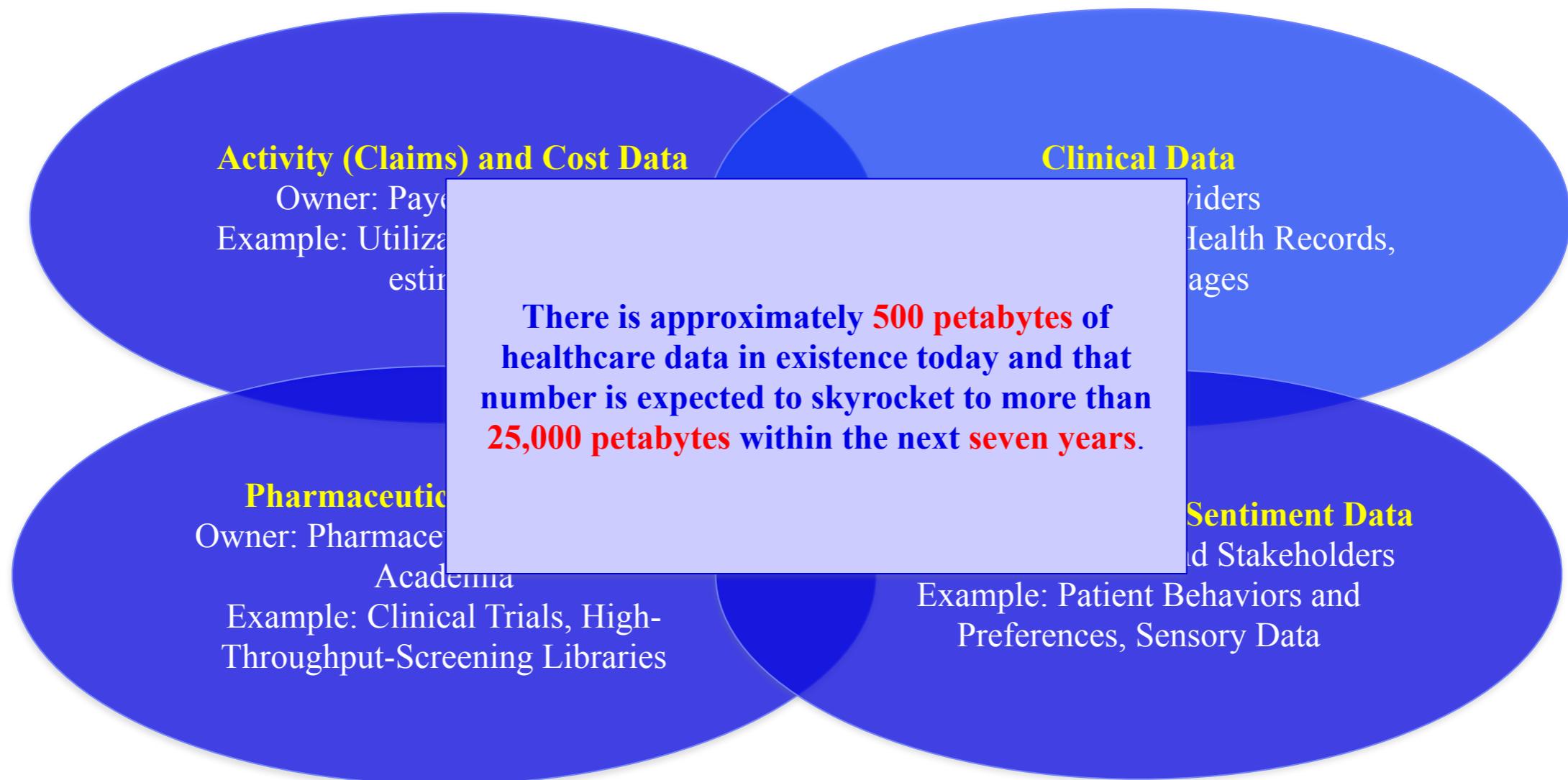
Roadmap

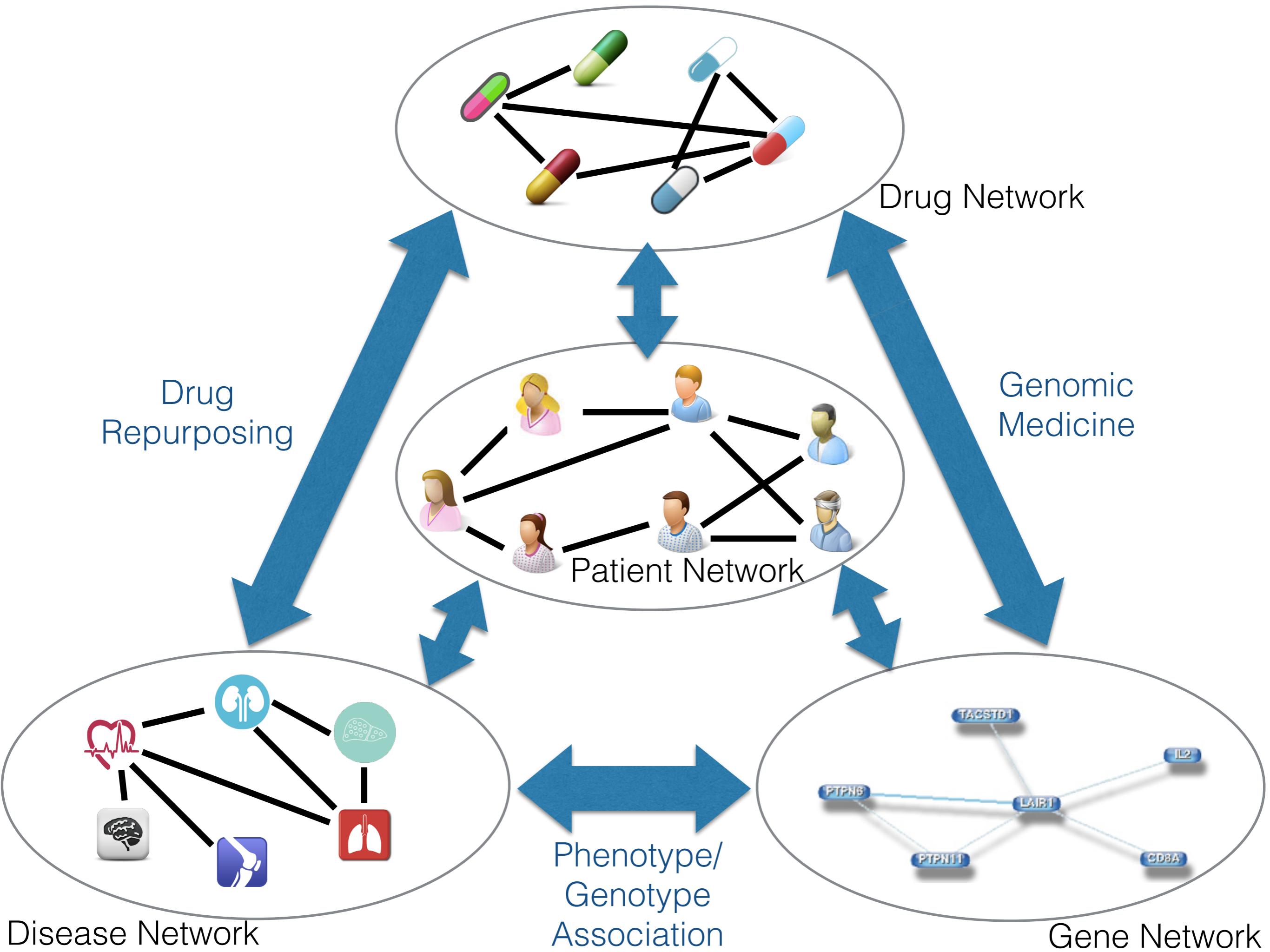
- Background
- Healthcare Data
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Healthcare Is in Crisis



Healthcare Data



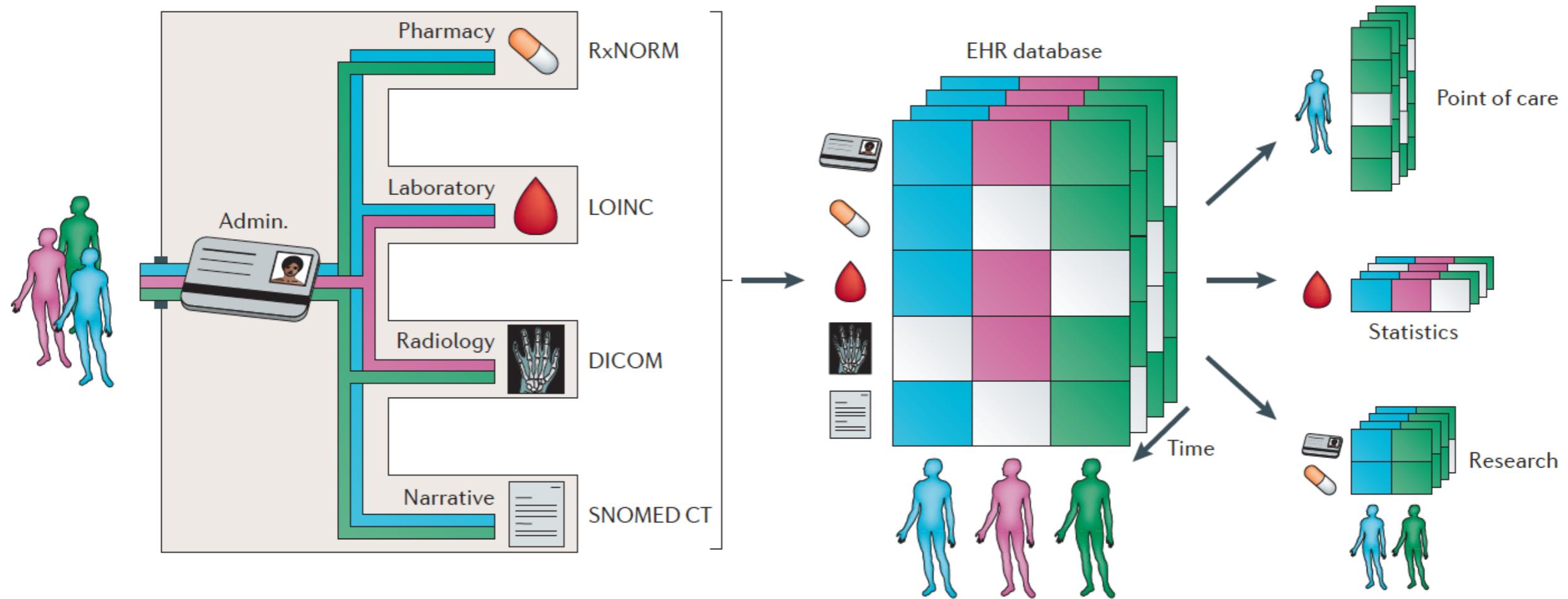


Roadmap

- Background
- Healthcare Data
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Electronic Health Records

An **Electronic Health Record** (EHR) is an evolving concept defined as a systematic collection of electronic health information about individual patients or populations



Diagnosis-ICD Codes

- ICD stands for International Classification of Diseases
- ICD is a hierarchical terminology of diseases, signs, symptoms, and procedure codes maintained by the World Health Organization (WHO)
- In US, most people use ICD-9, and the rest of world use ICD-10

(250) Diabetes mellitus

- (250.0) Diabetes mellitus without mention of complication
- (250.1) Diabetes with ketoacidosis
- (250.2) Diabetes with hyperosmolarity
- (250.3) Diabetes with other coma
- (250.4) Diabetes with renal manifestations
- (250.5) Diabetes with ophthalmic manifestations
- (250.6) Diabetes with neurological manifestations
- (250.7) Diabetes with peripheral circulatory disorders
- (250.8) Diabetes with other specified manifestations
- (250.9) Diabetes with unspecified complication

Procedure - CPT Codes

- CPT: Current Procedural Terminology (CPT)
- Describes medical, surgical, and diagnostic services and is designed to communicate uniform information about medical services and procedures among physicians, coders, patients, accreditation organizations, and payers for administrative, financial, and analytical purposes

Codes for Radiology: 70010-79999

(70010–76499) diagnostic imaging
(76506–76999) diagnostic ultrasound
(77001–77032) radiologic guidance
(77051–77059) breast mammography
(77071–77084) bone/joint studies
(77261–77799) radiation oncology
(78000–79999) nuclear medicine

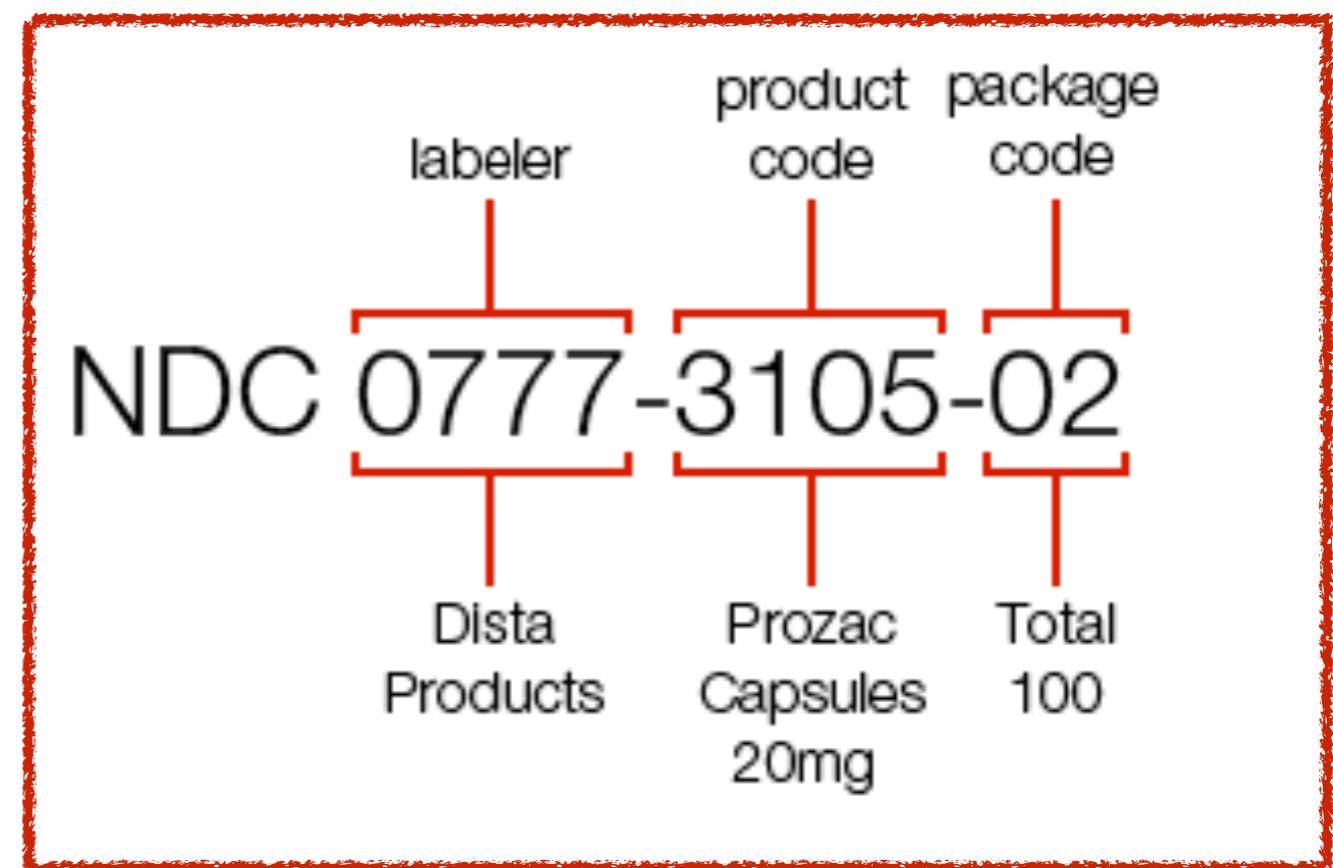
Lab - LOINC Codes

- **Logical Observation Identifiers Names and Codes (LOINC) is a database and universal standard for identifying medical laboratory observations**
- Component- what is measured, evaluated, or observed (example: urea,...)
- Kind of property- characteristics of what is measured, such as length, mass, volume, time stamp and so on
- Time aspect- interval of time over which the observation or measurement was made
- System- context or specimen type within which the observation was made (example: blood, urine,...)
- Type of scale- the scale of measure. The scale may be quantitative, ordinal, nominal or narrative
- Type of method- procedure used to make the measurement or observation

13362-9	Collection duration:Time:*:Urine:Qn:
35663-4	Protein:MCnc:XXX:Urine:Qn:
26801-1	Protein:MRat:12H:Urine:Qn:
19153-6	Specimen volume:Vol:XXX:Urine:Qn:
46952-8	Fluticasone propionate:MCnc:Pt:Urine:Qn:
6765-2	17-Hydroxypregnolone:MCnc:Pt:Ser/Plas:Qn:
19139-5	Pathologist name:PN:Pt:Provider:Nom:
22638-1	Path report.comments:Imp:Pt:Specimen:Nar:
19139-5	Pathologist name:ID:Pt:Provider:Nom:
33511-7	Appearance:Aper:Pt:XXX:Nom:
35265-8	Path report.addendum:Find:Pt:Specimen:Nar:
46608-6	Primary referring physician ID:ID:Pt:Provider:Nom:
74221-3	Referring physician address:Addr:Pt:Provider:Nom

Medication - NDC Codes

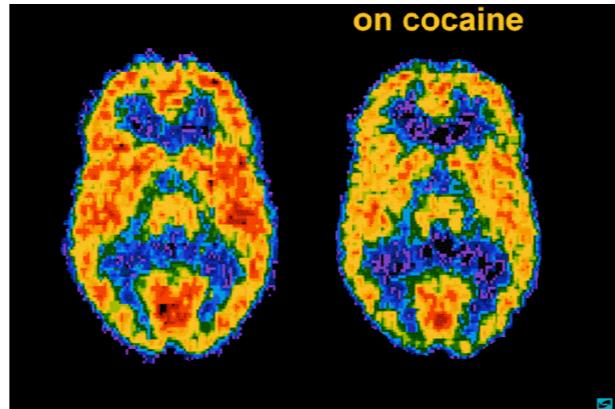
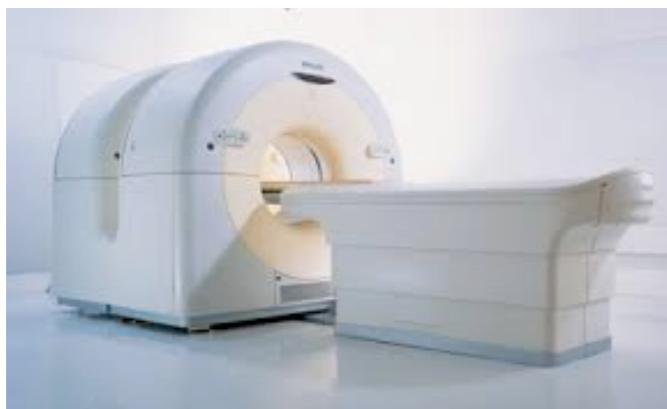
- **The National Drug Code (NDC) is a unique product identifier used in the United States for drugs intended for human use**
- 10-digit, 3-segment numeric identifier
 - The first segment, the labeler code, is 4 or 5 digits long and assigned by the Food and Drug Administration (FDA) upon submission of a Labeler Code Request. A labeler is any firm that manufactures, repacks or distributes a drug product.
 - The second segment, the product code, is 3 or 4 digits long and identifies a specific strength, dosage form, and formulation for a particular firm.
 - The third segment, the package code, is 1 or 2 digits long and identifies package forms and sizes.



Medical Imaging



**X-ray Computed
Tomography (CT)**

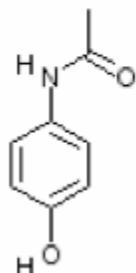


**Positron Emission
Tomography (PET)**

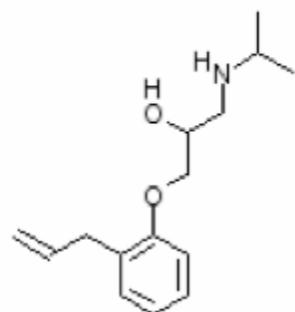


**Magnetic Resonance
Imaging (MRI)**

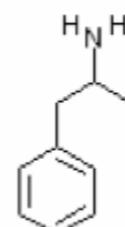
Drug: Chemical Compounds



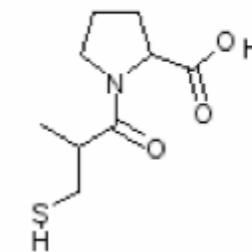
Acetaminophen



Alprenolol



Amphetamine

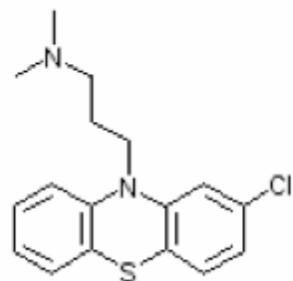


Captopril

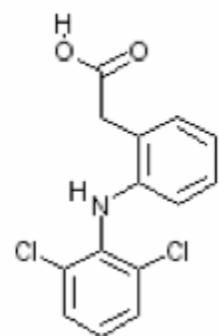
PubChem



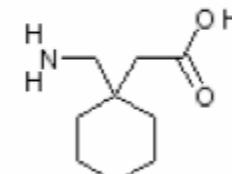
Go Advanced Search



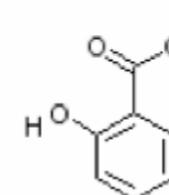
Chlorpromazine



Diclofenac



Gabapentin

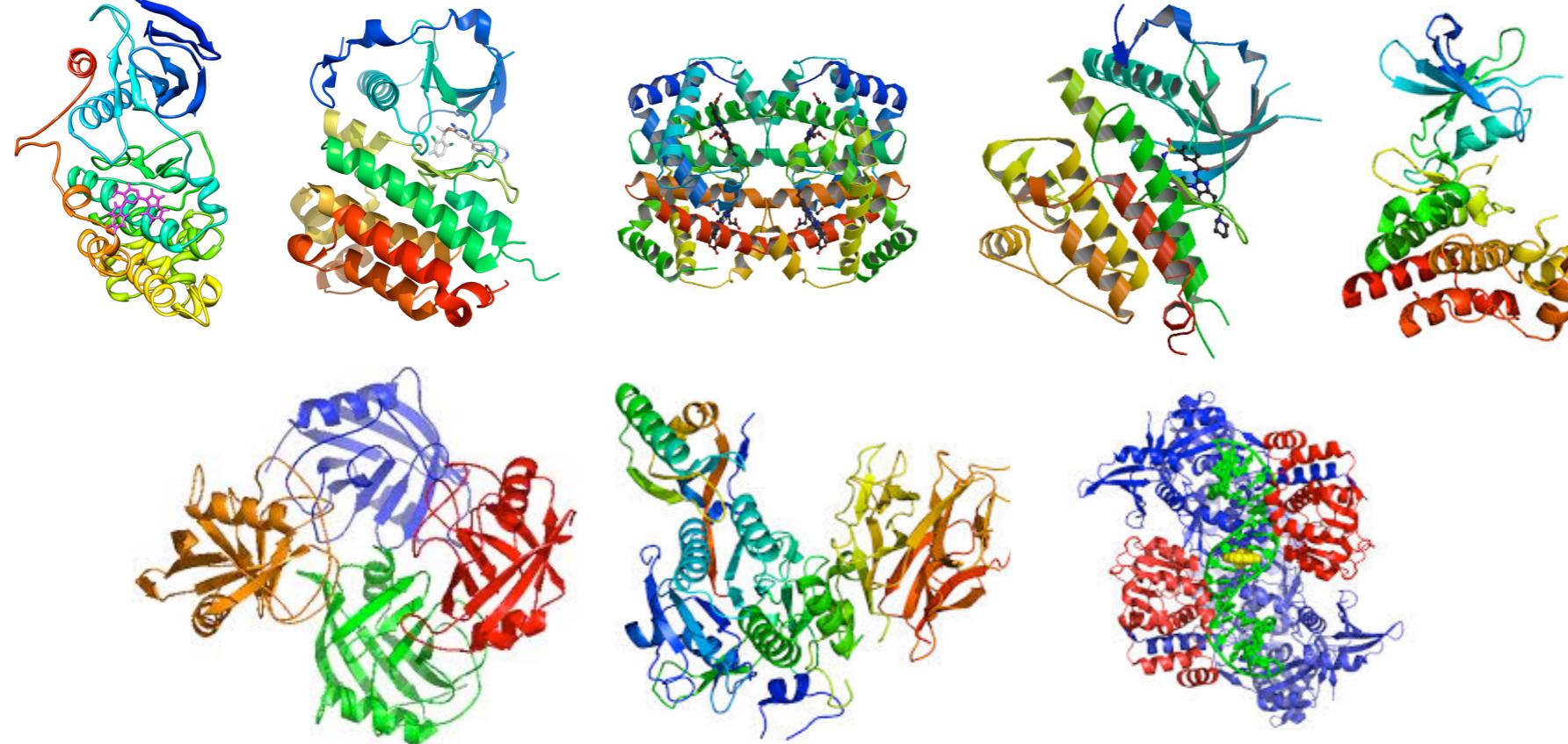


Salicylate



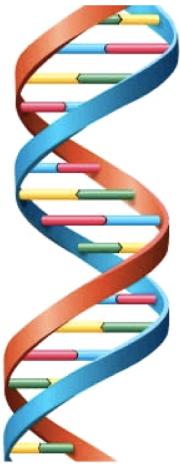
A chemical compound is a pure chemical substance consisting of two or more different chemical elements that can be separated into simpler substances by chemical reactions

Drug: Protein Targets

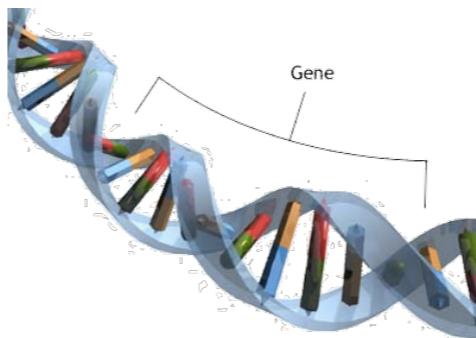


The term biological target is frequently used in pharmaceutical research to describe the native protein in the body whose activity is modified by a drug resulting in a desirable therapeutic effect. In this context, the biological target is often referred to as a drug target.

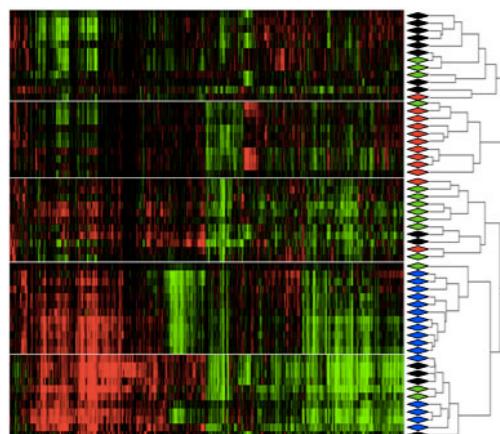
Gene



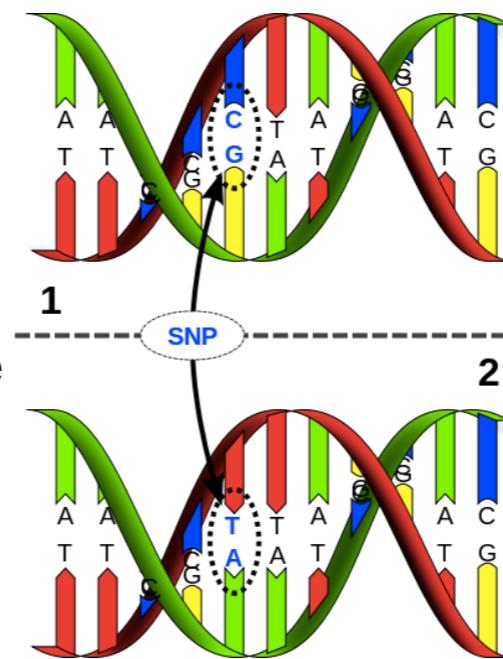
DNA: A long molecule that looks like a twisted ladder. It is made of four types of simple units and the sequence of these units carries information, just as the sequence of letters carries information on a page.



Gene: A segment of DNA. Genes are like sentences made of the "letters" of the nucleotide alphabet, between them genes direct the physical development and behavior of an organism. Genes are like a recipe or instruction book, providing information that an organism needs so it can build or do something - like making an eye or a leg, or repairing a wound.



Gene expression: The process in which the information encoded in a gene is converted into a form useful for the cell. The first step is transcription, which produces a messenger RNA molecule complementary to the DNA molecule on which a gene is encoded. For protein-coding genes, the second step is translation, in which the messenger RNA is read by the ribosome to produce a protein.

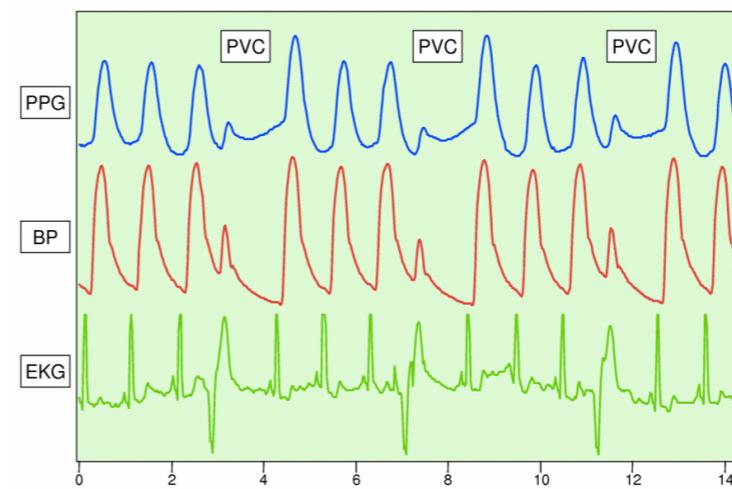


A **Single Nucleotide Polymorphism** (SNP, pronounced snip; plural snips) is a DNA sequence variation occurring commonly within a population (e.g. 1%) in which a single nucleotide — A, T, C or G — in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes.

Physiology



Physiology is the scientific study of function in living systems. A sub-discipline of biology, its focus is in how organisms, organ systems, organs, cells, and bio-molecules carry out the chemical or physical functions that exist in a living system



Vector Based Representation

A collection of numbers

Row Vector
 $\mathbf{v} = [v_1, v_2, \dots, v_d]$
Dimensionality

Transpose
 $v' = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_d \end{bmatrix}$

Patient Diagnosis Vector:
d is the number of distinct diagnosis code, x_i represents the frequency of the i-th diagnosis code in his/her historical records

Integer

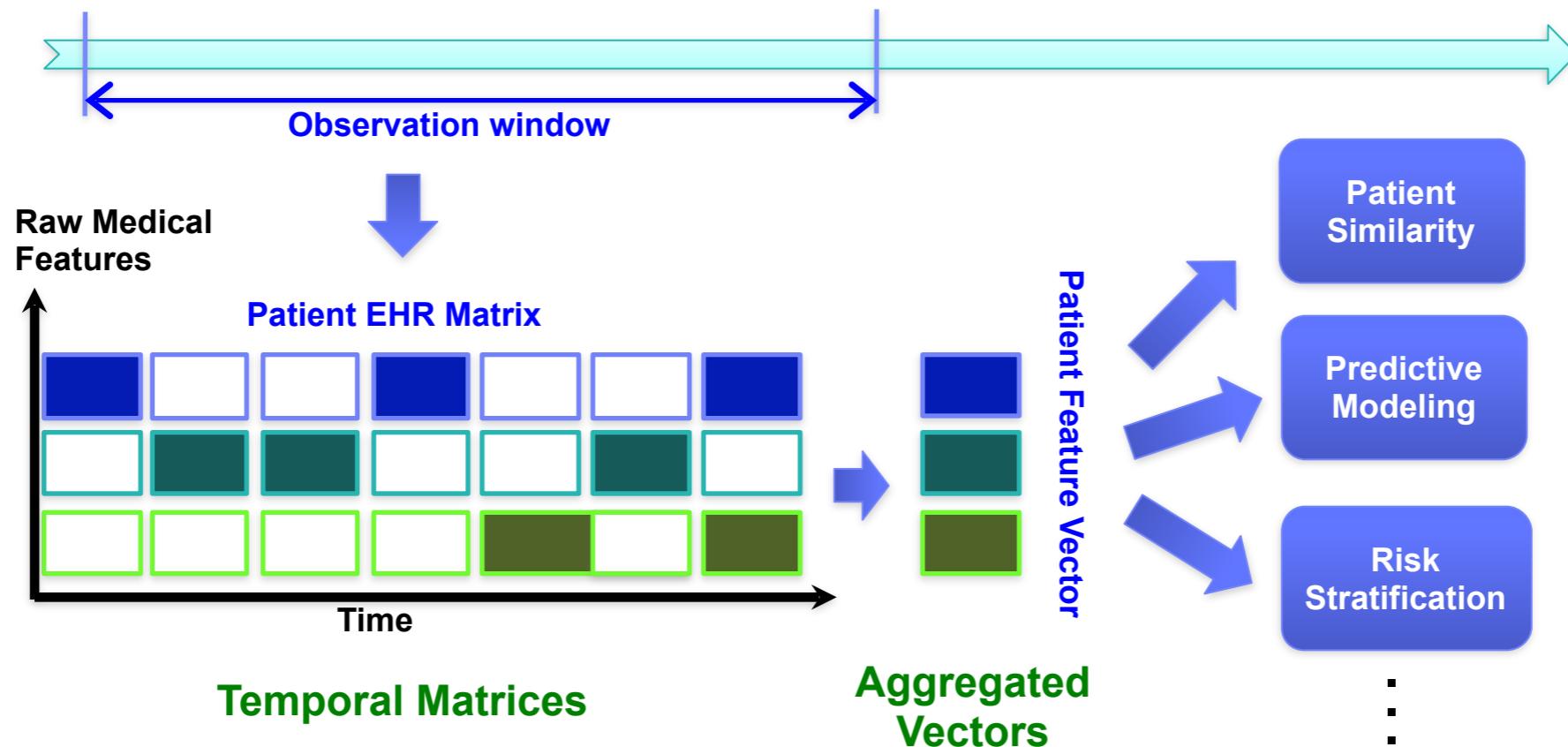
Drug Compound Vector:
d is the number of distinct chemical compounds, x_i represents whether or not the i-th compound appears

Binary

Gene Expression Vector:
d is the number of distinct samples, x_i represents the gene expression value on the i-th sample

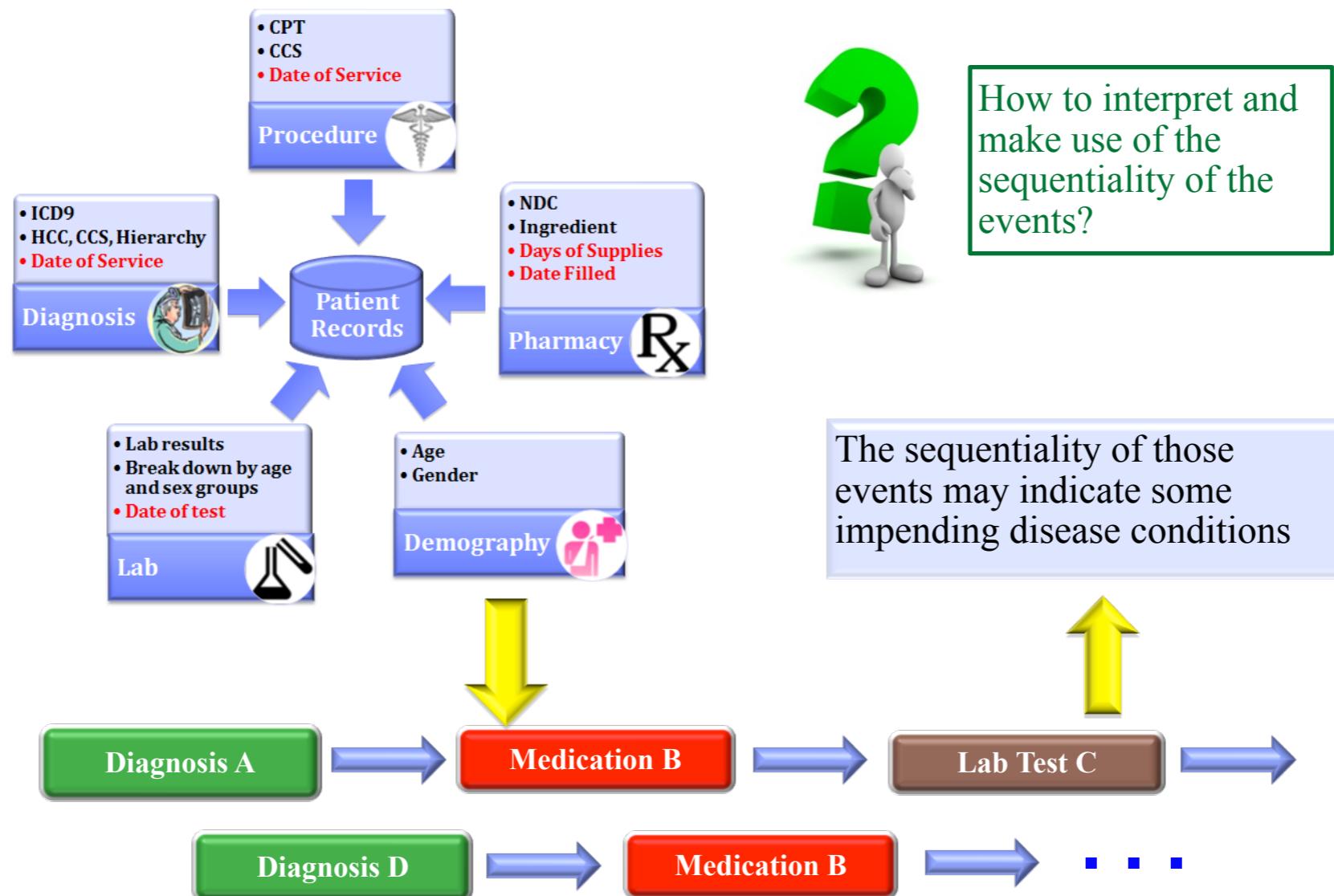
Continuous

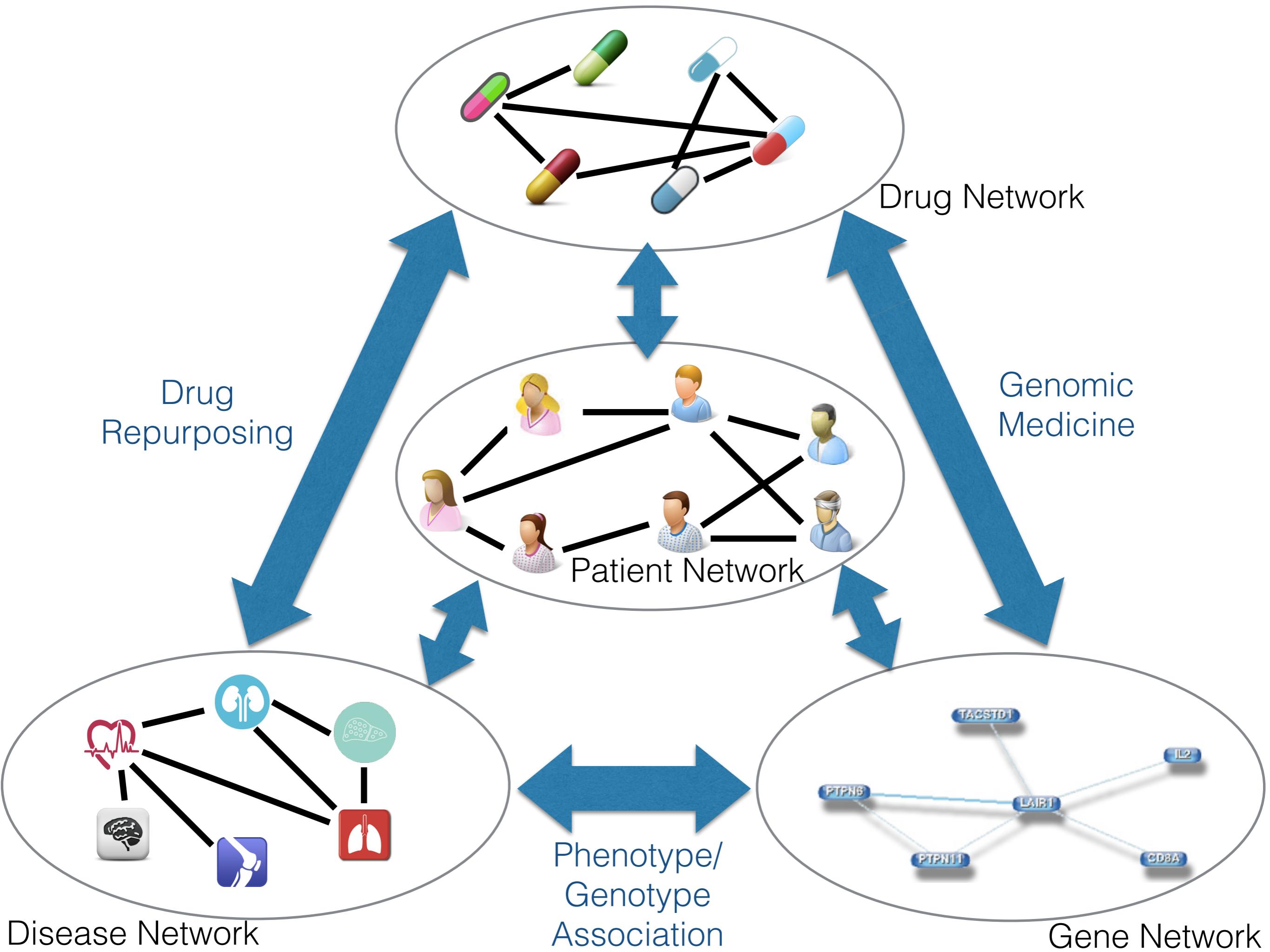
Matrix Based Representation



- Jimeng Sun, Fei Wang, Jianying Hu, Shahram Edabollahi: Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 14(1): 16-24 (2012)
- Fei Wang, Jimeng Sun, Shahram Ebadollahi: Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. Statistical Analysis and Data Mining 5(1): 54-69 (2012)
- J.Wu, J. Roy,W. F. Stewart, Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches, Medical Care 48 S106–S113 (2010)

Sequence Based Representation



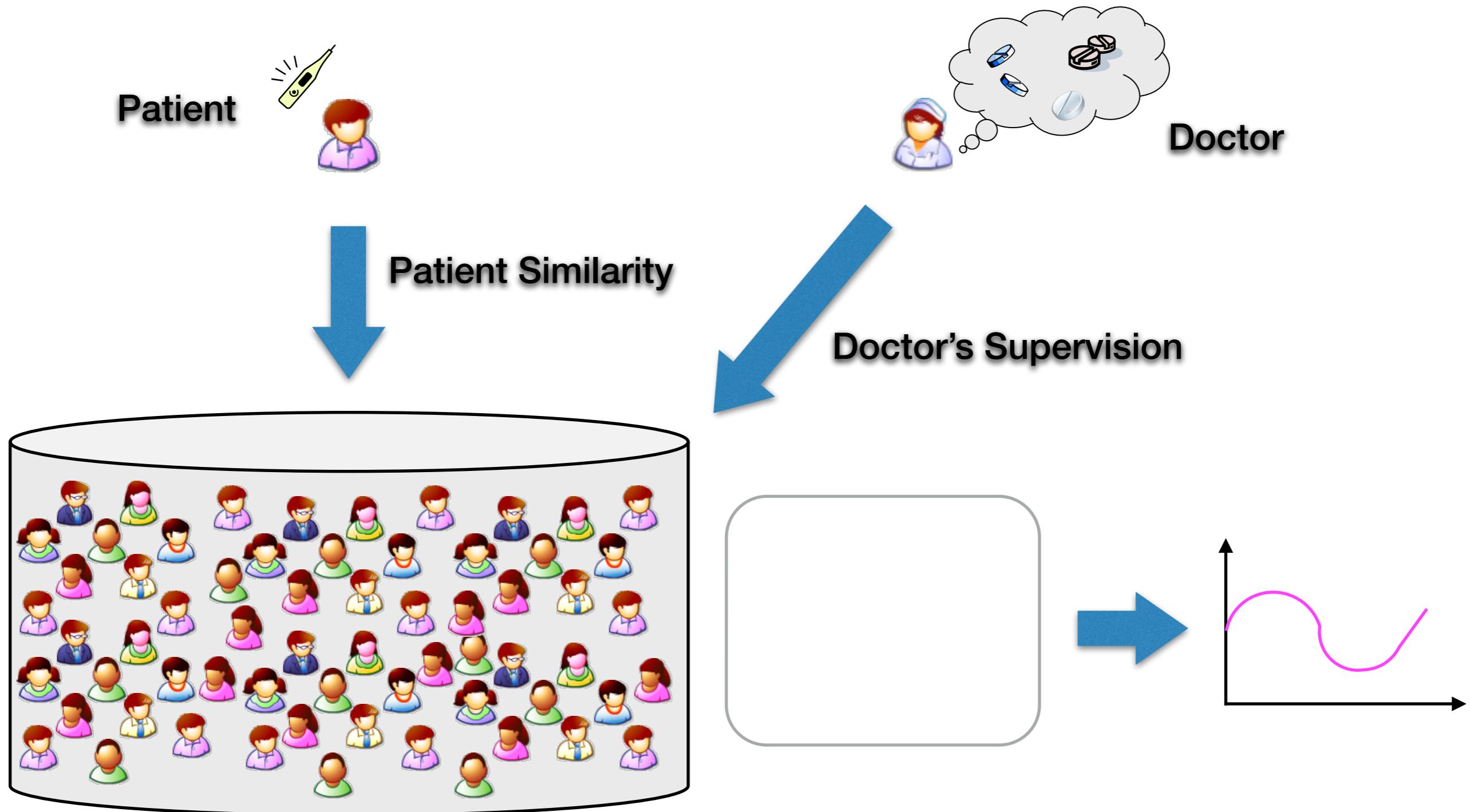


Roadmap

- Background
- Healthcare Data
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

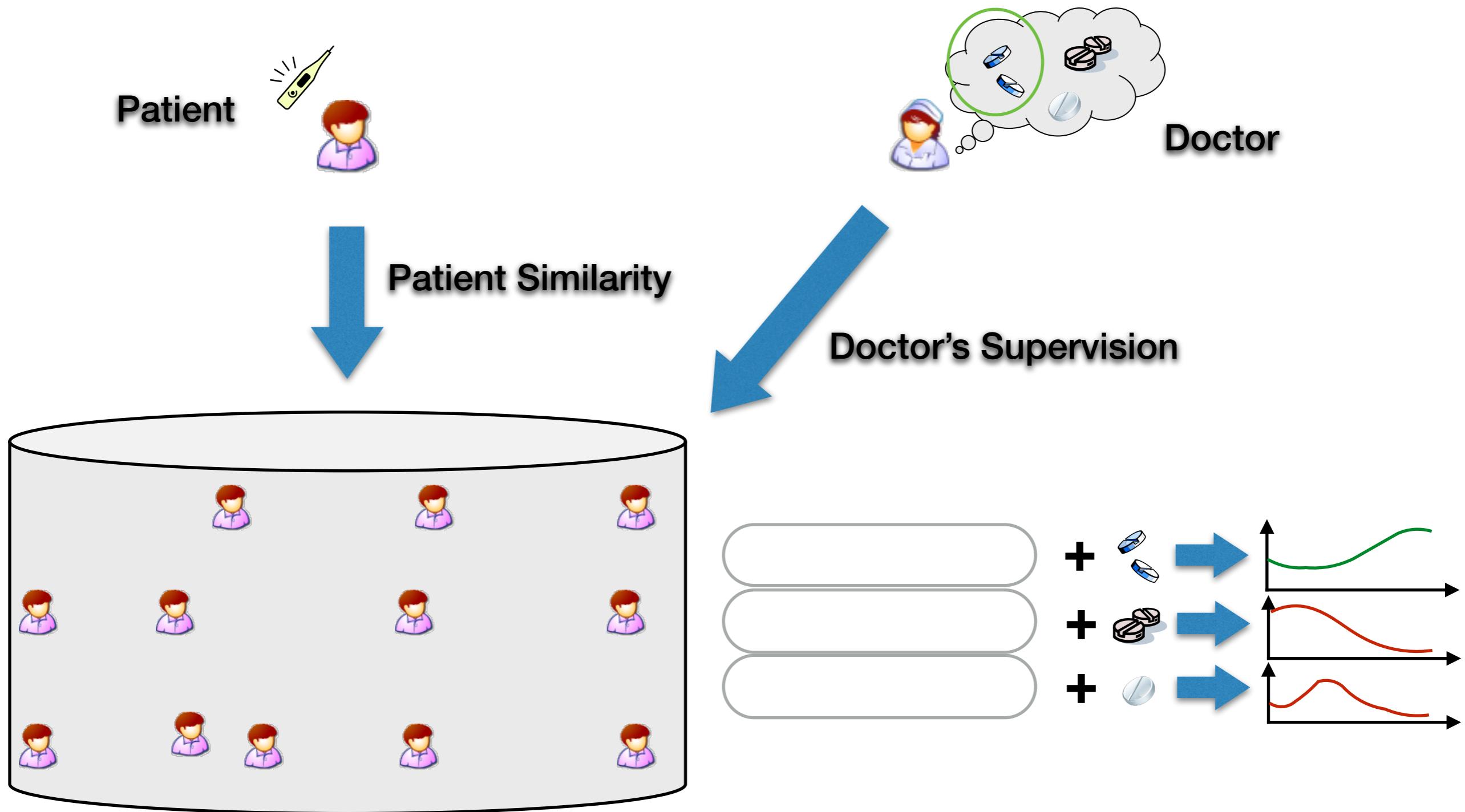
Patient Similarity Problem

Prognostication/Outcome Analysis

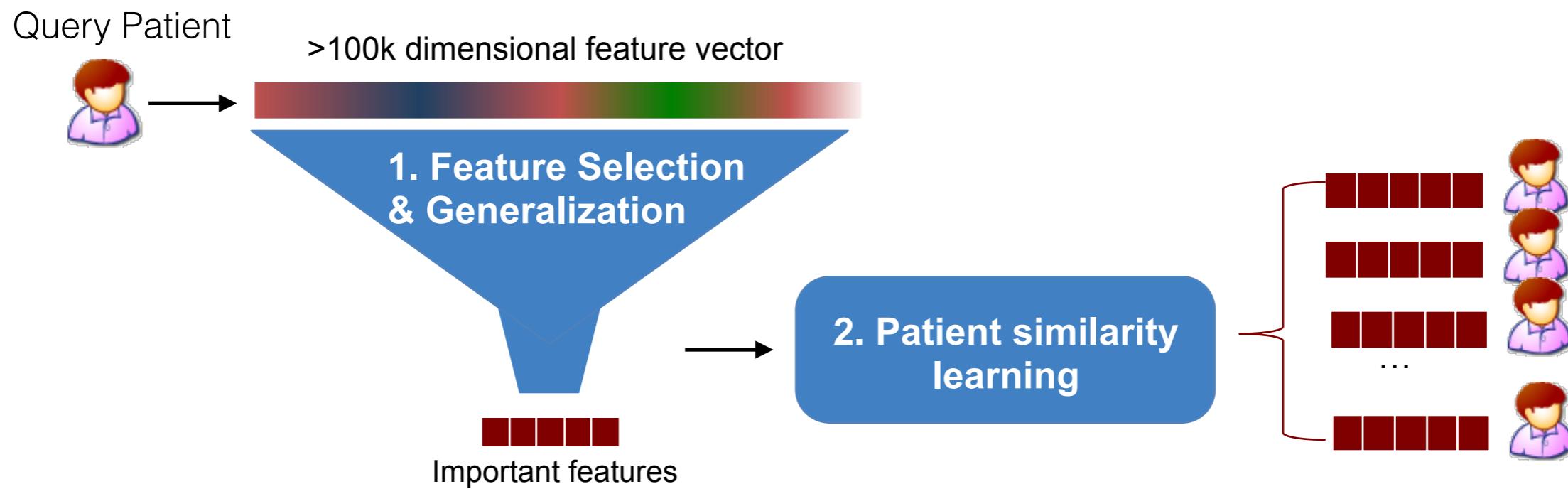


Patient Similarity Problem

Personalized Treatment Recommendation

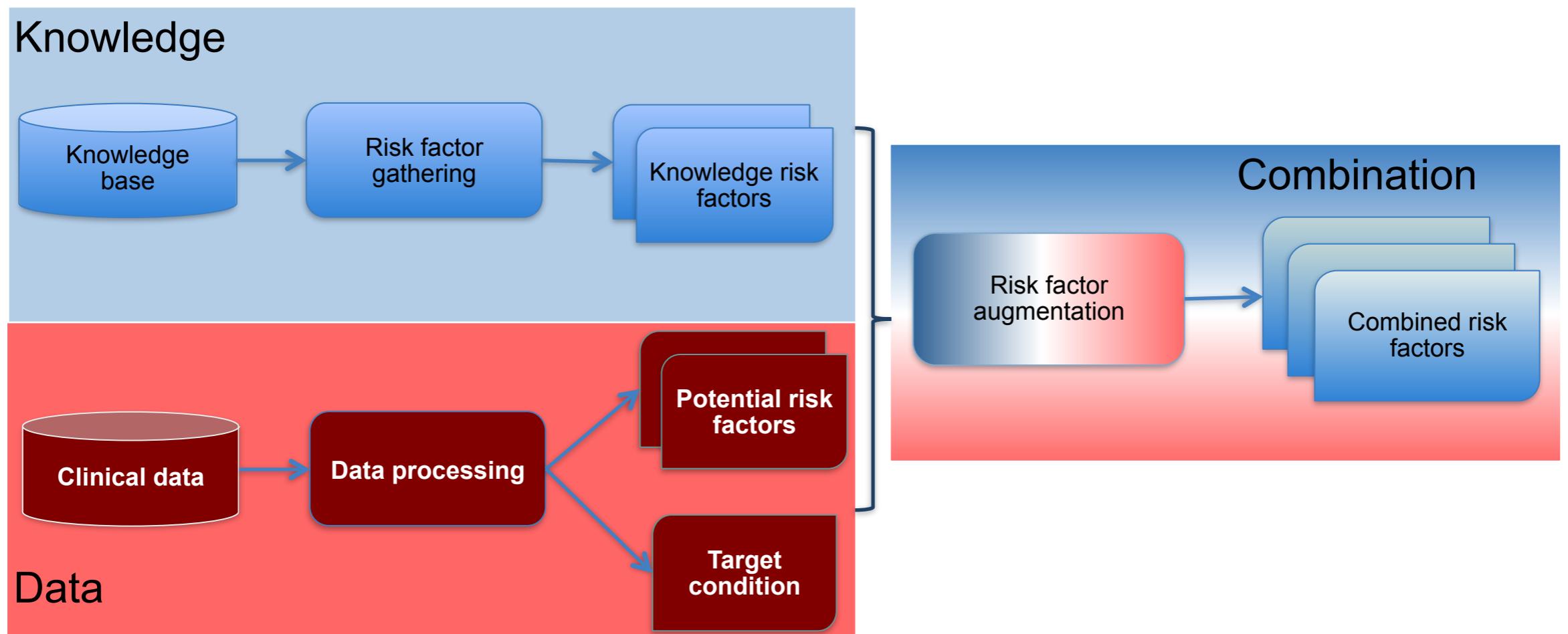


Our Approach



- For a clinical context,
 1. What are important features?
 2. What is the right similarity measure?

Feature Selection



Jimeng Sun, Jianying Hu, Dijun Luo, Marianthi Markatou, Fei Wang, Shahram Ebadollahi, Steven E. Steinhubl, Zahra Daar, Walter F. Stewart. Combining Knowledge and Data Driven Insights for Identifying Risk Factors using Electronic Health Records. AMIA2012.

Dijun Luo, Fei Wang, Jimeng Sun, Marianthi Markatou, Jianying Hu, Shahram Ebadollahi, SOR: Scalable Orthogonal Regression for Low-Redundancy Feature Selection and its Healthcare Applications. SDM'12

Scalable Orthogonal Regression

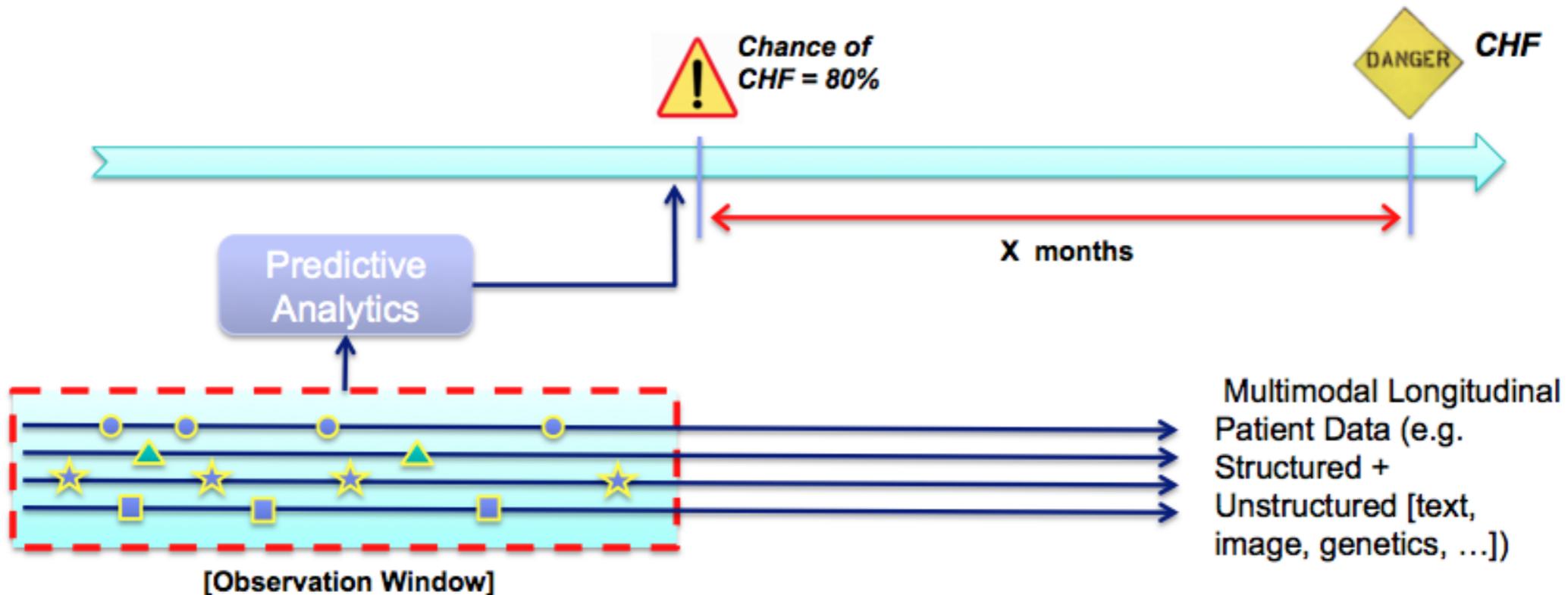
- Incorporating knowledge driven risk factors
- Accurate prediction
- Minimal redundancy:
 - Little correlation between the selected data driven risk factors and existing knowledge driven risk factor
 - Little correlation among the additional risk factors from data, to further ensure quality of the additional factors

$$f(\alpha) = \boxed{\mathcal{L}(\mathbf{y}, \mathbf{X}\alpha)} + \frac{\beta}{4} \left[\sum_{ij \in \mathcal{D}} (\alpha_i \mathbf{x}_i^T \mathbf{x}_j \alpha_j)^2 + \sum_{i \in \mathcal{D}, j \in \mathcal{K}} (\alpha_i \mathbf{x}_i^T \mathbf{x}_j \alpha_j)^2 \right] + \lambda \|\alpha\|_1$$

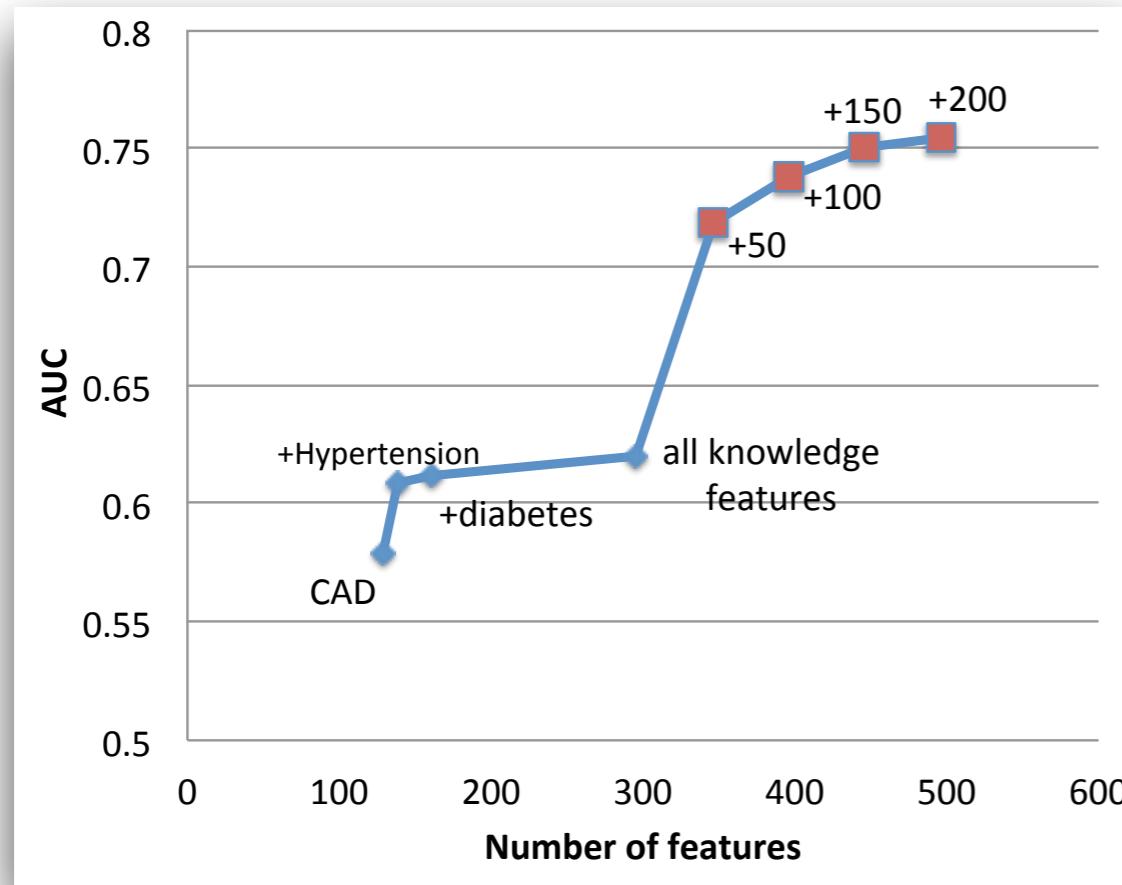
Model Accuracy Correlation between data-driven features Correlation between data- and knowledge-driven features Sparsity Penalty

Early Detection of CHF

- **Goal:**
 - Build a model for predicting HF onset x months before the HF diagnosis
- **Data: Longitudinal patient records**
 - Structured data:
 - Demographics, Outpatient diagnoses, Problem List , Vitals, Medication, Labs
 - Unstructured text : encounter notes
- **Challenge faced by our clinical partners:**
 - How to systematically collect and evaluate many weak and non-specific indicators and identify the ones that combined are truly predictive



CHF Prediction Results



- 4644 case patients, 45,981 control patients
- Over 20k features of different types (diagnoses, demographics, Framingham symptoms, lab results, medication, vital)
- Novel feature selection algorithm enabling integration of knowledge driven and data driven risk factors
- Investigation of different observation windows (30 – 900 days) and prediction windows (1 – 720 days)
- Investigation of multiple classification models (logistic regression, random forest, kNN, cox regression ...)

- AUC significantly improves as complementary data driven risk factors are added into existing knowledge based risk factors.
- A significant AUC increase occurs when we add first 50 data driven features

Announcement (10/9/13)
NIH grant on early CHF detection

\$2 Million Awarded to Sutter Health, IBM and Geisinger Health System to Study Heart Failure Prediction

Three-year collaboration will develop groundbreaking, big data analytic methods to improve care and reduce costs for treatment of heart disease

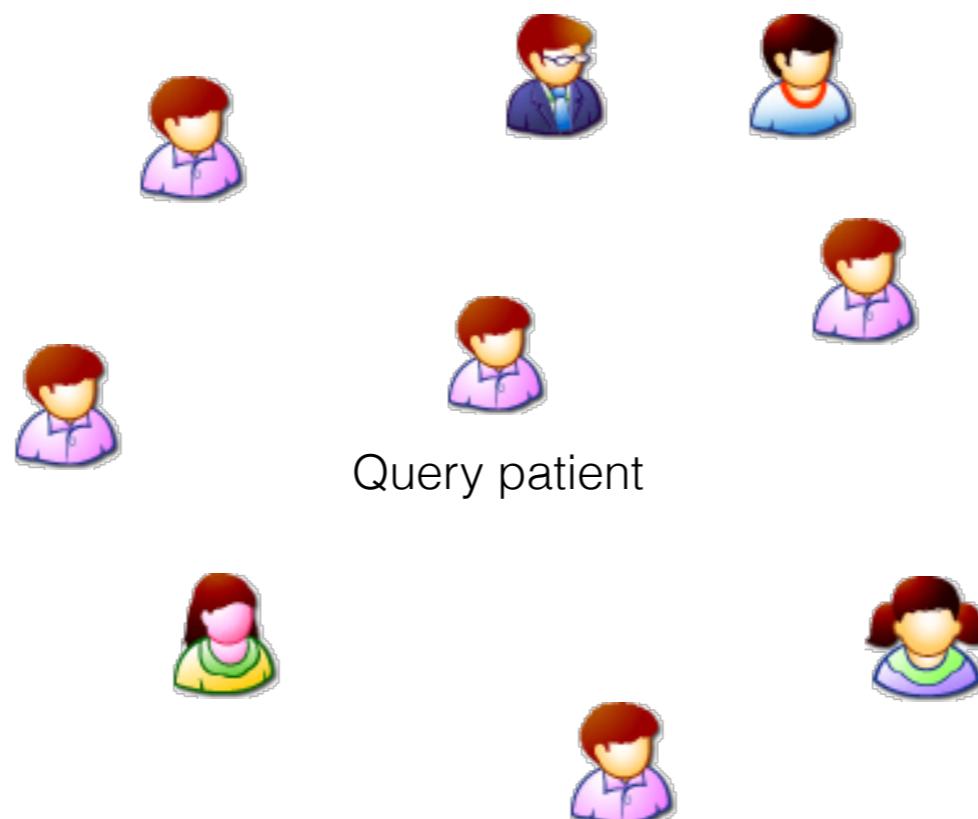
Top 10 Features

Category	Feature	Relevancy to HF
Diagnosis	Dyslipidemia	✓
Medication	Thiazides-like Diuretics	✓
Lab	Antihypertensive Combinations	✓
Symptom	Aminopenicillins	✓
	Bone density regulators	✗
	Natriuretic Peptide	✓
	Rales	✓
	Diuretic Combinations	✓
	S3Gallop	✓
	NSAIDS	✓

- 9 out of 10 are considered relevant to HF
- The data driven features are complementary to the existing knowledge-driven features

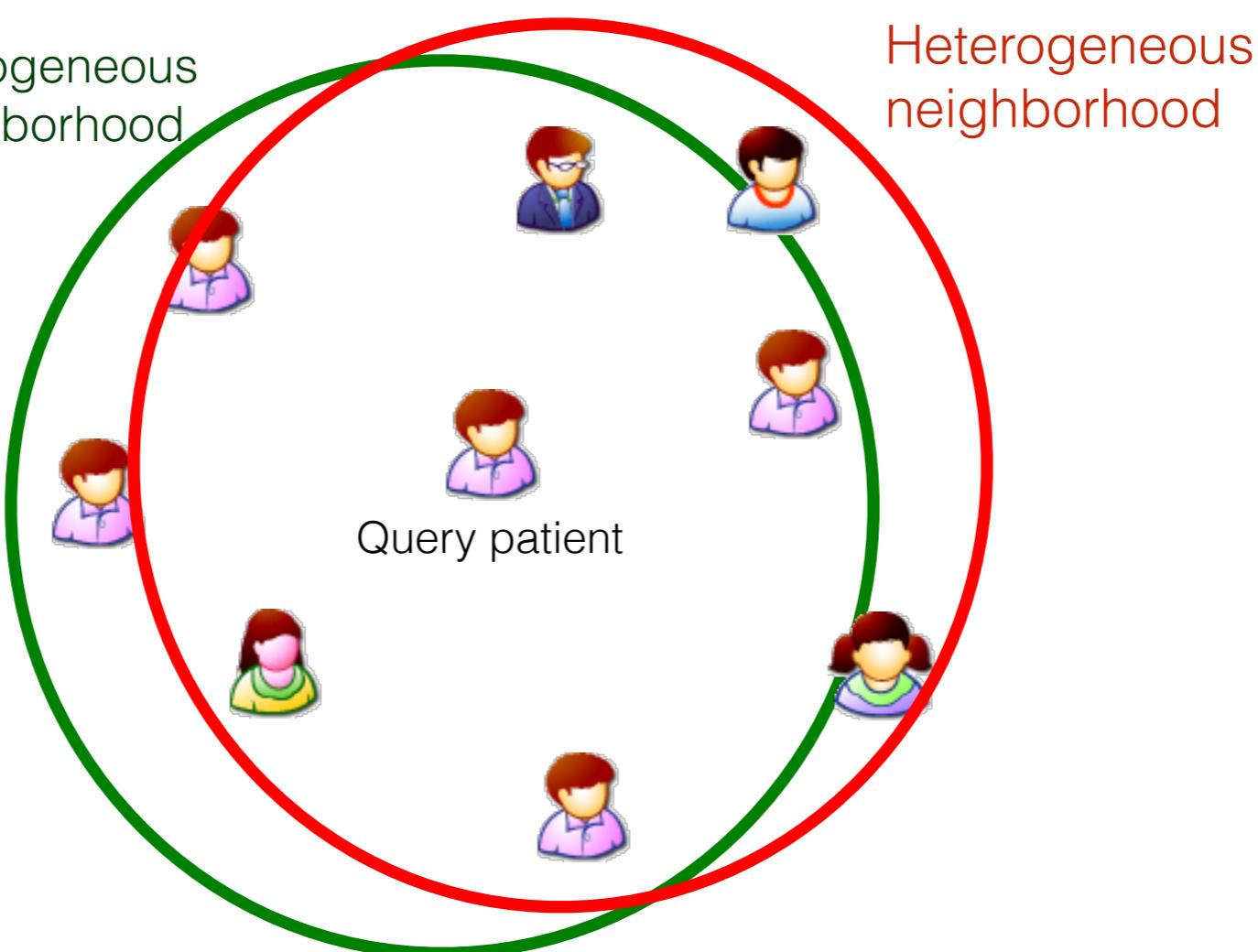
Patient Similarity

Under a specific clinical context



Patient Similarity

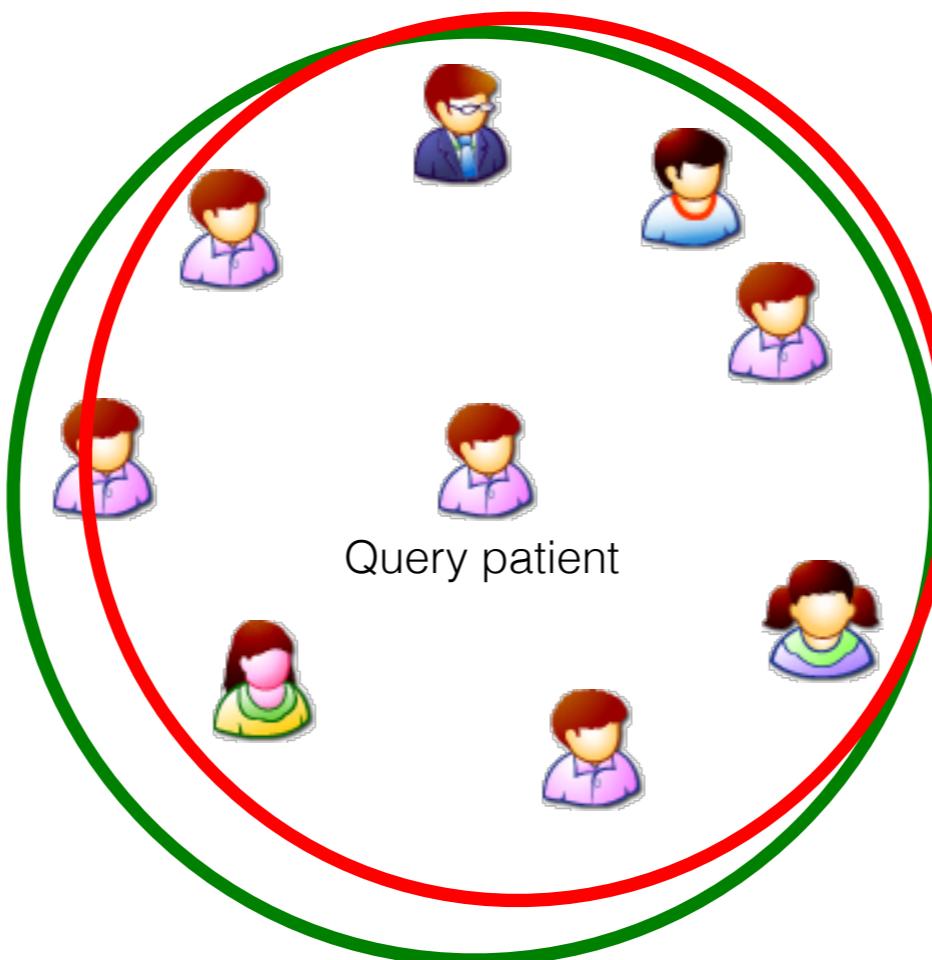
Under a specific clinical context



- Homogeneous neighbors: true positives
- Heterogeneous neighbors: false positives

Patient Similarity

Under a specific clinical context



- Shrink homogeneous neighborhood
- Grow heterogeneous neighborhood

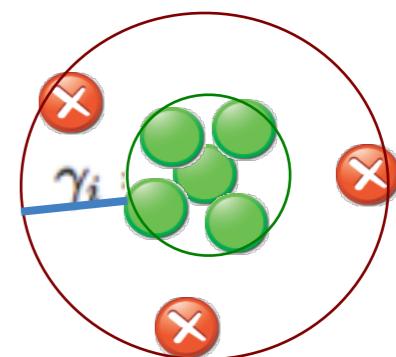
Locally Supervised Metric Learning

Goal: Learn a generalized Mahalanobis distance for a specific clinical context (target label)

$$d_{\Sigma}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\top} \Sigma (\mathbf{x}_i - \mathbf{x}_j)} \quad \Sigma = \mathbf{W}\mathbf{W}^{\top}$$

Margin for \mathbf{x}_i

$$\gamma_i = \underbrace{\sum_{k: \mathbf{x}_k \in \mathcal{N}_i^e} \|\mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_k\|^2}_{\text{Total distance to heterogeneous neighbors}} - \underbrace{\sum_{j: \mathbf{x}_j \in \mathcal{N}_i^o} \|\mathbf{W}^T \mathbf{x}_i - \mathbf{W}^T \mathbf{x}_j\|^2}_{\text{Total distance to homogeneous neighbors}}$$



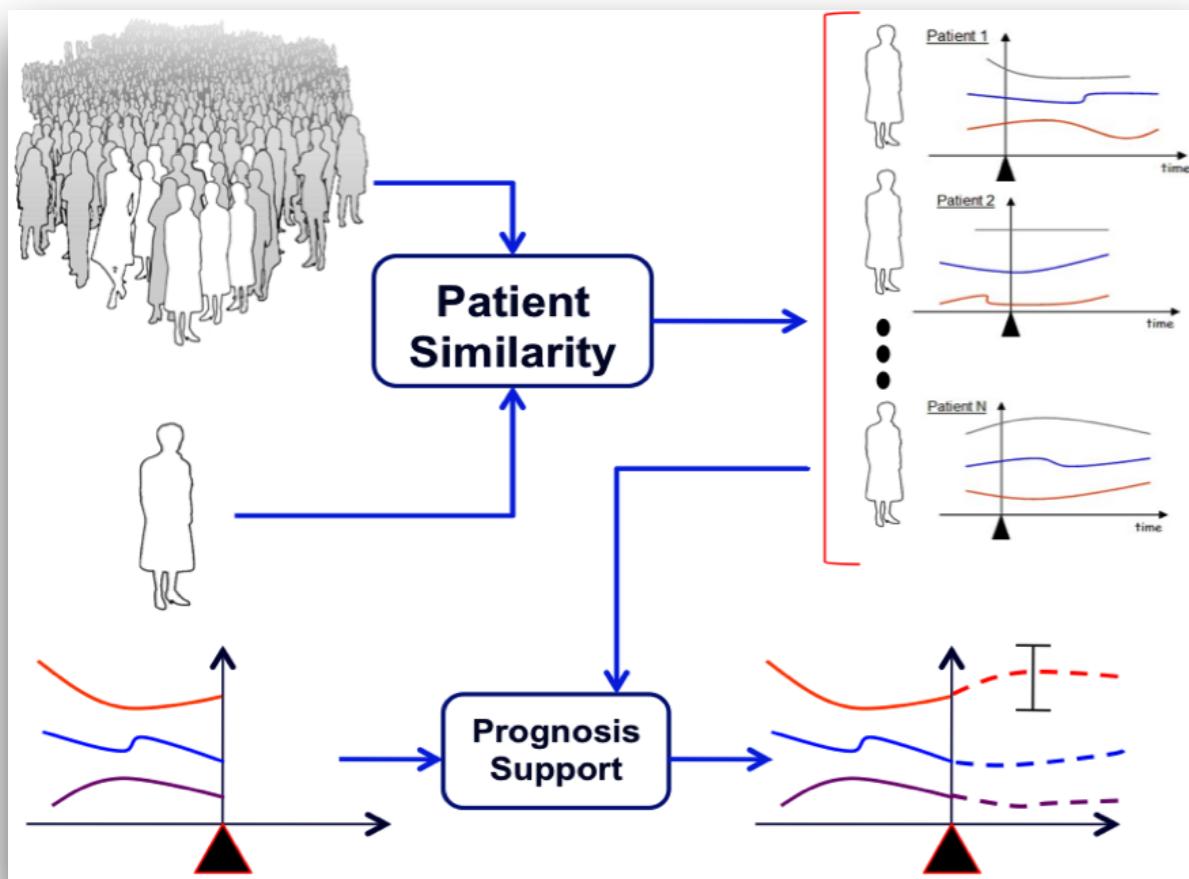
Maximize the total margin

\mathcal{N}_i^o Homogeneous neighborhood for \mathbf{x}_i

\mathcal{N}_i^e Heterogeneous neighborhood for \mathbf{x}_i

$$\begin{aligned} \gamma = & \sum_i \sum_{k: \mathbf{x}_k \in \mathcal{N}_i^e} \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_k) (\mathbf{x}_i - \mathbf{x}_k)^T \mathbf{W} \\ & - \sum_i \sum_{j: \mathbf{x}_j \in \mathcal{N}_i^o} \mathbf{W}^T (\mathbf{x}_i - \mathbf{x}_j) (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W} \end{aligned}$$

Near-Term Prognostication in ICU



Physiological Streams

ABP, systolic ABP, diastolic ABP, SpO₂, heart rate

Patients

(H Group) 590 patients experienced at least one occurrence of Acute Hypotensive Episode (AHE),
(C Group) 910 patients did not

Data

2 hours centered around decision time T₀

(H Group) T₀ is one hour before AHE

(C Group) T₀ is randomly chosen

Results

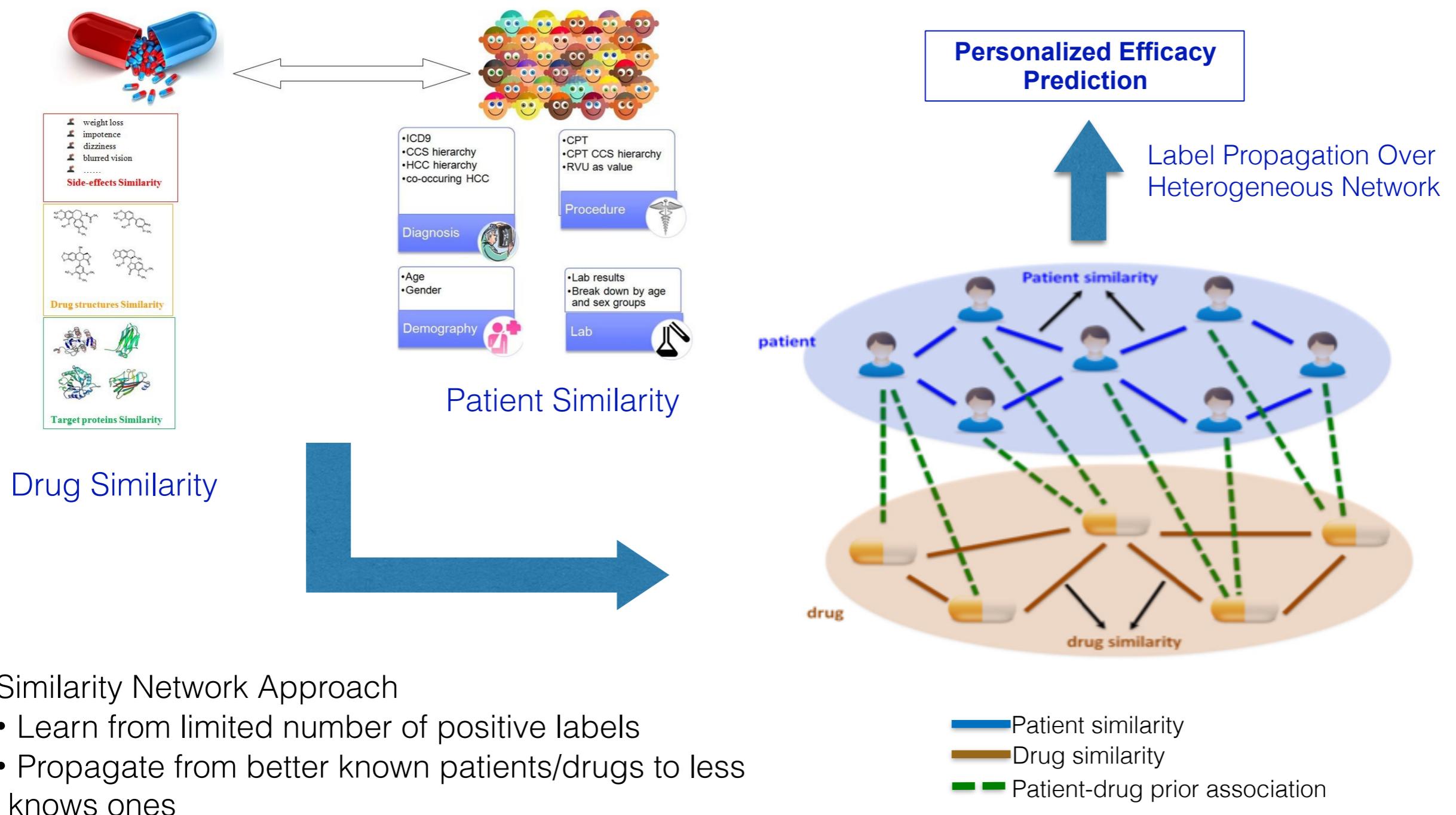
(Challenge09) 0.4982 (LSML) 0.8551

SP₀₂ measures the amount of oxygen affixed to hemoglobin cells within the circulatory system.

(Challenge09) X. Chen, D. Xu, G. Zhang, and R. Mukkamala. *Forecasting acute hypotensive episodes in intensive care patients based on peripheral arterial blood pressure waveform*. Computers in Cardiology. 2009.

(LSML) Shahram Ebadollahi, Jimeng Sun, David Gotz, Jianying Hu, Daby Sow, and Chalapathy Neti. *Predicting patient's trajectory of physiological data using temporal trends in similar patients: A system for Near-Term prognostics*. AMIA 2010. (Best Paper Finalist)

Personalized Medicine

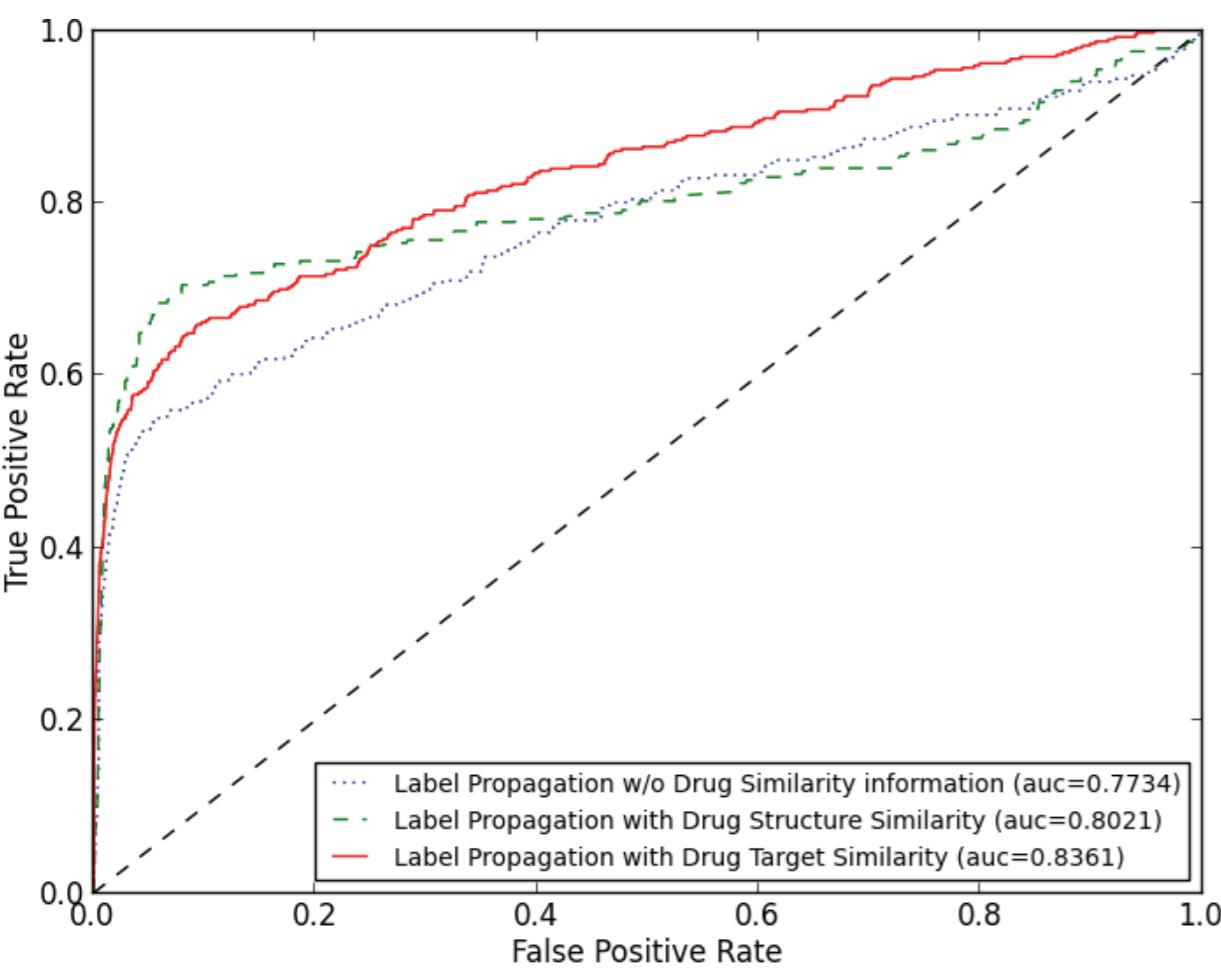


Similarity Network Approach

- Learn from limited number of positive labels
- Propagate from better known patients/drugs to less known ones

- Patient similarity
- Drug similarity
- Patient-drug prior association

Personalized Medicine

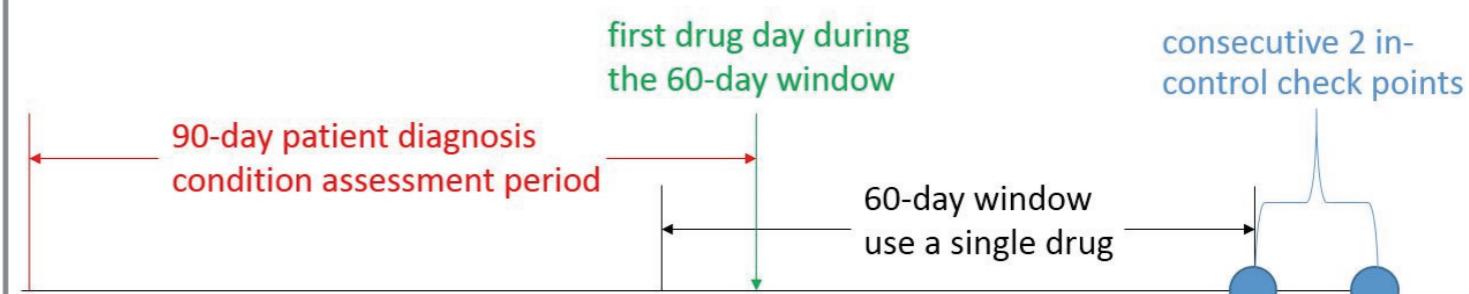


Condition: Dyslipidemia

Criterion

Low-Density Lipoprotein (LDL) level is below 130 mg/dL, is considered to be “well-controlled”

Evaluation Period



Data

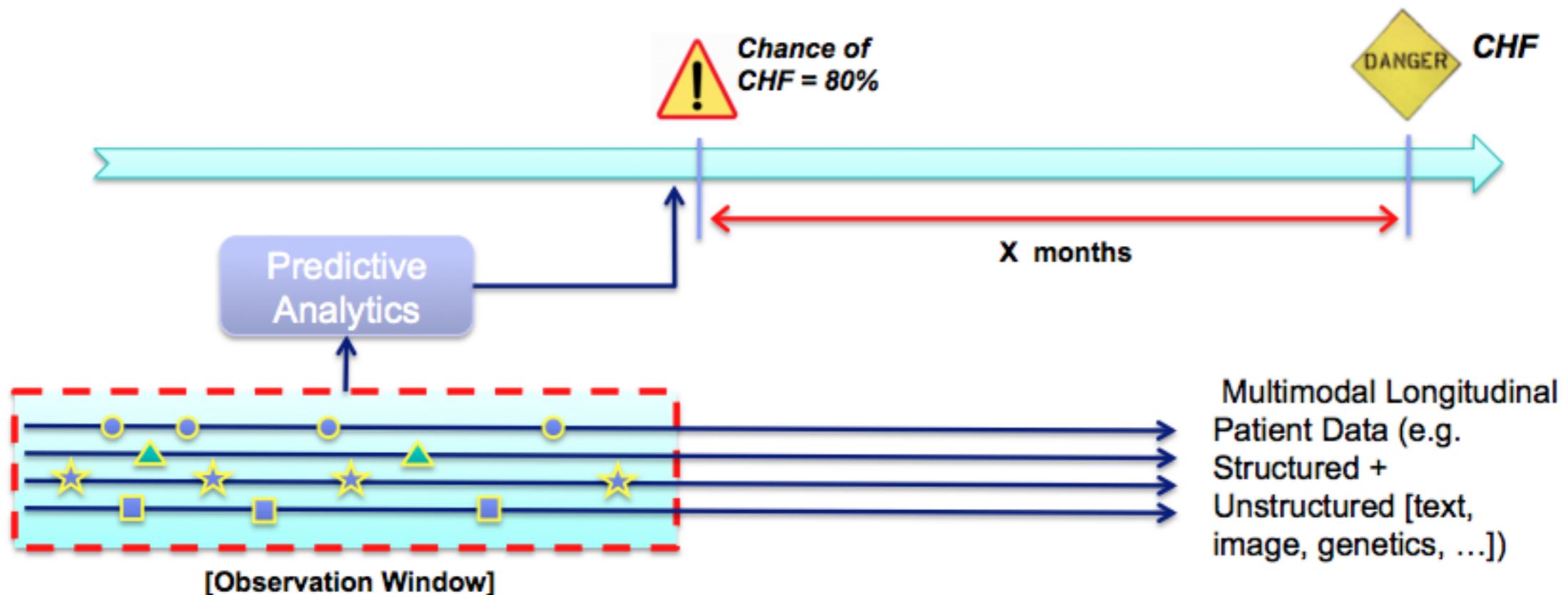
1219 Distinct Patients

4 Drugs: Atorvastatin, Lovastatin, Pravastatin, Simvastatin

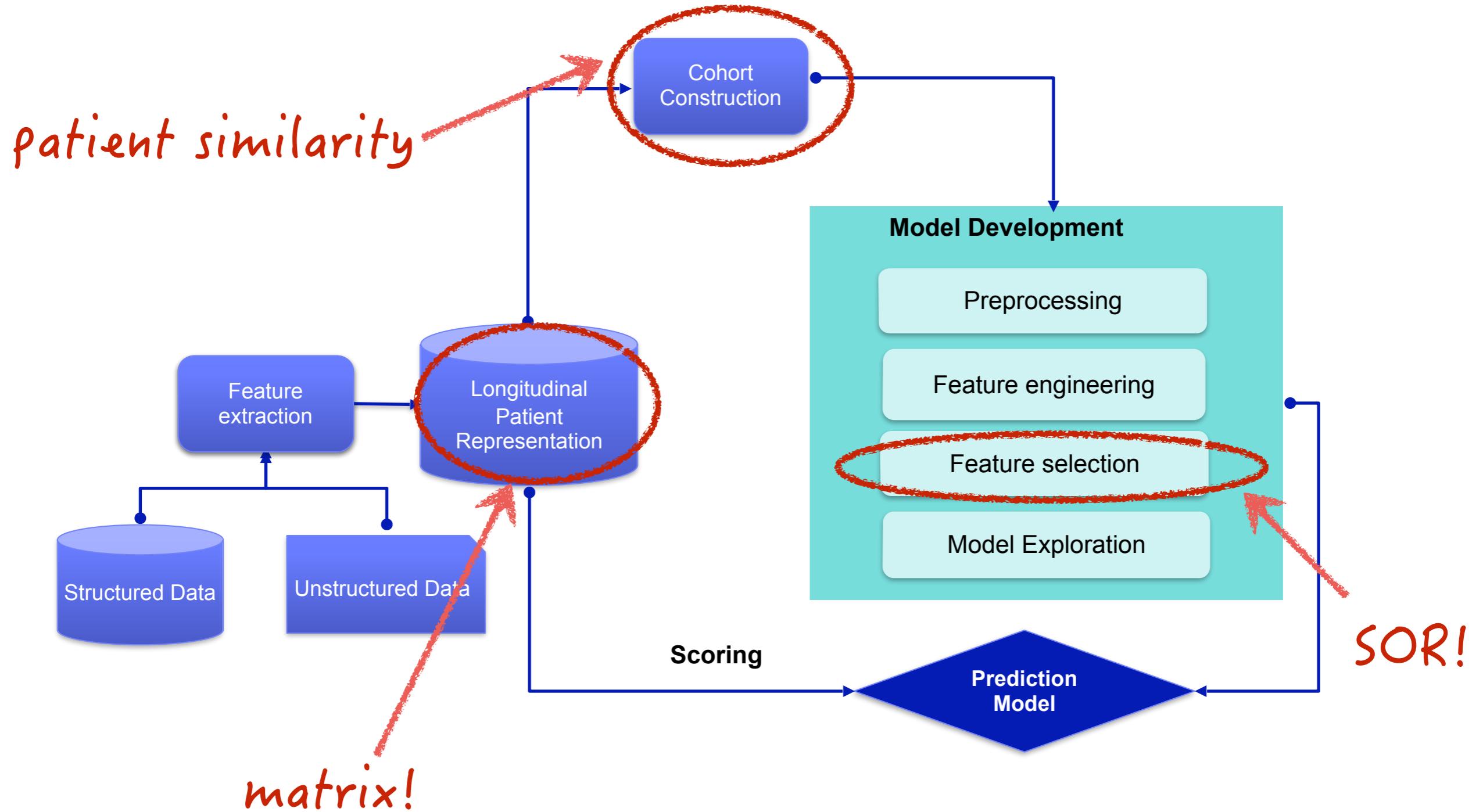
Roadmap

- Background
- Healthcare Data
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Predictive Modeling Scenario Revisited

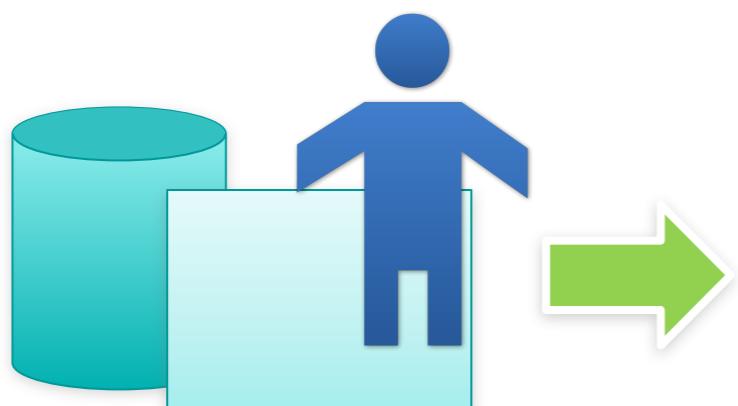


Predictive Modeling Pipeline



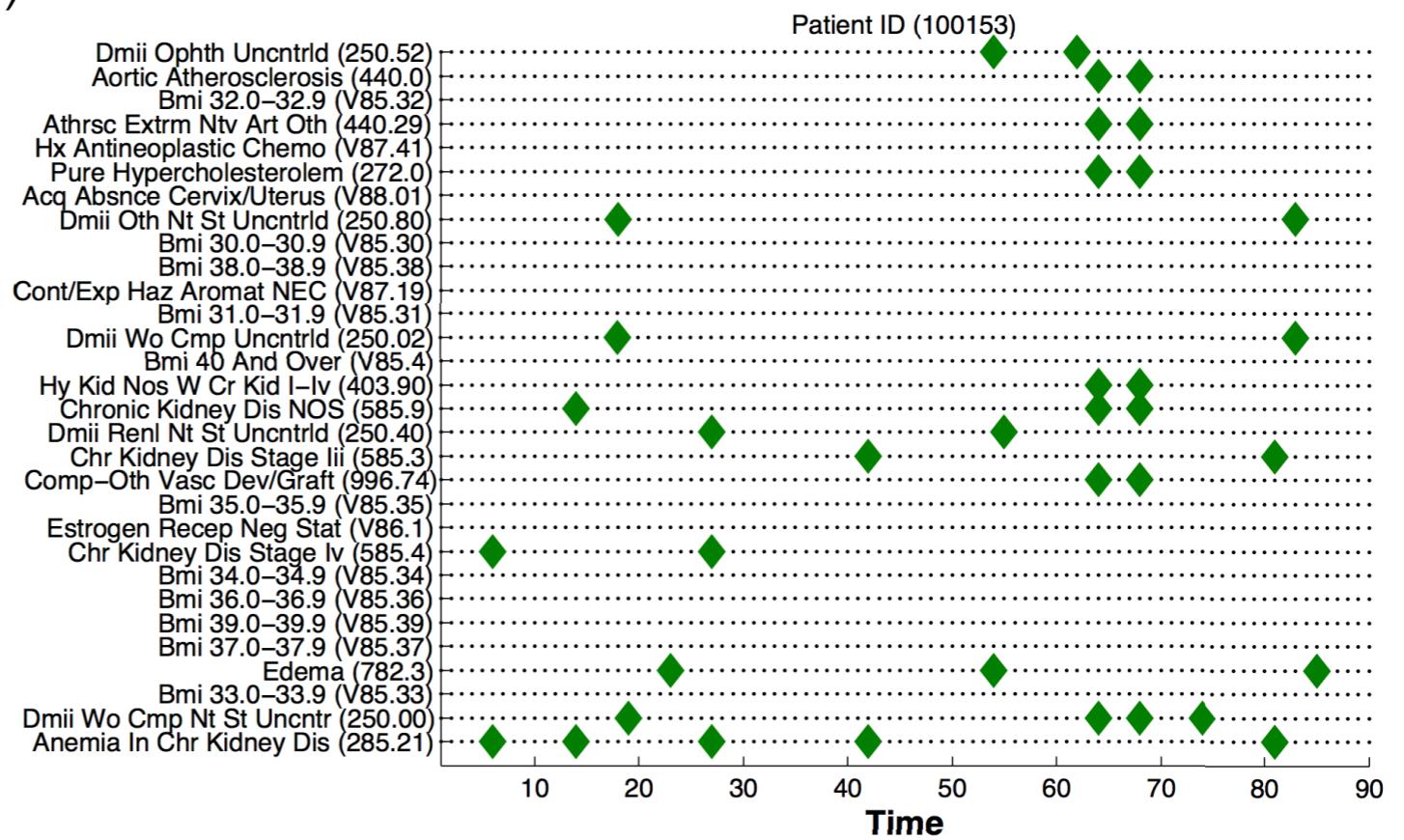
EMR Sparsity

- Longitudinal patient matrices.
- Challenges
 - Different start time, end time, and length of records.
 - Extremely sparse (low-density)



Patient
Longitudinal
EMR Data

Diagnosis Codes
CHF Cohort Density
0.0034%



Is Sparsity Good?

Sparsity may cause problems in phenotyping patients.

Patient 1	...	t	t+1	t+2	...
Heart Failure		X		X	
Head Injury			X		
...					
Patient 2	...	t	t+1	t+2	...
Heart Failure		X	X	X	
Head Injury					
...					

Questions 1

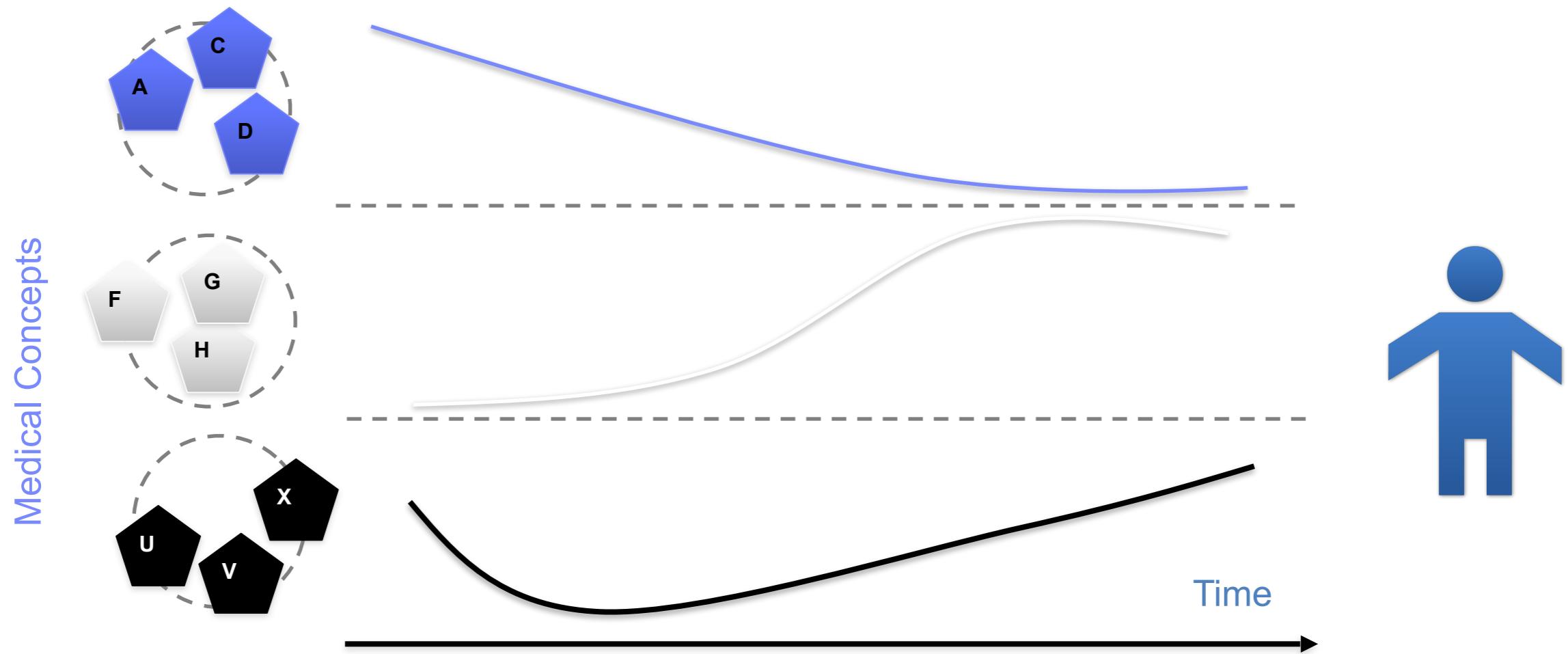
Is heart failure cured at
this time?

Question 2

Does patient 1 have less risk
when predicting future heart
failure than patient 2?

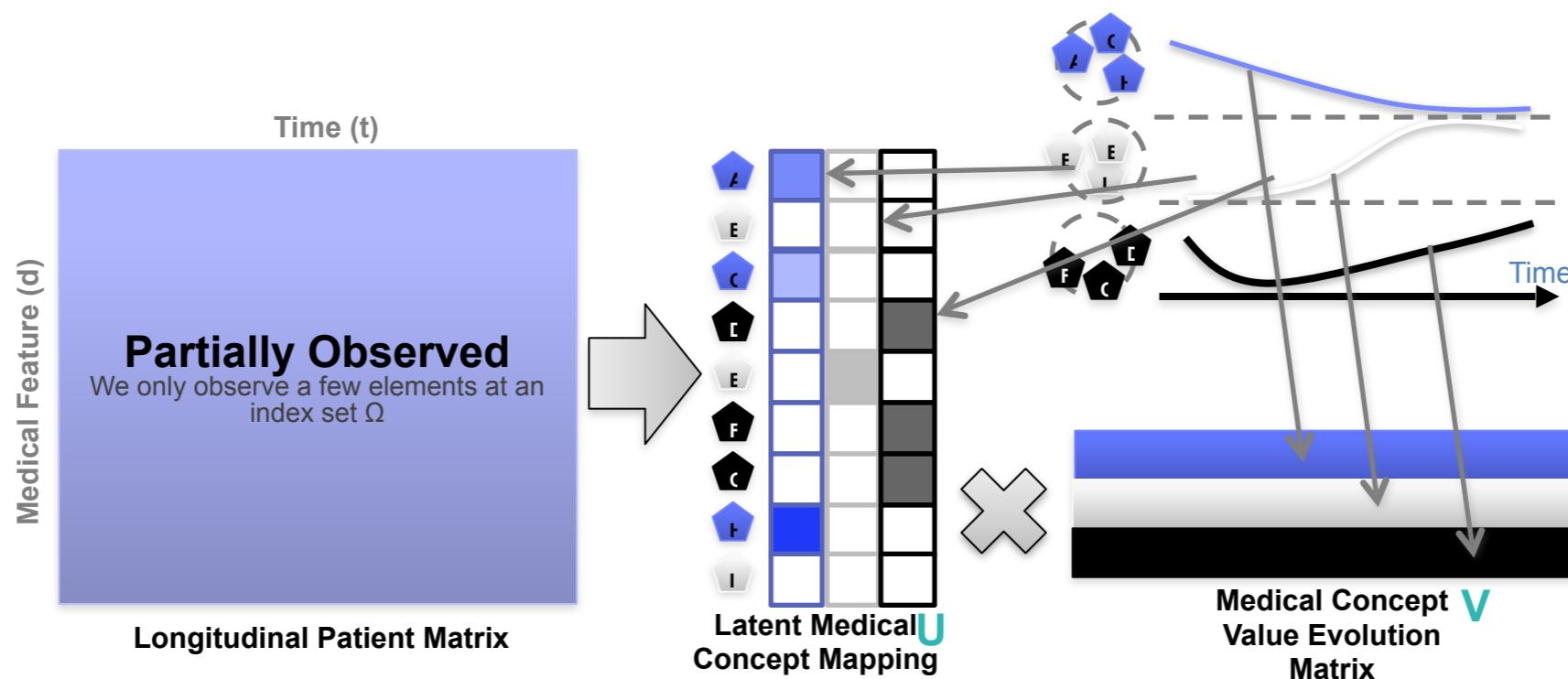
Medical Concepts

- In the disease progression, biomarkers evolve continuously over time.
- We assume that:
 - There exist some high-level *medical concepts*, which consist of raw medical features.
 - For each patient, the medical concepts *evolve smoothly*.



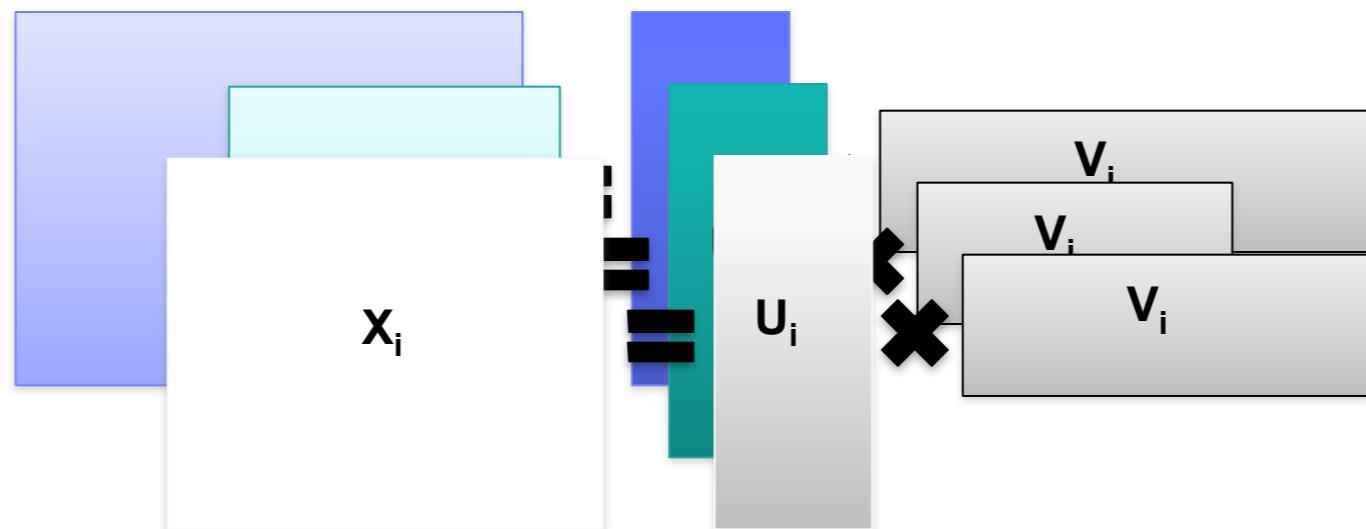
Patient Record Densifier

- Assume the *full* longitudinal patient matrix can be approximated by a low rank matrix
 - Macro-phenotype mapping matrix U : *Sparse, Non-negative*
 - Concept value evolution matrix V : *Temporal Smoothness*



Individual Basis Approach

- Assume that each patient has different medical concepts from other patients



- Formulation

$$\min_{\{S_i\}, \{U_i\}, \{V_i\}} \sum_{i=1}^n \frac{1}{2t_i} \|S_i - U_i V_i\|_F^2 + \lambda_1 \|U_i\|_1 + \lambda_2 \sum_{i=1}^n \frac{1}{2t_i} \|V_i\|_F^2 + \lambda_3 \sum_{i=1}^n \frac{1}{2t_i} \|V_i R_i\|_F^2$$

subject to: $\mathcal{P}_{\Omega_i}(S_i) = \mathcal{P}_{\Omega_i}(X_i)$, $U_i \geq 0, \forall i$

Matrix Completion

Completion via matrix factorization. Enforce a low rank factorization U_i, V_i and encourage the values of $U_i V_i$ at the observed location to be close to the observed ones.

Meaningful Medical Concepts
Medical concepts involve non-negative components of medical features.

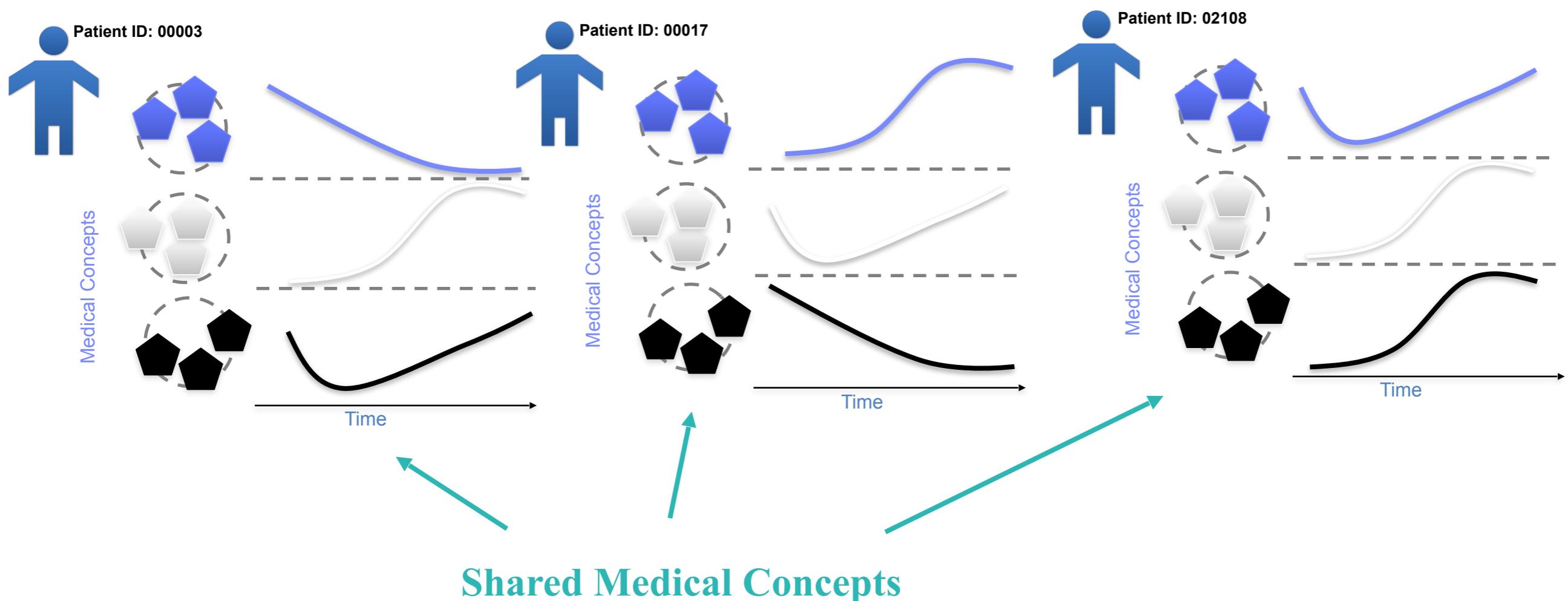
Sparsity
Controllable sparsity that encourages a few medical features in each concept.

Overfitting Control
Prevent V_i from overfitting.

Temporal Smoothness
Couple the columns of V_i and force them to close to each other.

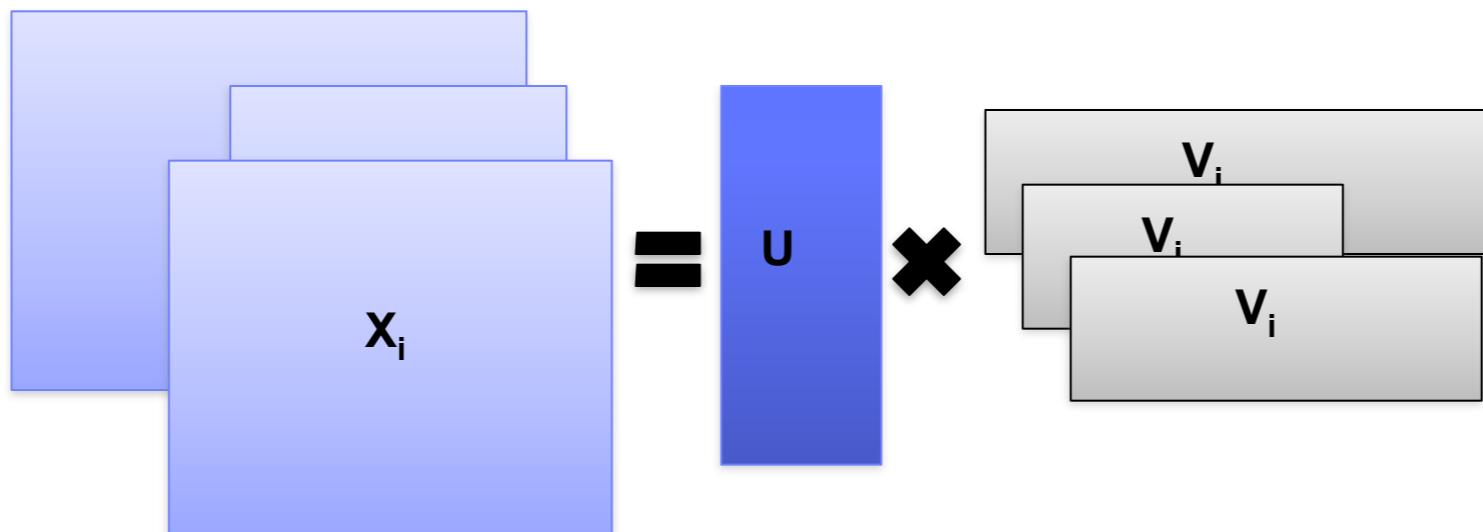
Shared Basis Approach

- If patients are similar to each other, then the patients may share the same set of medical concepts, while each of them may evolve differently.



Shared Basis Approach

- Assume that each patient has shared medical concepts from other patients



- Formulation

Shared Medical Concept Mapping

$$\min_{\{S_i\}, \{U\}, \{V_i\}} \sum_{i=1}^n \frac{1}{2t_i} \|S_i - UV_i\|_F^2 + \lambda_1 \|U\|_1 + \lambda_2 \sum_{i=1}^n \frac{1}{2t_i} \|V_i\|_F^2 + \lambda_3 \sum_{i=1}^n \frac{1}{2t_i} \|V_i R_i\|_F^2$$

subject to: $\mathcal{P}_{\Omega_i}(S_i) = \mathcal{P}_{\Omega_i}(X_i), U \geq 0$

Clustered Basis Approach

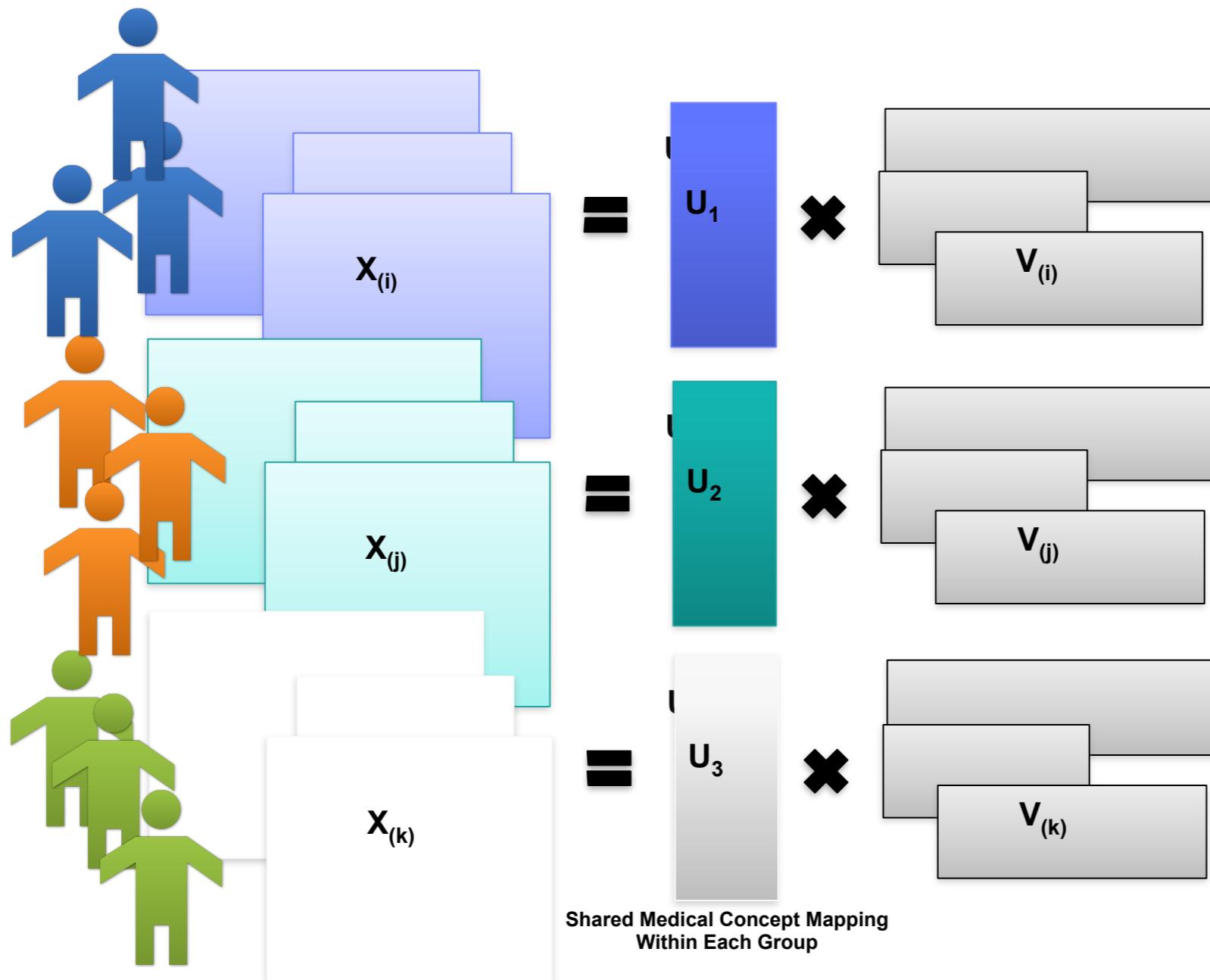
- Patients may exhibit some groups
 - Inside each group: homogeneous, share medical concepts.
 - Outside group: heterogeneous
 - Transfer knowledge within each group



- Each group may consist of patients of similar health condition:
 - Affected by the same set of diseases or share comorbidity.
 - Similar in terms of blood type, geographic, age, sex.
 - Leverage external domain knowledge.

Clustered Basis Approach

Simultaneously clustering patients and densification

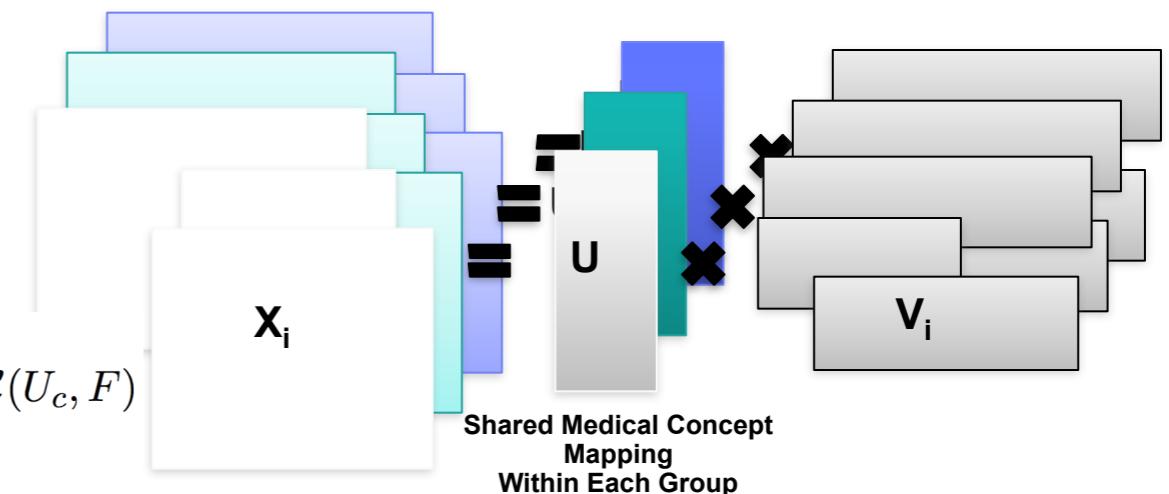


Clustered Basis Approach

- Model-based patient clustering in Pacifier
 - Spectral K-Means objective on medical concepts

$$\min_{F: F^T F = I_p} \frac{1}{2k} \sum_{c=1}^k (\text{tr}(U_c^T U_c) - \text{tr}(F^T U_c^T U_c F)) := \min_{F: F^T F = I_p} \frac{1}{2k} \sum_{c=1}^k \mathcal{C}(U_c, F)$$

- Formulation of Pacifier-CBA



Patient Clustering

$$\min_{F, \{U_{(i)}, V_{(i)}, S_{(i)}\}} \frac{1}{n} \sum_{i=1}^n \frac{1}{2t_i} \|S_{(i)} - U_{(i)} V_{(i)}\|_F^2 + \lambda_c \frac{1}{2k} \sum_{c=1}^k \mathcal{C}(U_c, F) + \mathcal{T}(\{V_{(i)}\})$$

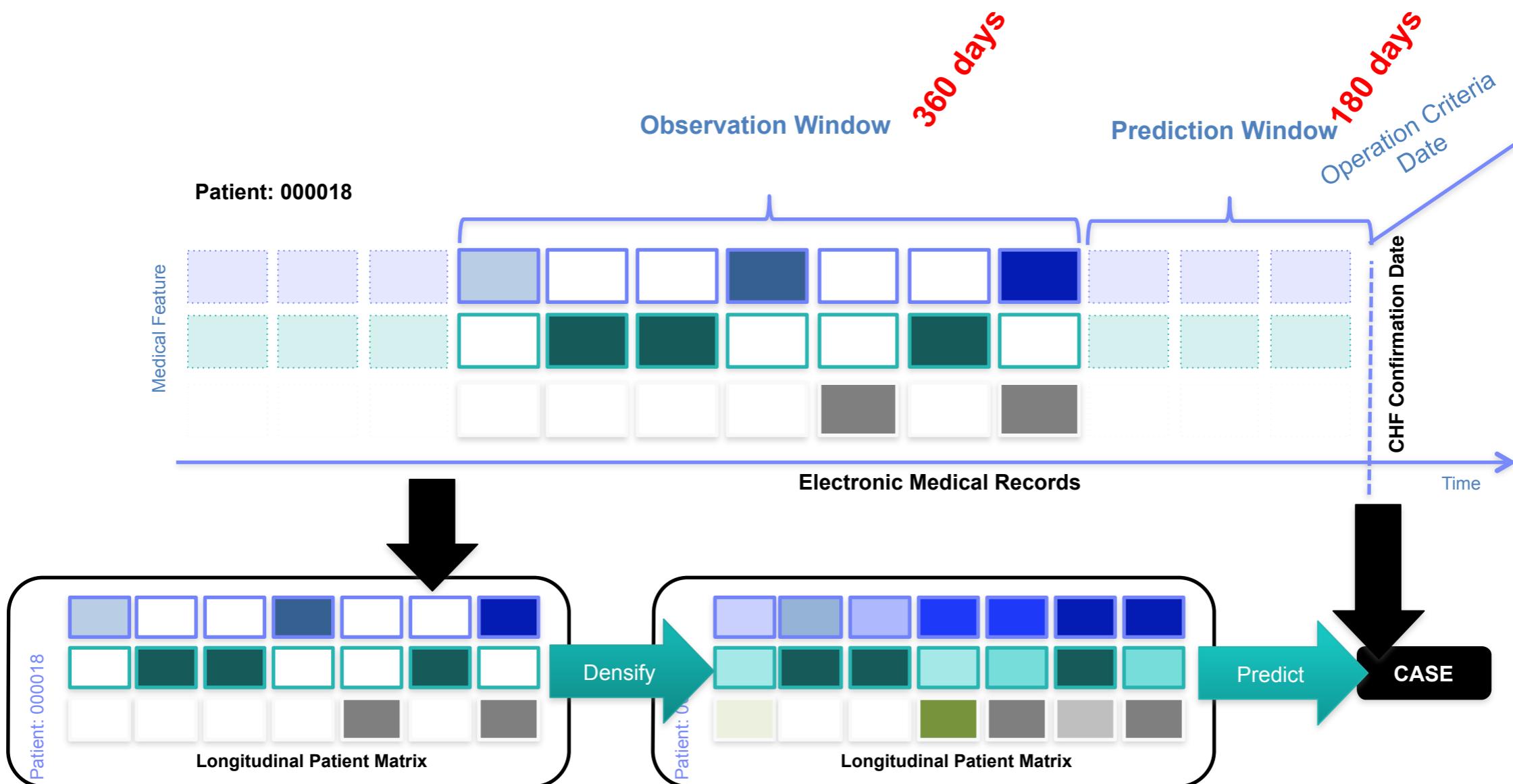
subject to: $\mathcal{P}_{\Omega_{(i)}}(S_{(i)}) = \mathcal{P}_{\Omega_{(i)}}(X_{(i)}), \{U_{(i)}\} \in \mathcal{U}, F^T F = I_p$

Matrix Completion via Factorization **Meaningful Medical Concepts** **Temporal Smoothness**

$$\mathcal{U} := \{U | U \geq 0, \|U\|_1 \leq \tau\}$$

$$\mathcal{T}(\{V_{(i)}\}) = \lambda_2 \frac{1}{n} \sum_{i=1}^n \frac{1}{2t_i} \|V_{(i)}\|_F^2 + \lambda_3 \frac{1}{n} \sum_{i=1}^n \frac{1}{2t_i} \|V_{(i)} R_{(i)}\|_F^2.$$

Application in CHF Onset Prediction



Experimental Settings

Densification Configuration

Competing methods

Baseline - Zero Imputation (RAW)

Basic Imputation methods

Row Average (RowAvg)

Next Occurrence Carry Backward (NOCB)

Last Occurrence Carry Forward (LOCF)

Interpolation of Previous and Next (PvNxInpl)

Parameter selection

10-fold cross-validation

Classification Configuration

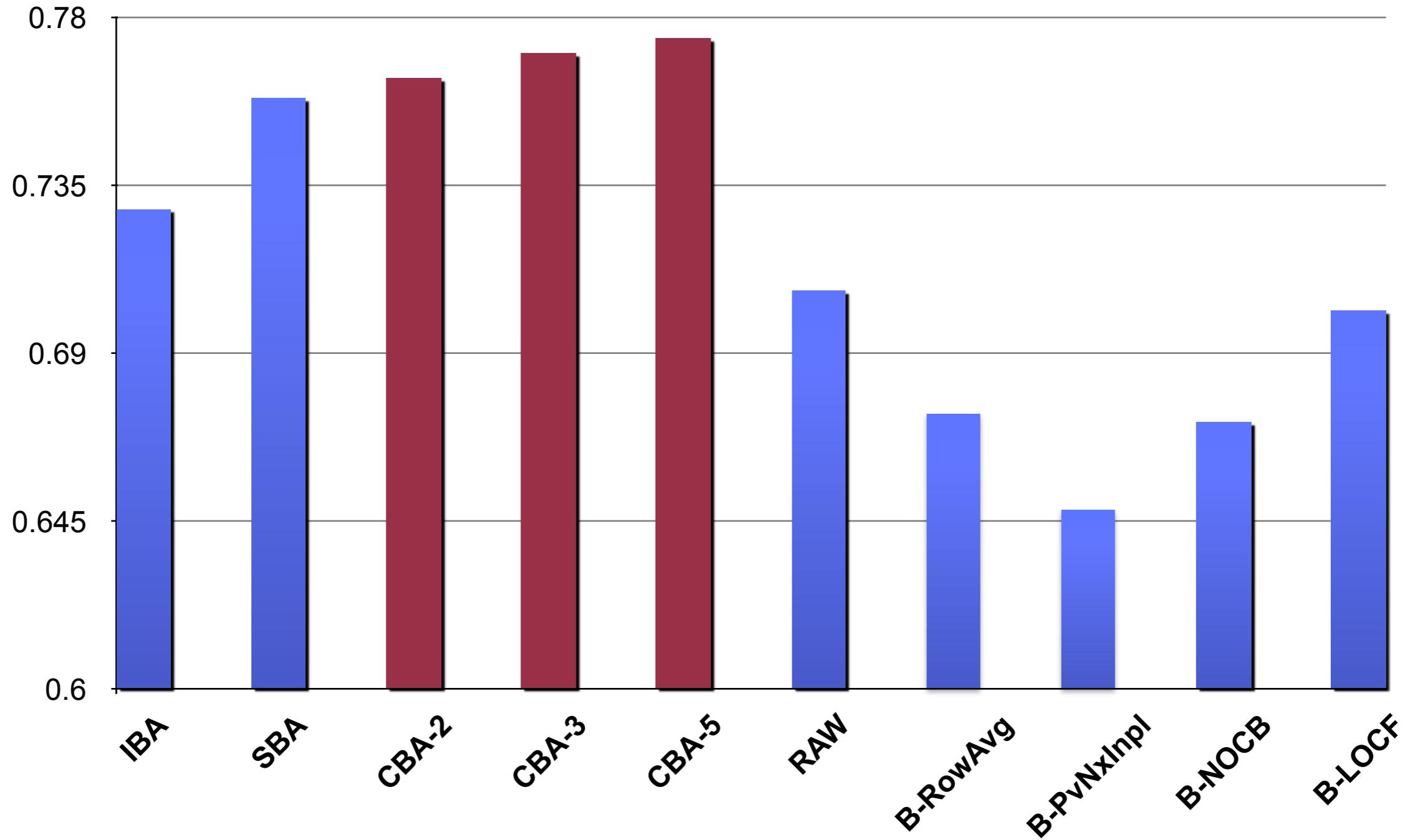
Training: 90% Percent

Classifier: sparse logistic regression

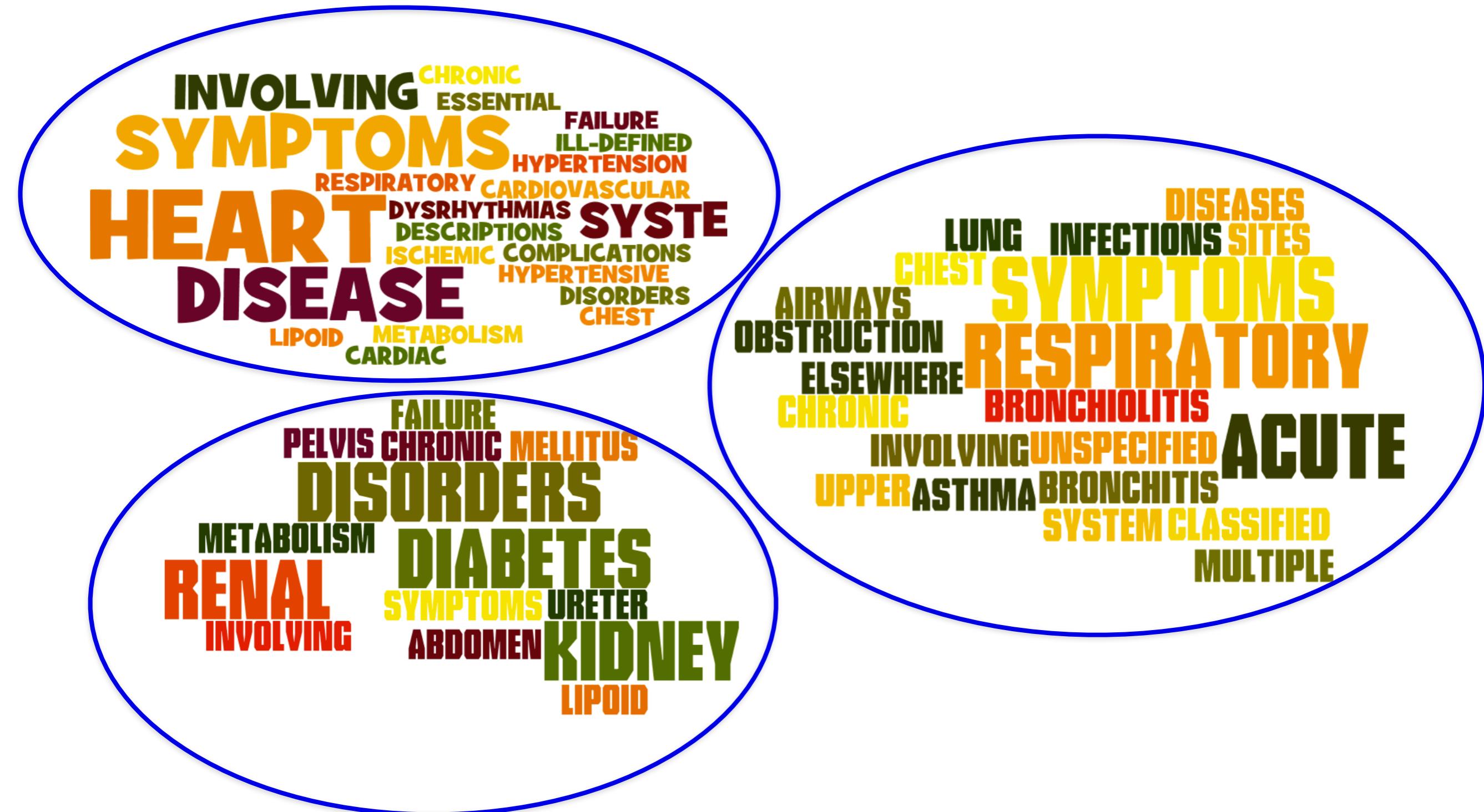
Model selection scheme: 5-fold cross validation

Metric: AUC

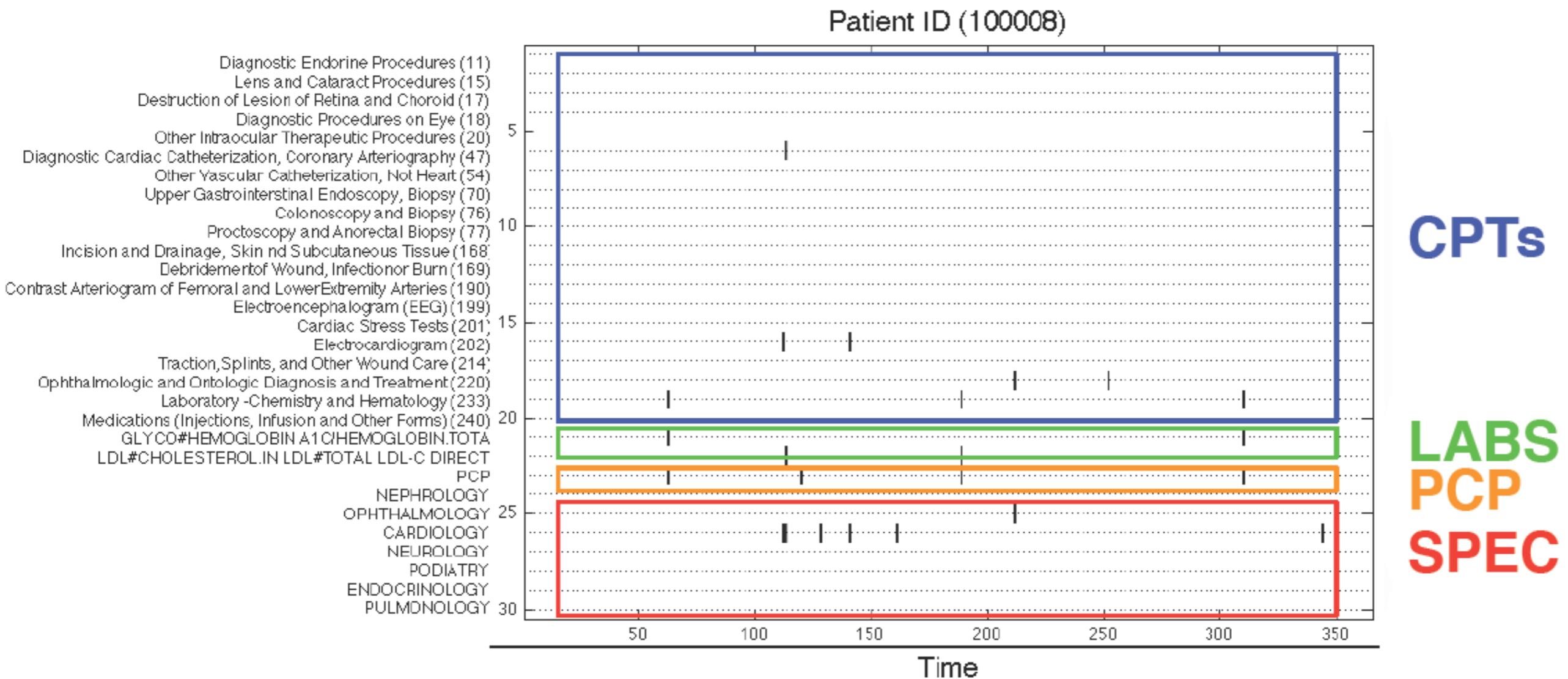
Performance



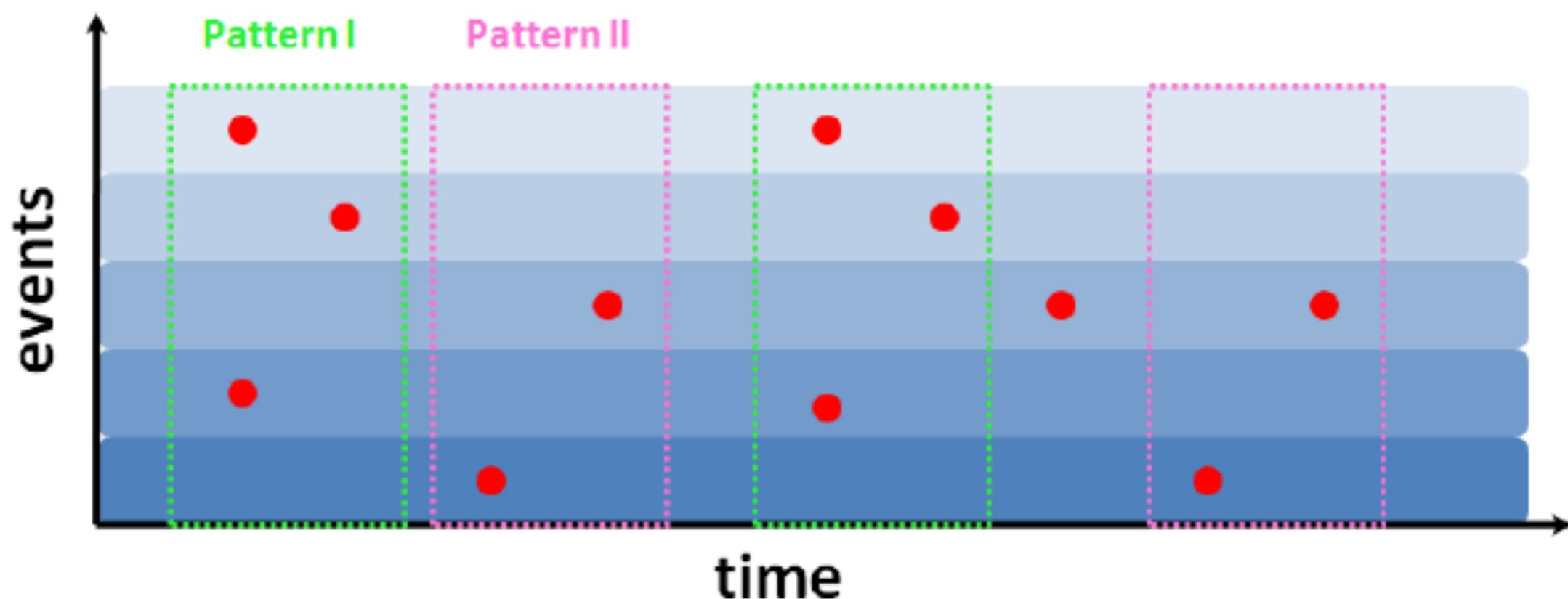
Learned Medical Concepts



Matrix Representation of EMR Revisited



Temporal Patterns

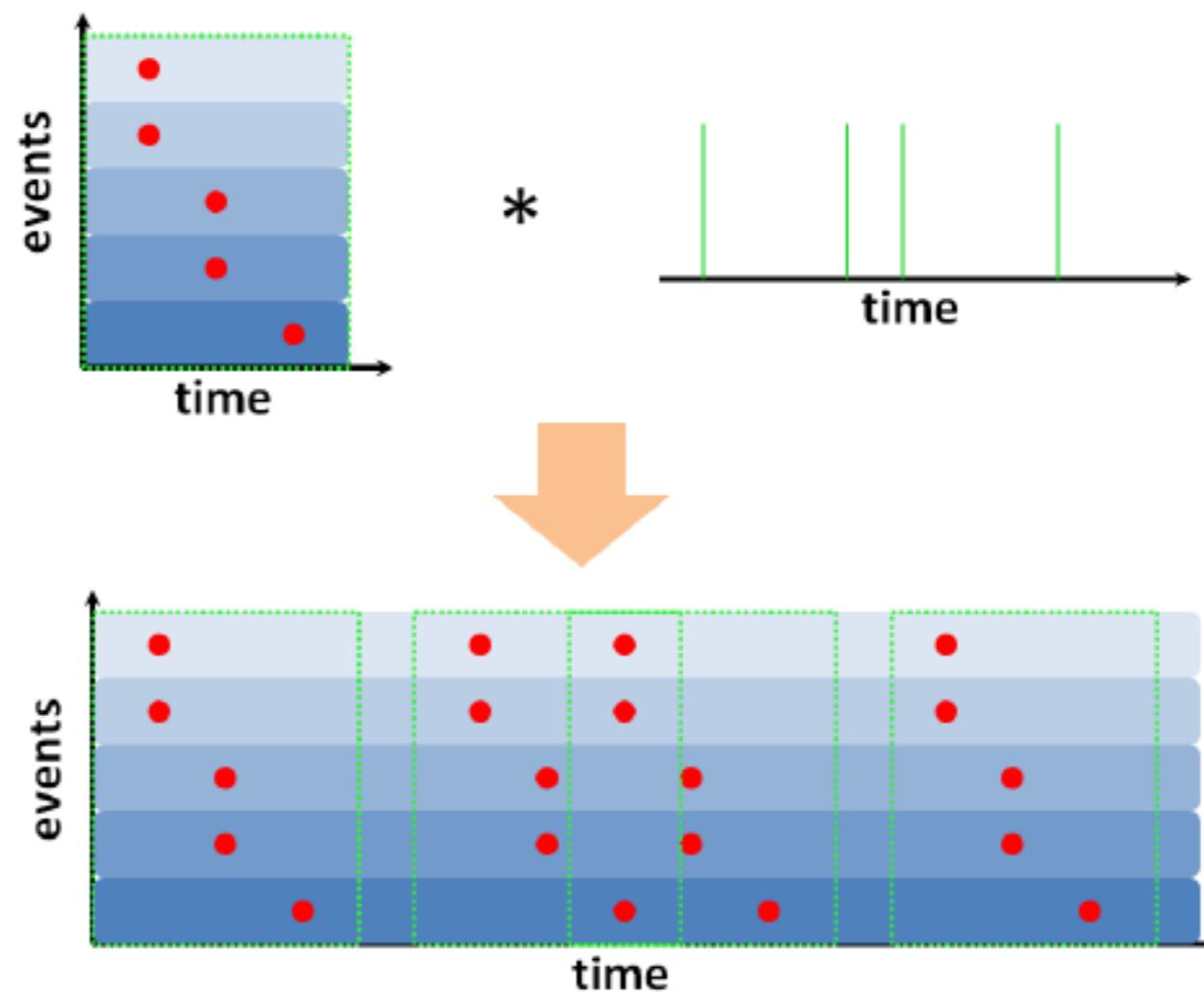


One-Side Convolution

Definition (One-Side Convolution). *The one-side convolution of $\mathbf{F} \in \mathbb{R}^{n \times m}$ and $\mathbf{g} \in \mathbb{R}^{t \times 1}$ is an $n \times t$ matrix with*

$$(\mathbf{F} * \mathbf{g})_{ij} = \sum_{k=1}^t g_{j-k+1} F_{ik}$$

Note that $g_j = 0$ if $j \leq 0$ or $j > t$, and $F_{ik} = 0$ if $k > m$.



One-Side Convolutional NMF

$$\begin{aligned} \min_{\mathcal{F}, \mathcal{G}} \quad & \mathcal{J} \\ s.t. \quad & \forall r = 1, \dots, R; c = 1, \dots, C \\ & \mathbf{F}^{(r)} \geq 0, \mathbf{g}_c^{(r)} \geq 0 \end{aligned}$$

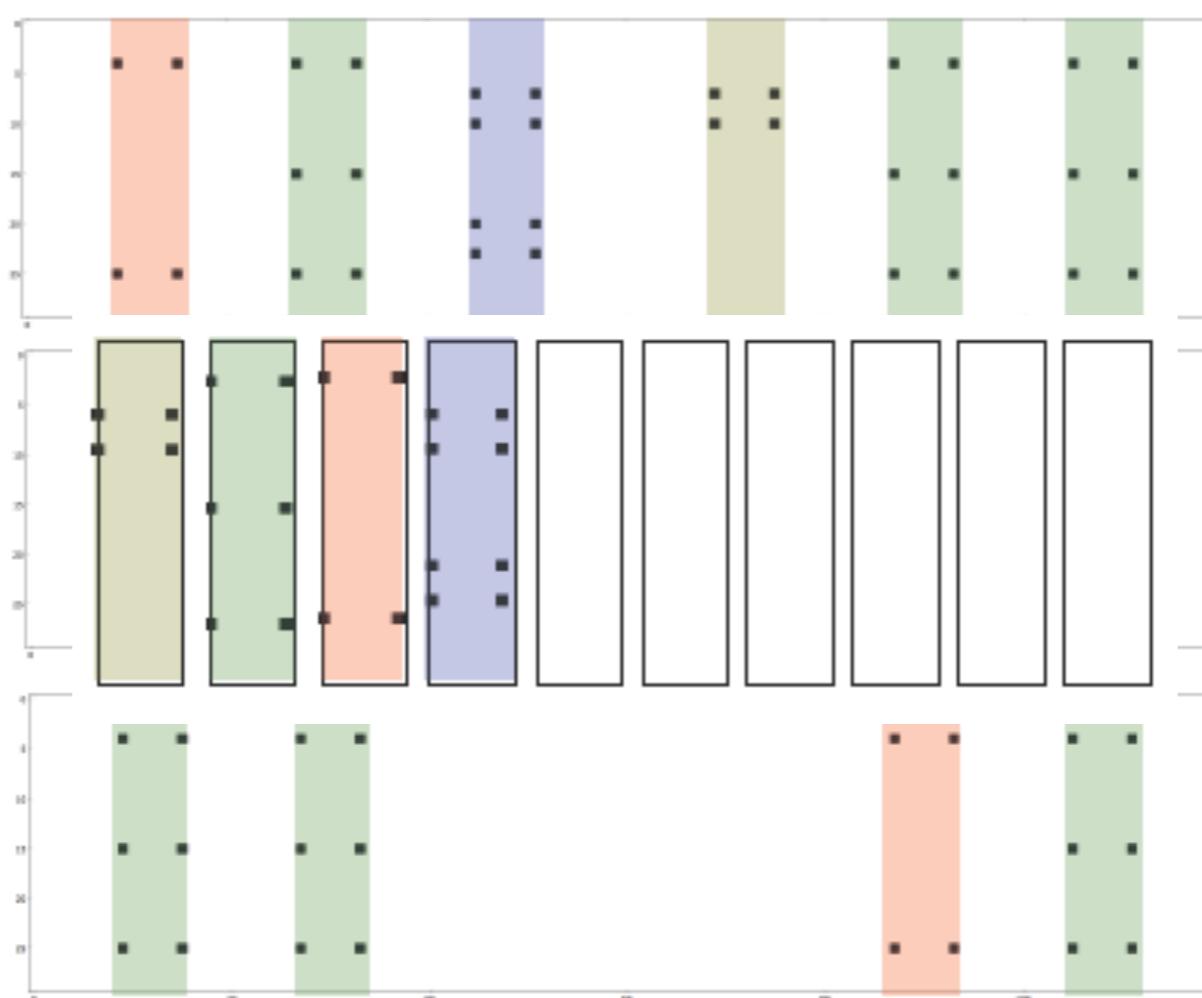
$$\begin{aligned} F_{ik}^{(r)} &\leftarrow F_{ik}^{(r)} \left(\frac{\sum_{c=1}^C \sum_{j=1}^t A_{cij}^{\beta-1} X_{cij} Y_{cij}^{\beta-2} g_{c_j-k+1}^{(r)}}{\sum_{c=1}^C \sum_{j=1}^t A_{cij} Y_{cij}^{\beta-1} g_{c_j-k+1}^{(r)} + \lambda_1} \right)^{\eta(\beta)} \\ g_{c_k}^{(r)} &\leftarrow g_c^{(r)} \left(\frac{\sum_{i=1}^n \sum_{j=1}^t A_{cij}^{\beta-1} X_{cij} Y_{cij}^{\beta-2} F_{i,j-k+1}^{(r)}}{\sum_{i=1}^n \sum_{j=1}^t A_{cij} Y_{cij}^{\beta-1} F_{i,j-k+1}^{(r)} + \lambda_2} \right)^{\eta(\beta)} \end{aligned}$$

$$\mathcal{J} = \sum_{c=1}^C d_\beta \left(\mathbf{A}_c \odot \mathbf{X}_c, \mathbf{A}_c \odot \left(\sum_{r=1}^R \mathbf{F}^{(r)} * \mathbf{g}_c^{(r)} \right) \right) + \lambda_1 \sum_{r=1}^R \|\mathbf{F}^{(r)}\|_1 + \lambda_2 \sum_{c=1}^C \sum_{r=1}^R \|\mathbf{g}_c^{(r)}\|_1 \quad \eta(\beta) = \begin{cases} \frac{1}{2-\beta}, & \beta < 1 \\ 1, & 1 \leq \beta \leq 2 \\ \frac{1}{\beta-1}, & \beta > 2 \end{cases}$$

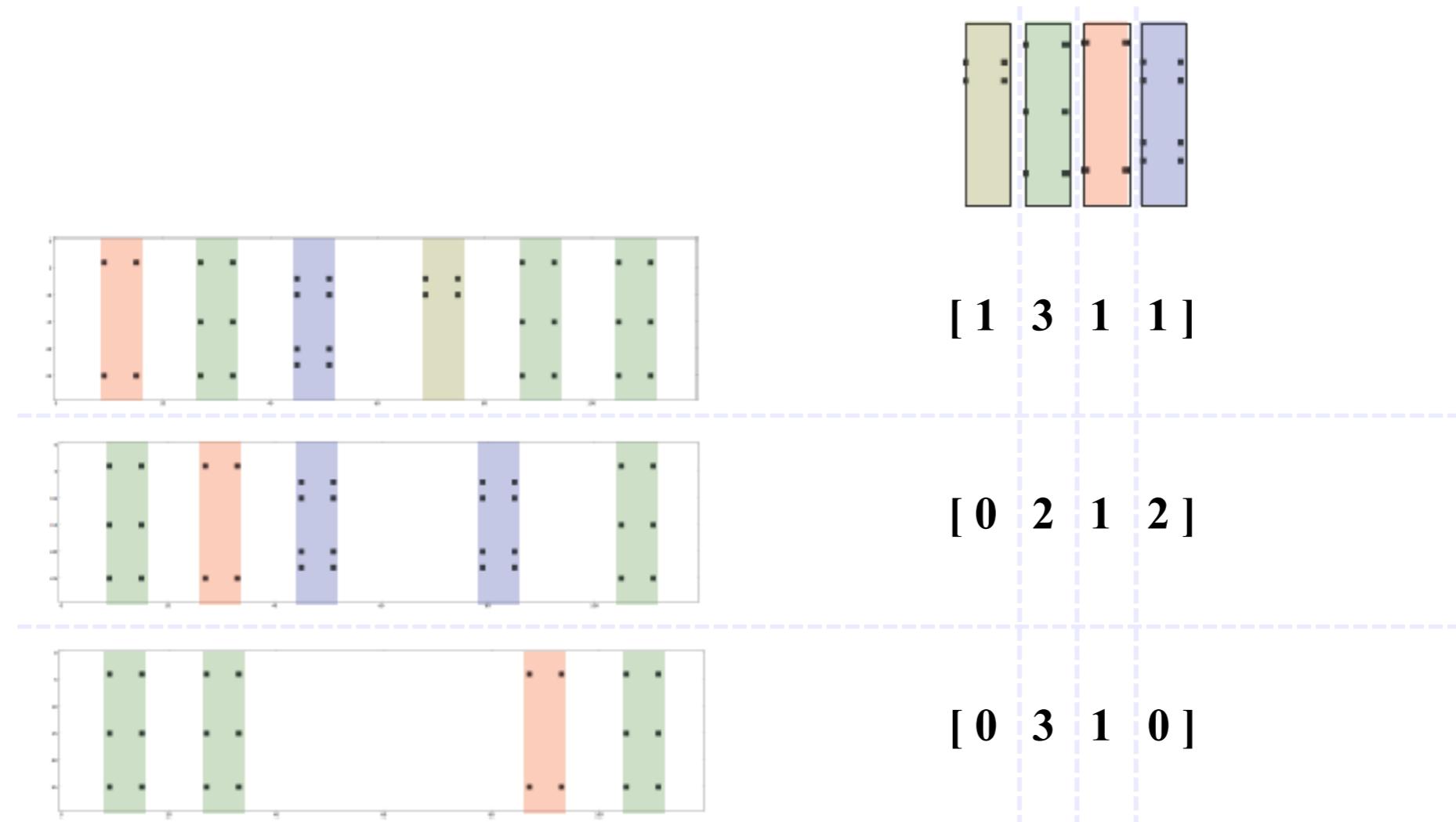
Definition (β -divergence) The β -divergence between two matrices \mathbf{A} and \mathbf{B} with the same size is

$$d_\beta(\mathbf{A}, \mathbf{B}) = \frac{1}{\beta(\beta-1)} \sum_{ij} \left(A_{ij}^\beta + (\beta-1)B_{ij}^\beta - \beta A_{ij} B_{ij}^{\beta-1} \right)$$

Synthetic Example

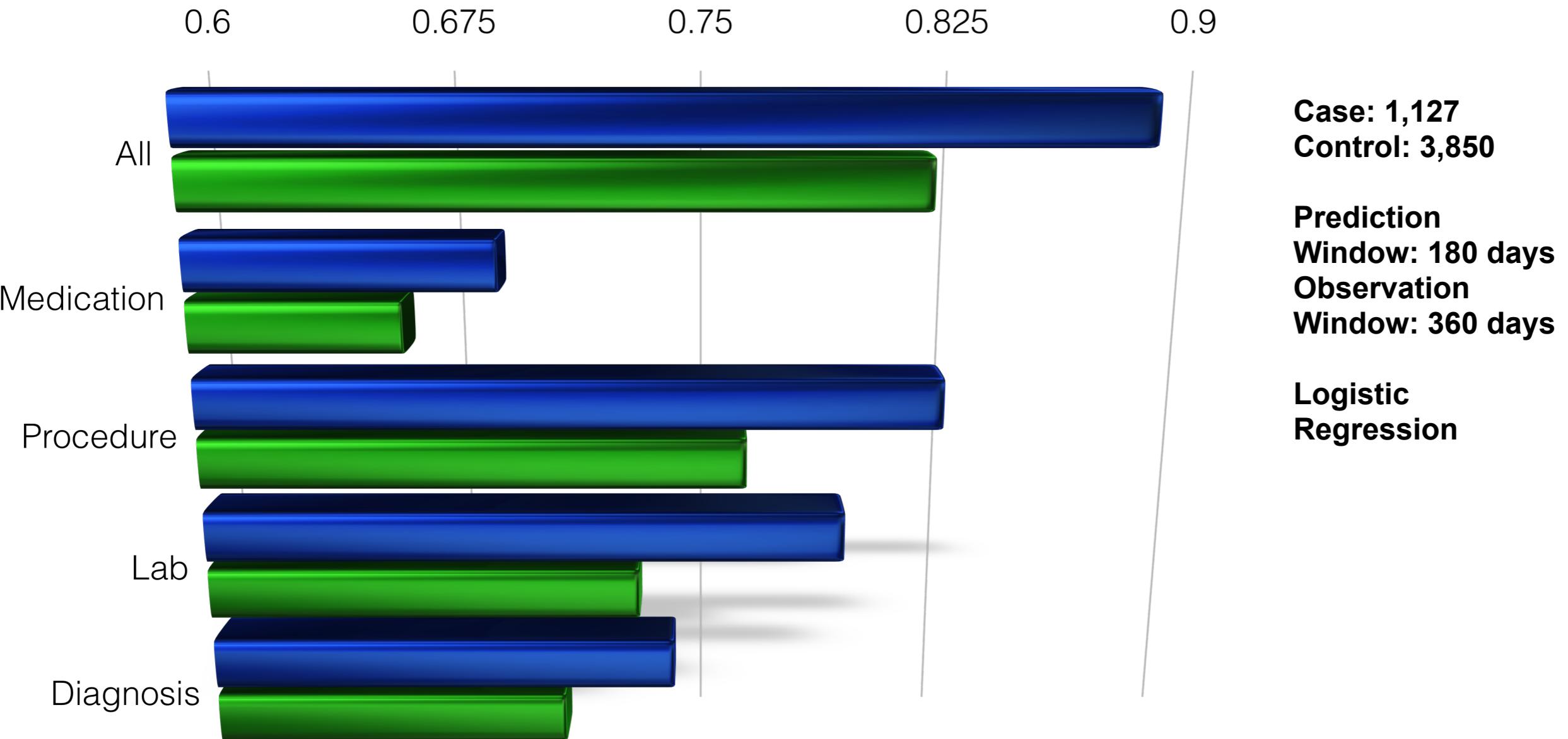


Bag-of-Pattern Representation



CHF Onset Prediction

■ Temporal+Static ■ Static



Roadmap

- Background
- Healthcare Data
- Patient Similarity Analytics
- Predictive Modeling
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

What is clinical pathway?



Karen Zander & Kathleen (Kathy) Bower - Principals of the Center For Case Management and board members of our European Pathway Association.

(1980's Karen Zander & Kathleen Bower, New England Medical Center, Boston)

Clinical pathways are designed to provide optimal patient care for groups of patients

- A tool to incorporate local and national guidelines into everyday practice
- Better Service, Less Cost

What is Clinical Pathway

Eligibility & exclusion criteria

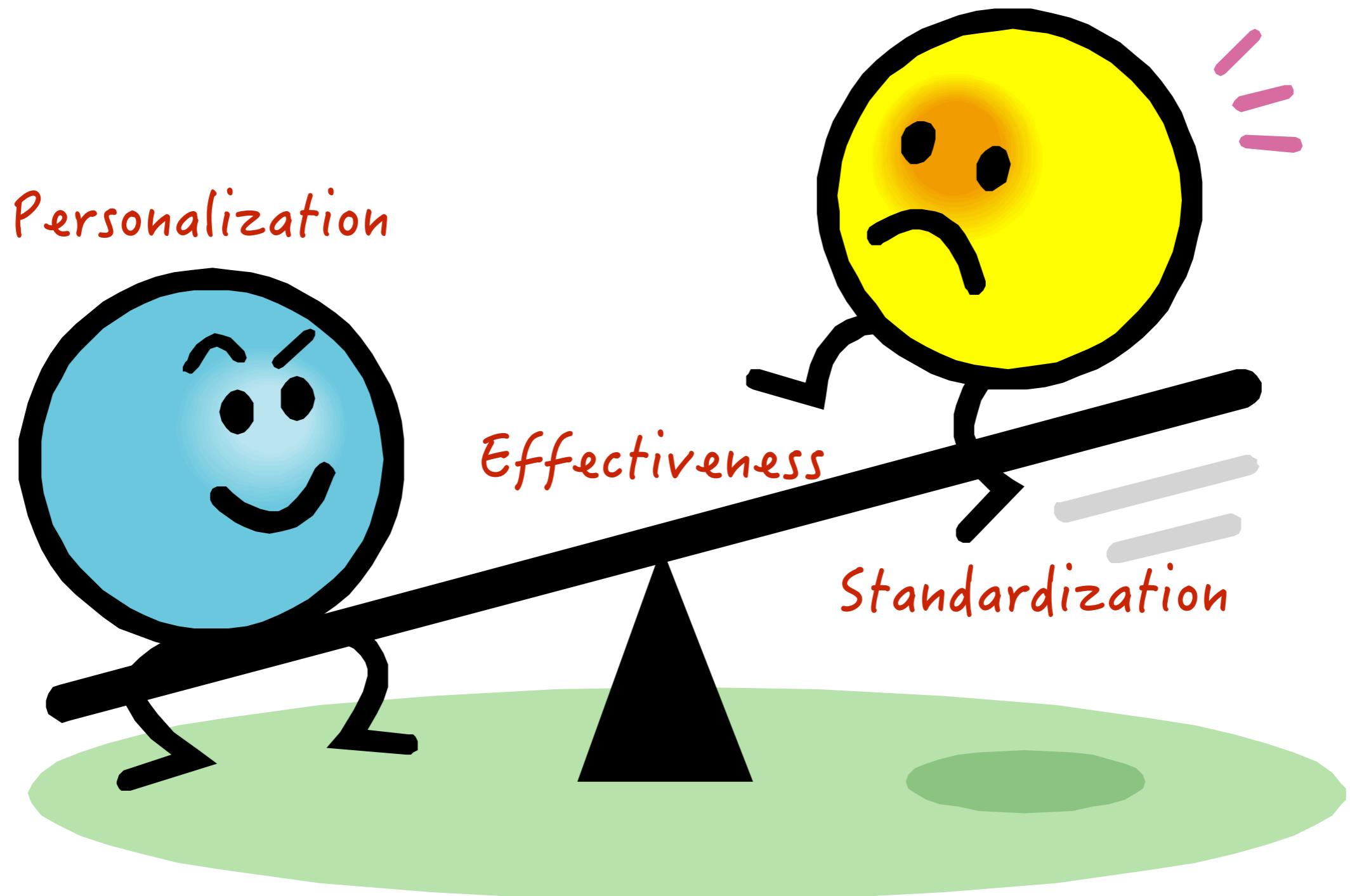
Bronchial lung cancer clinical pathway			
Suitable for: First-listed diagnosis is bronchial lung cancer (ICD-10: C34; D02.2) Local excision of pulmonary / Lobectomy / Pneumonectomy + Systematic lymph node dissection / Thoracotomy surgery (ICD-9-CM: 3-32.29/32.3-32.5)			
Patient Name:	Gender:	Age:	
Admission Date:	Discharge Date:	Expected LOS:	12-21 days
Major diagnosis and treatment	Day 1 <ul style="list-style-type: none"> <input type="checkbox"/> Medical history inquiry and physical examination <input type="checkbox"/> Write patient record <input type="checkbox"/> Issue laboratory orders and check request form <input type="checkbox"/> Attending rounds <input type="checkbox"/> Initially set the treatment plan 	Day 2-6 (preoperative day) <ul style="list-style-type: none"> <input type="checkbox"/> Higher authority physician rounds <input type="checkbox"/> Preoperative preparation <input type="checkbox"/> Clinical stage and preoperative evaluation <input type="checkbox"/> Preoperative discussion and surgical planning <input type="checkbox"/> Complete the consultation within relevant sections according to patient condition <input type="checkbox"/> Resident completes medical records including the course of the disease and preoperative log summary, the superior physician records. <input type="checkbox"/> Sign the informed consent procedure, goods agreement at their own expense, blood transfusion consent, consent authorization 	Day 4-7 (operative day) <ul style="list-style-type: none"> <input type="checkbox"/> Indwelling catheter before surgery <input type="checkbox"/> Surgery <input type="checkbox"/> Surgeon completes the operation record <input type="checkbox"/> Resident completes postoperative course <input type="checkbox"/> Higher authority physician rounds <input type="checkbox"/> Observation of vital signs <input type="checkbox"/> Account of illness and postoperative precautions to patients and their families
Doctor's major advice	Long-term medical advice: <ul style="list-style-type: none"> <input type="checkbox"/> Thoracic surgery Secondary care <input type="checkbox"/> Normal diet Temporary medical advice: <ul style="list-style-type: none"> <input type="checkbox"/> Blood, urine, stool routine examination <input type="checkbox"/> Coagulation, blood type, liver and kidney function, electrolytes examination, infectious disease screening, tumor markers check <input type="checkbox"/> Lung function, arterial blood gas analysis, ECG echocardiography <input type="checkbox"/> Sputum cytology, bronchoscopy + biopsy <input type="checkbox"/> Imaging: lateral chest X-ray, chest CT, abdominal ultrasound or CT, whole body bone scan, brain MRI or CT <input type="checkbox"/> When necessary: PET-CT or SPECT, mediastinoscopy, 24-hour ambulatory ECG, percutaneous lung biopsy, etc. 	Long-term medical advice: <ul style="list-style-type: none"> <input type="checkbox"/> Atomizing inhalation Temporary medical advice: <ul style="list-style-type: none"> <input type="checkbox"/> Schedule for tomorrow under general anesthesia: <input type="radio"/> Local excision of pulmonary <input type="radio"/> Lobectomy <input type="radio"/> Pneumonectomy <input type="radio"/> Thoracotomy surgery <input type="checkbox"/> No food or water intake 6 hours before surgery <input type="checkbox"/> Enema the night before surgery <input type="checkbox"/> Preoperative skin preparation <input type="checkbox"/> Preparation of blood transfusion <input type="checkbox"/> Sedative drugs (where appropriate) <input type="checkbox"/> Preparation of antibacterial drugs in surgery <input type="checkbox"/> Other special advices 	Long-term medical advice: <ul style="list-style-type: none"> <input type="checkbox"/> General thoracic surgery postoperative care <input type="checkbox"/> Premium or first level nursing <input type="checkbox"/> Liquid food intake 6 hours after clear-headed <input type="checkbox"/> Oxygen inhalation <input type="checkbox"/> Body temperature, ECG, blood pressure, respiration, pulse, blood oxygen saturation monitoring <input type="checkbox"/> Record the amount of chest drainage <input type="checkbox"/> Continued catheterization, record 24-hour intake and output <input type="checkbox"/> Atomizing inhalation <input type="checkbox"/> Prophylactic antibiotics <input type="checkbox"/> Analgesic Temporary medical advice: <ul style="list-style-type: none"> <input type="checkbox"/> Other special advices
Major nursing care	<ul style="list-style-type: none"> <input type="checkbox"/> Introduce the ward environment, facilities and equipment <input type="checkbox"/> Admission nursing assessment <input type="checkbox"/> Aid smoking cessation 	<ul style="list-style-type: none"> <input type="checkbox"/> Education, preoperative skin preparation <input type="checkbox"/> Inform of no water and food intake <input type="checkbox"/> Respiratory exercises 	<ul style="list-style-type: none"> <input type="checkbox"/> Observe changes in condition <input type="checkbox"/> Postoperative care of psychological and life <input type="checkbox"/> Maintain patency of airway

pathway title

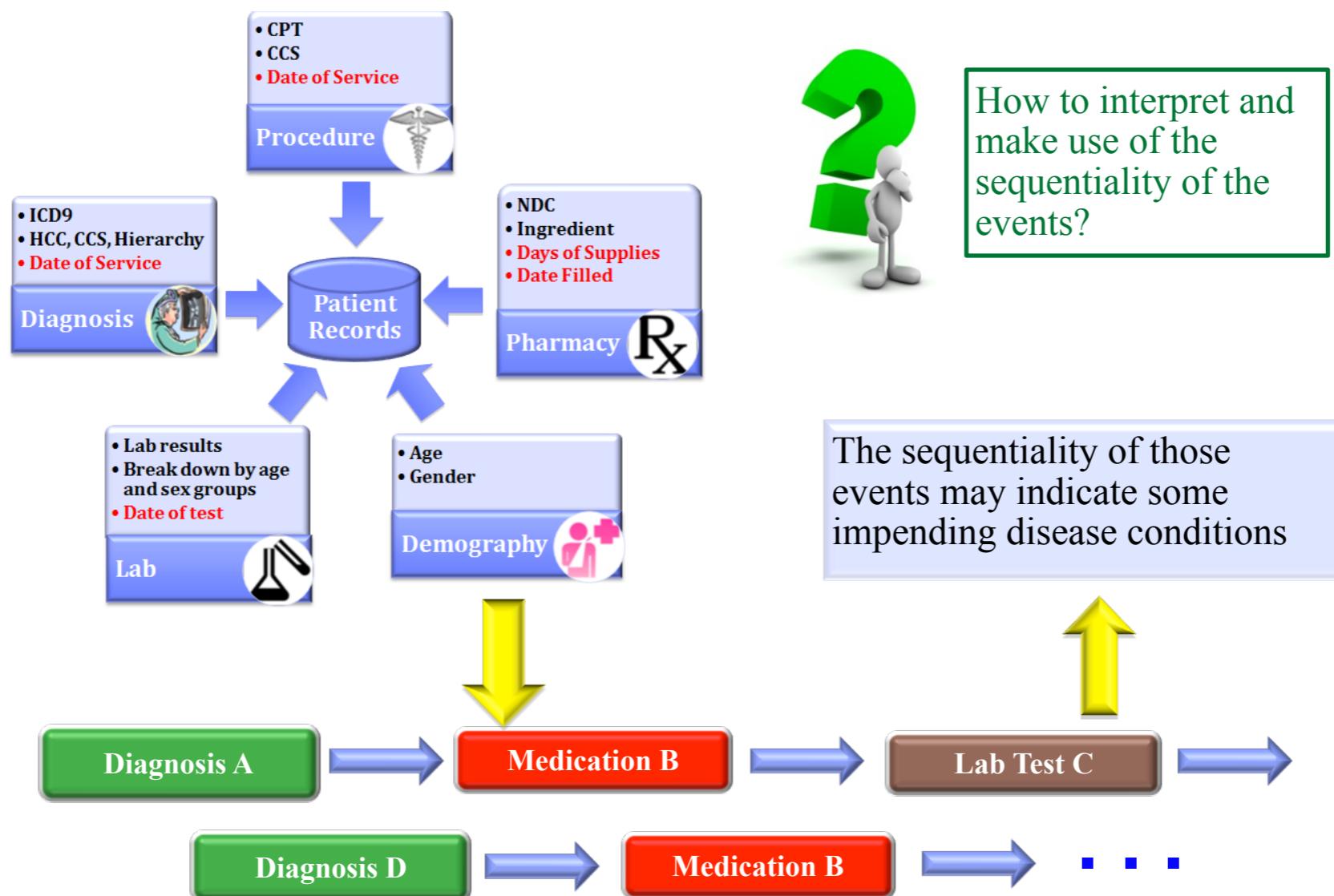
patient information

actions & advices

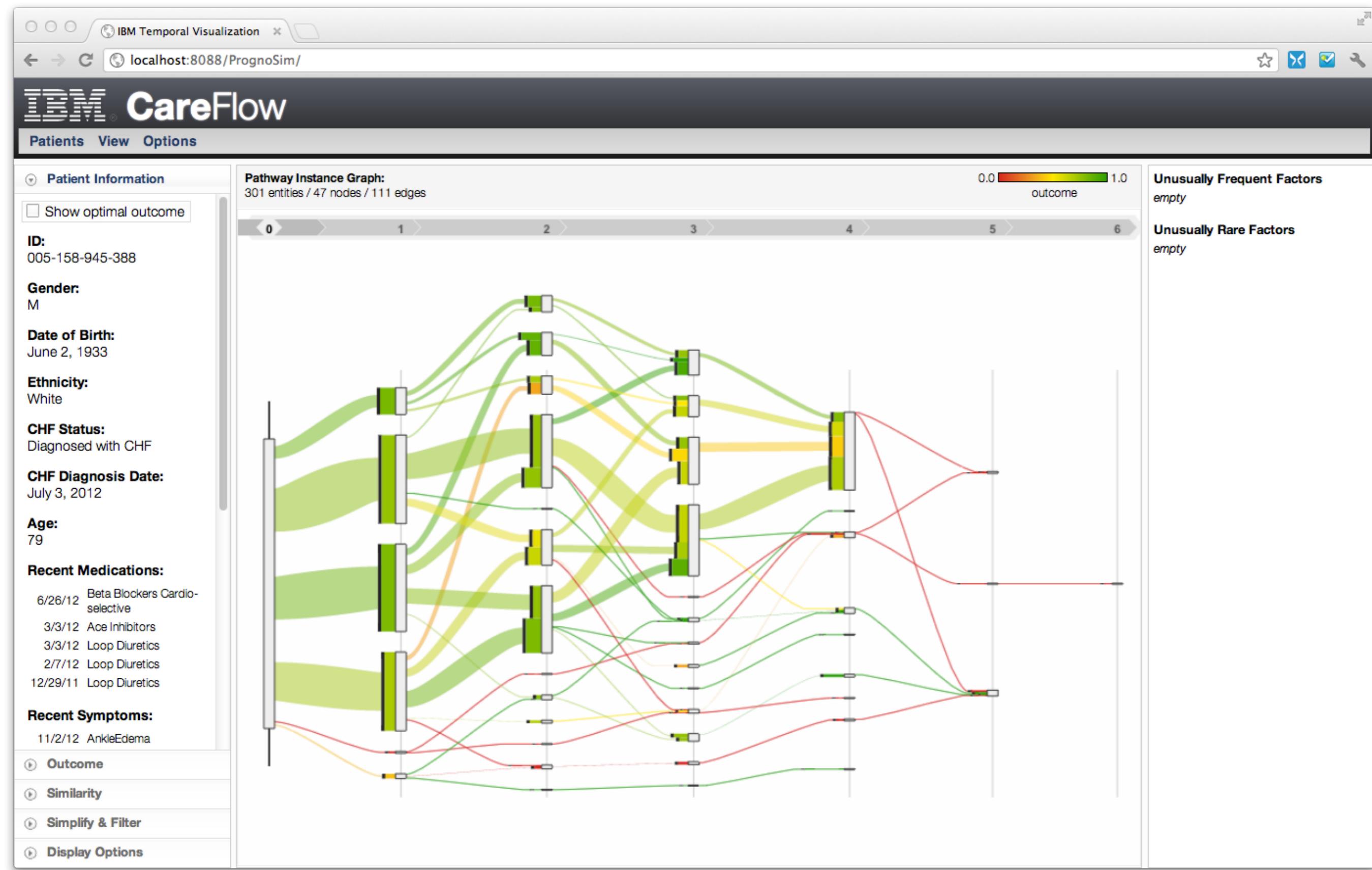
Intelligent Clinical Pathway



Sequence Representation



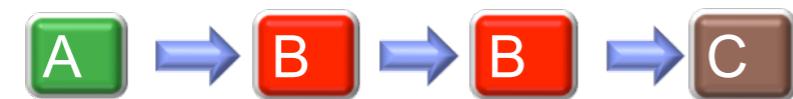
CareFlow



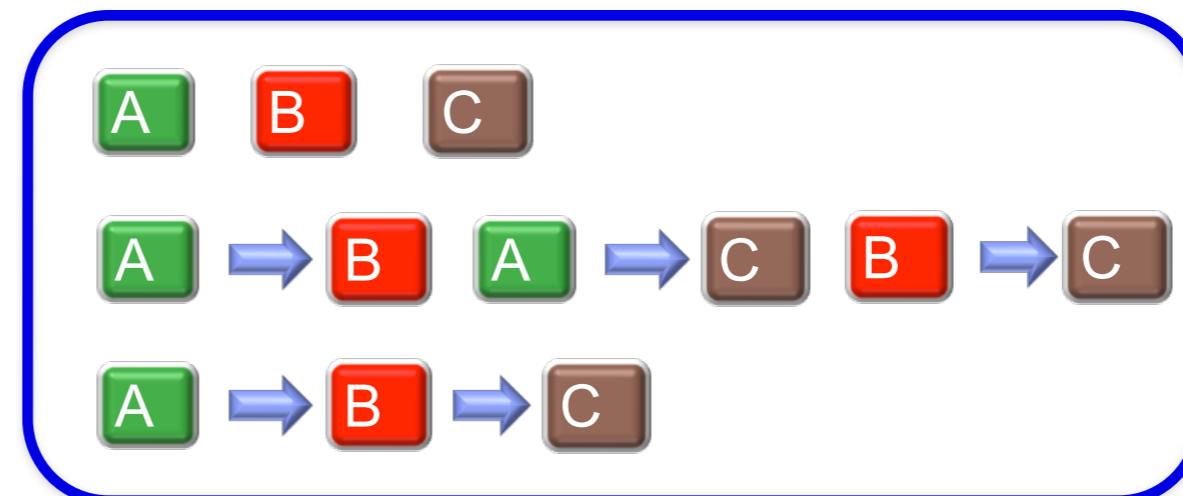
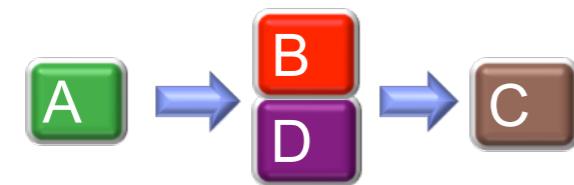
Sequential Pattern Mining

Given a set of sequences and *support* threshold, find the complete set of *frequent* subsequences

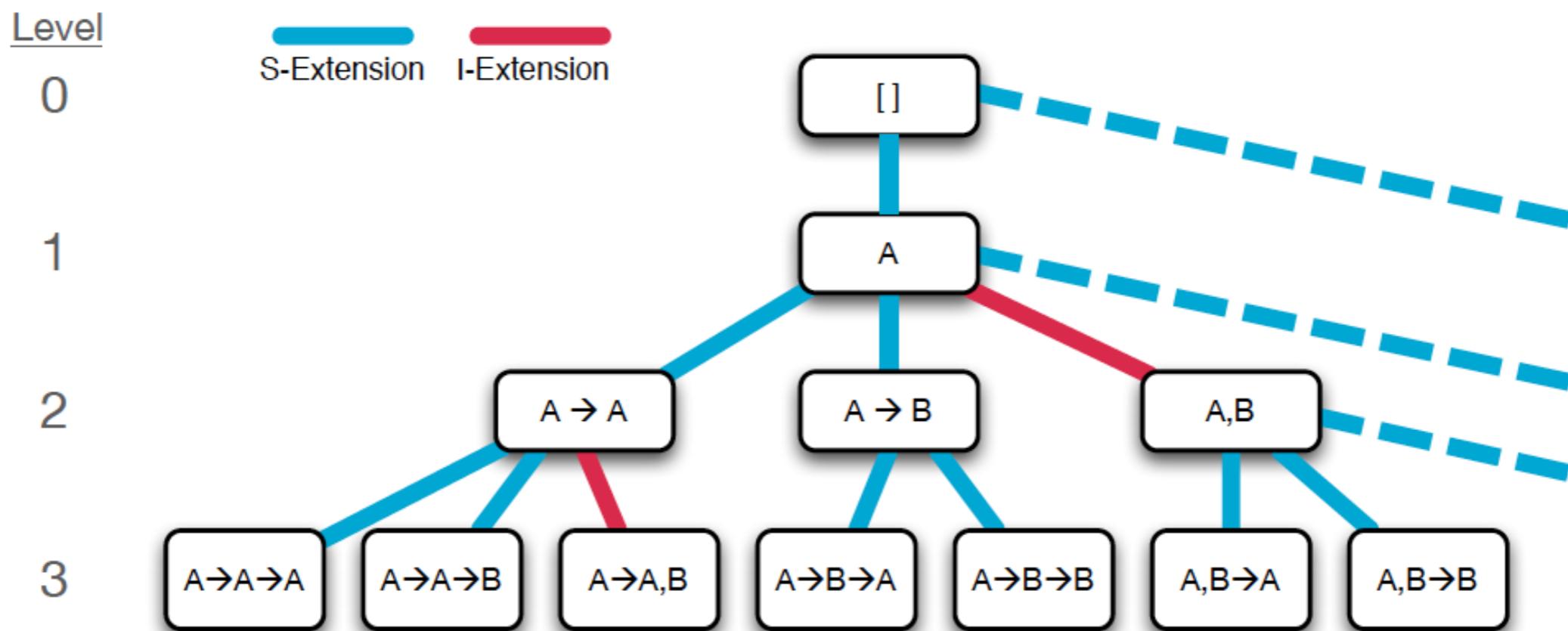
Event sequences



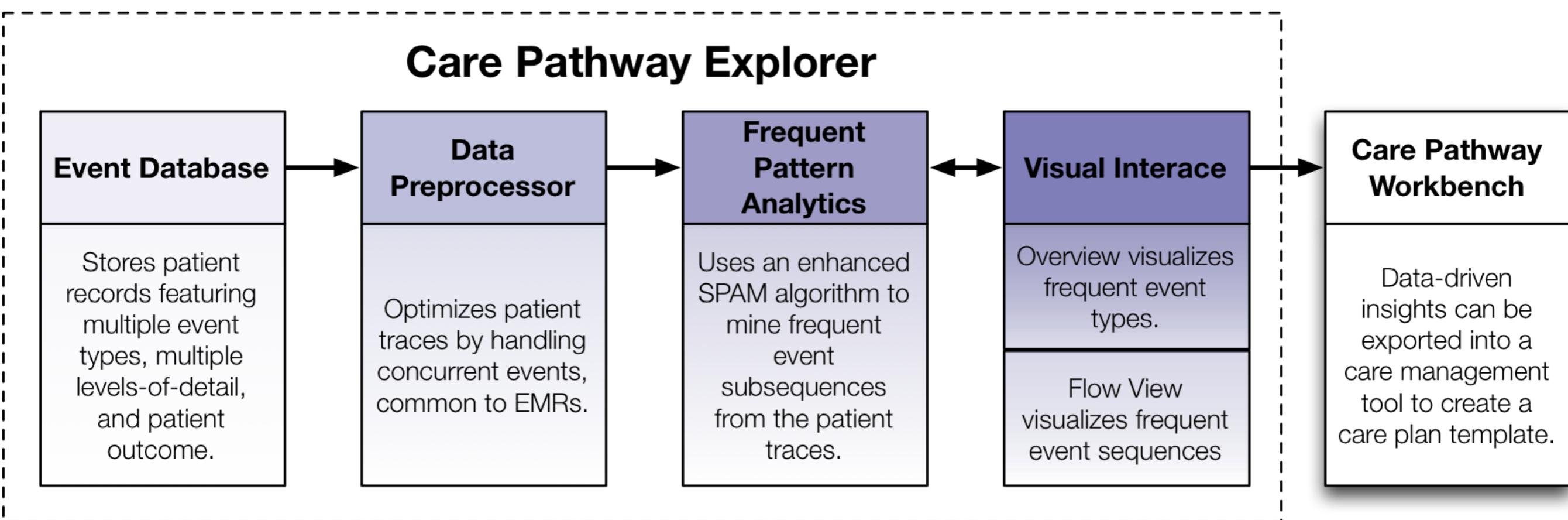
Frequent patterns with support 0.6



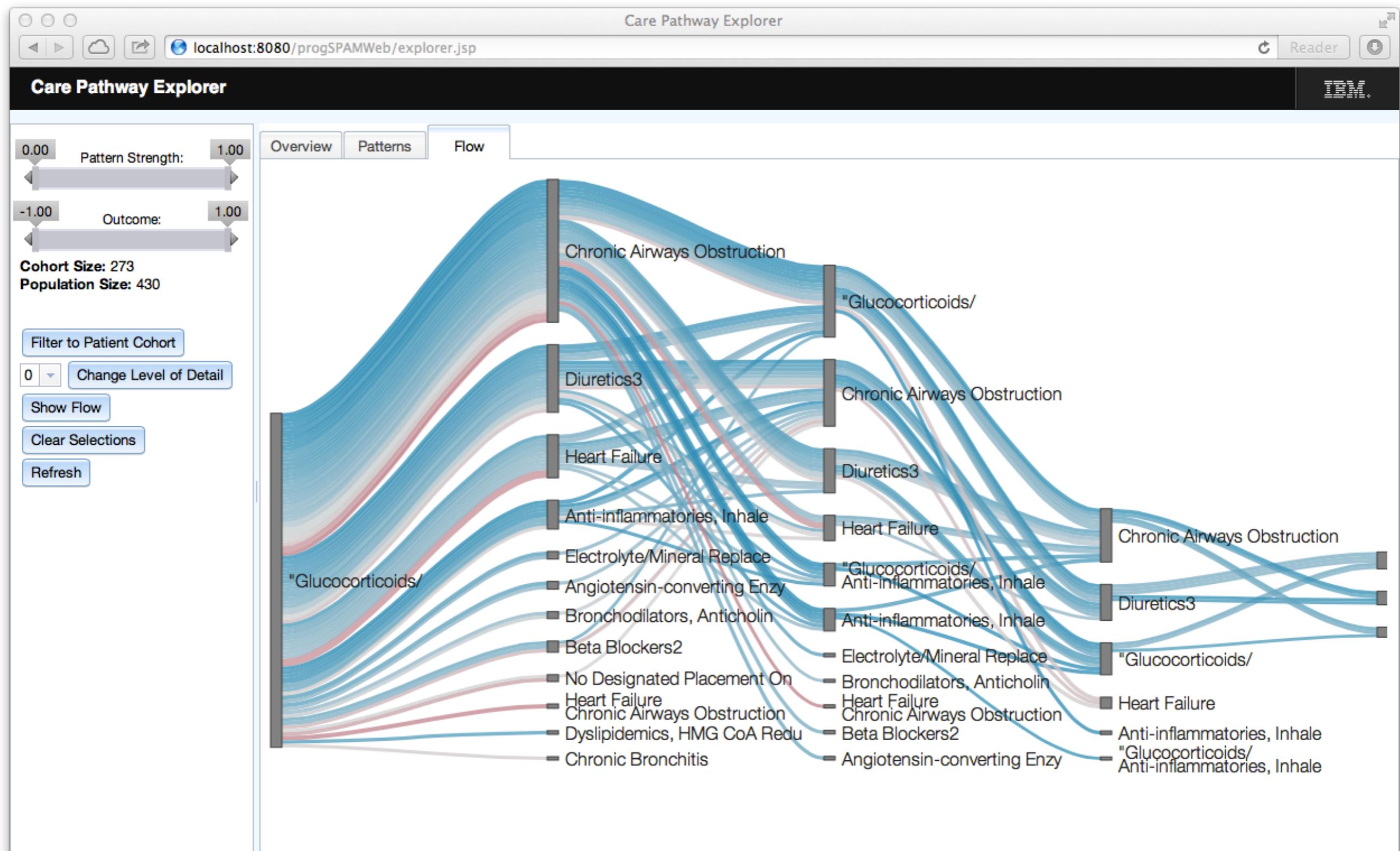
Pattern Growing



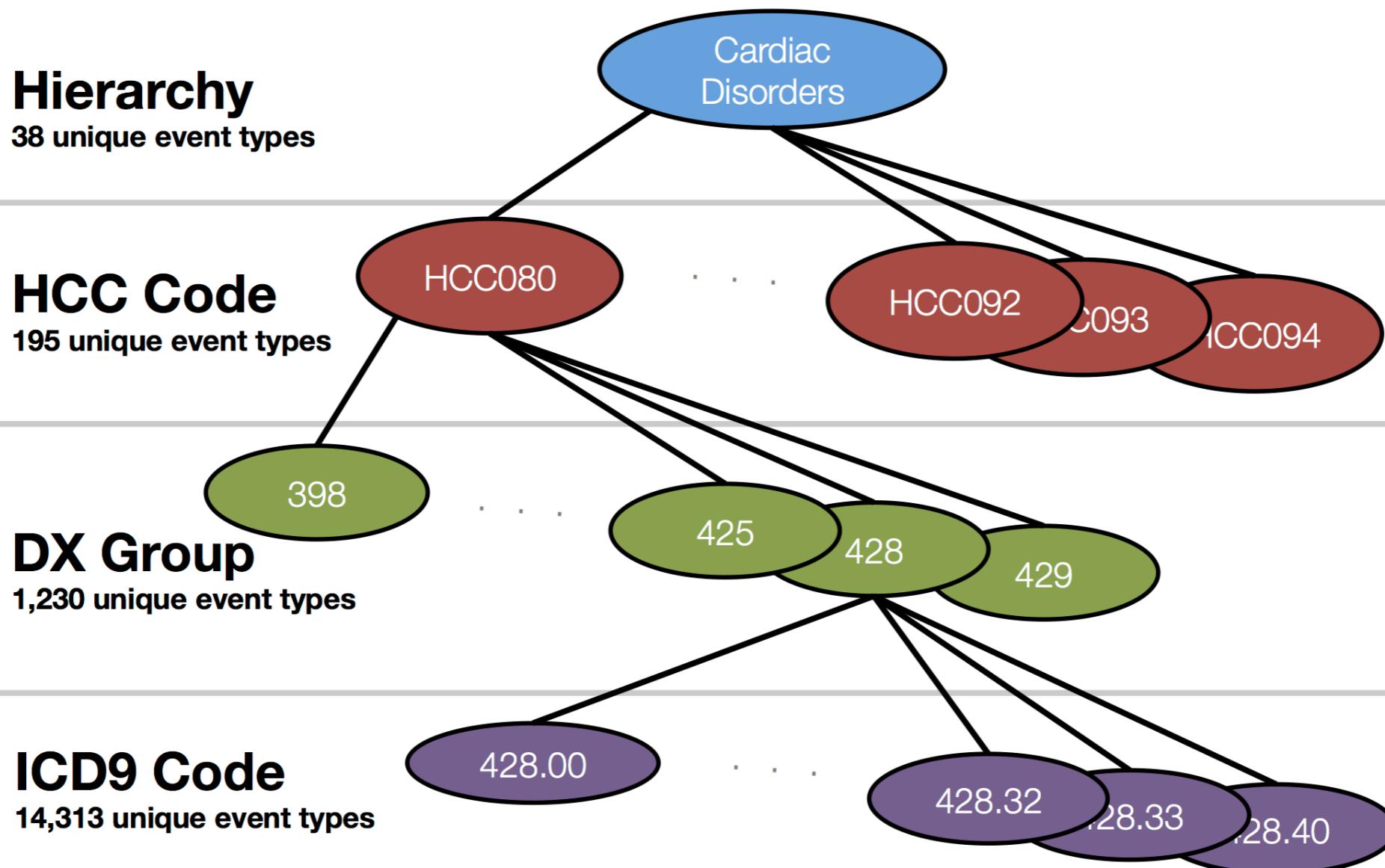
Care Pathway Explorer



User Interface



Hierarchical Exploration



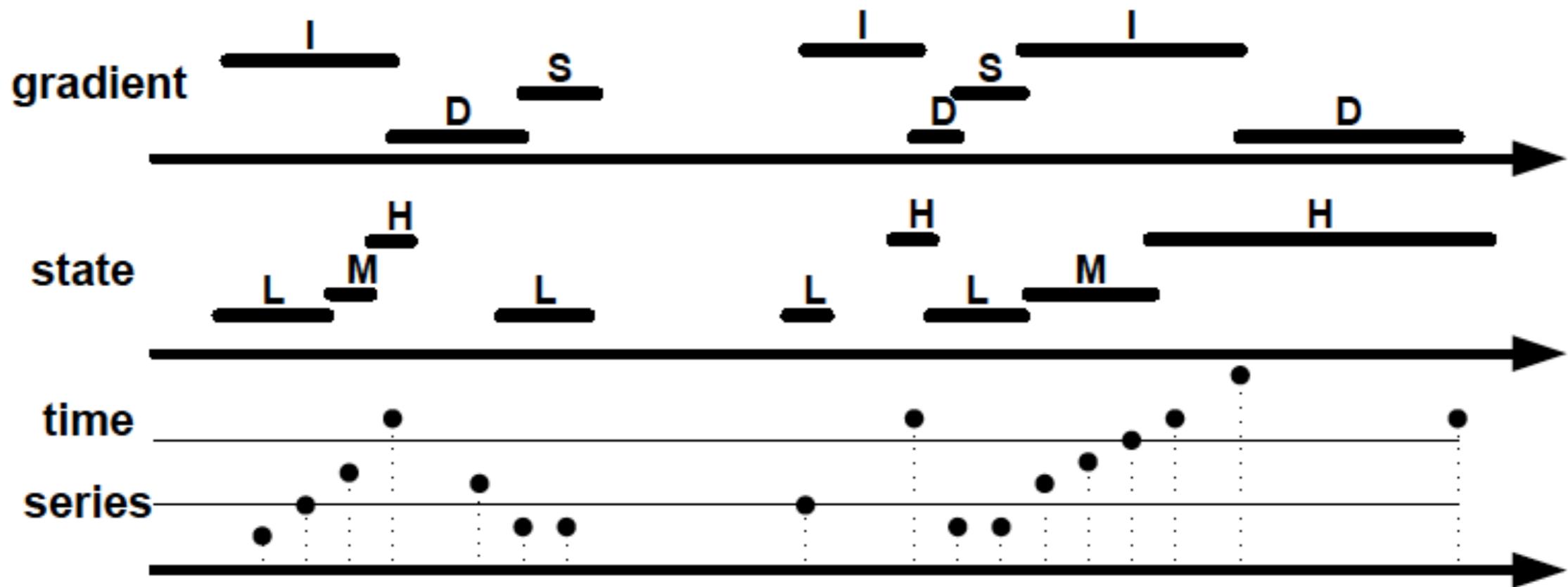
Hierarchical Exploration

Collecting all concurrent event sets, detecting frequently co-occurred event subsets from them with pre-defined support threshold

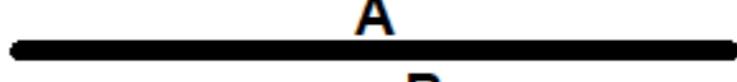
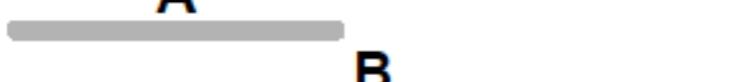
Event Package	Events
Metabolic Panel	BUN
	Creatinine
	GFR estimated
	Glucose
	Potassium
	Sodium
Hepatic Panel	HCT
	HGB

Event Package	Events
Diabetes Related Procedures	SUP-BLOOD GLUCOSE TST STRIP, 50
	SPRING-PWRD DEV FOR LANCET,EACH
Diabetes Related Diagnosis	SUP-LANCETS,PER BOX
	DIABETES MELLITUS
	DISORDERS OF LIPOID METABOLISM

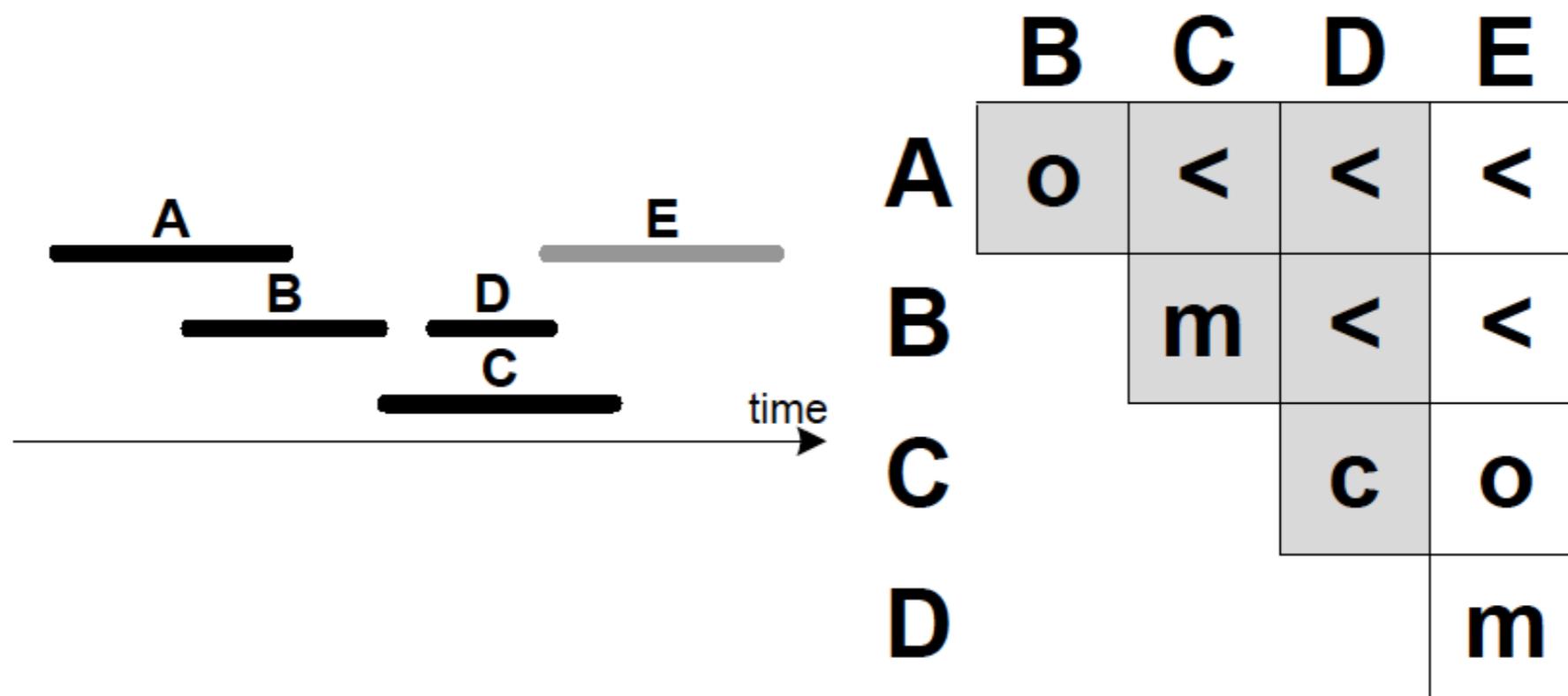
Interval Representation



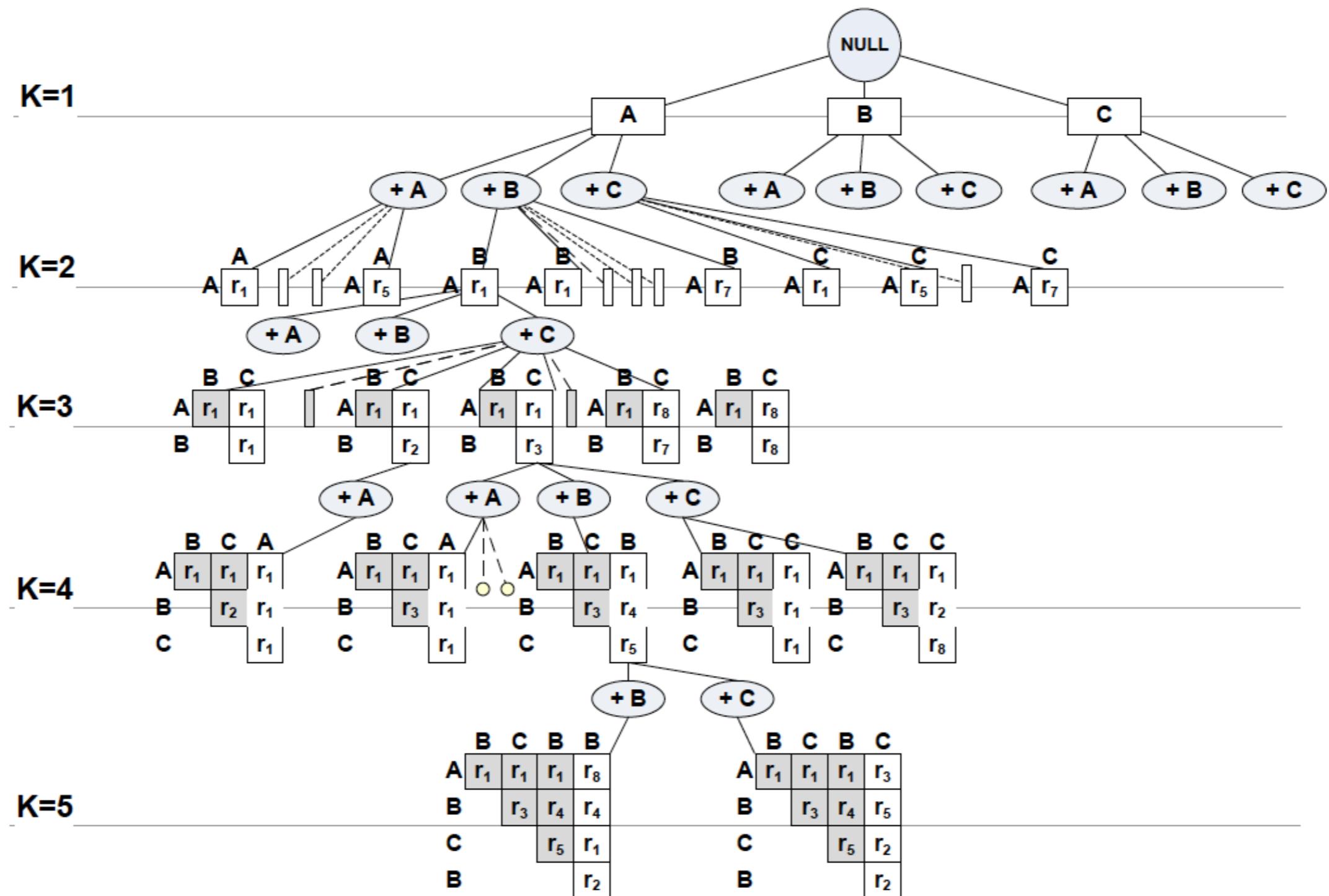
Allen's Event Relations

before (<)		$B.s - A.e > \varepsilon \wedge B.s - A.e < \text{max}$
meets (m)		$ B.s - A.e \leq \varepsilon$
overlaps (o)		$B.s - A.s > \varepsilon \wedge A.e - B.s > \varepsilon$
contains (c)		$B.s - A.s > \varepsilon \wedge A.e - B.e > \varepsilon$
finish-by (fi)		$B.s - A.s > \varepsilon \wedge B.e - A.e \leq \varepsilon$
equal (=)		$ B.s - A.s \leq \varepsilon \wedge B.e - A.e \leq \varepsilon$
starts (s)		$ B.s - A.s \leq \varepsilon \wedge B.e - A.e > \varepsilon$

Time Interval Related Pattern



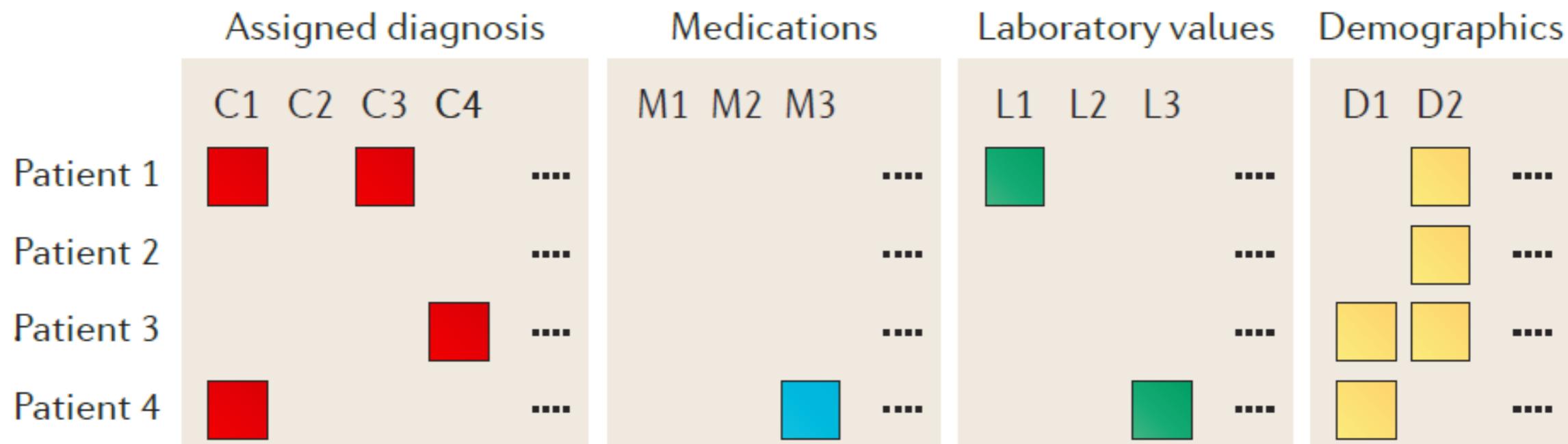
KarmaLego Enumeration Tree



Roadmap

- Background
- Healthcare Data
- Patient Similarity Analytics
- EMR Densification
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Longitudinal Information is Important



PID	DAY_ID	CLINICAL_EVENT	OP_DATE	ICD9_LONGNAME
000000	74053	305.1	74726	Tobacco Use Disorder
000000	74053	496	74726	Chronic Airway Obstruction, Not Elsewhere Classified
000000	74053	733	74726	Osteoporosis, Unspecified
000000	74053	724.2	74726	Lumbago
000000	74091	733	74726	Osteoporosis, Unspecified
000000	74148	733	74726	Osteoporosis, Unspecified
000000	74148	782.3	74726	Edema
000000	74148	780.79	74726	Other Malaise And Fatigue

Chronic Disease

Reported Cases of Common Chronic Diseases 2003	
Millions (As percent of population)	
Cancers:	10.6 (3.6%)
Diabetes:	13.7 (4.7%)
Heart Disease:	19.1 (6.6%)
Hypertension:	36.8 (12.6%)
Stroke:	2.4 (0.8%)
Mental Disorders:	30.3 (10.4%)
Pulmonary Conditions:	49.2 (16.9%)
Total Reported Cases:	162.2 (55.8%)

United States Economic Impact 2003	
(Annual Costs in billions)	
Treatment Expenditures:	\$277.0B
Lost Productivity:	\$1,046.7B
Total Costs:	\$1,323.7B

Projected Annual Costs 2023			
	Current Alternative Course	Future	Costs Avoided
Treatment Expenditures	\$790.1B	\$572.4B	\$217.6B (27.6%)
Lost Productivity	\$3,363.0B	\$2,458.0B	\$905.1B (26.9%)
Total	\$4,130.0B	\$2,996.7B	\$1,133.3B (27.4%)



Real GDP in 2050 (In billions, 2003 dollars)	
Current Course:	\$32,229.2B
Alternative Future:	\$37,897.5B
Potential Gain in GDP:	\$5,668.4B (17.6%)

Disease Progression Modeling

- Chronic disease is a global burden
 - Hundreds of millions of people
 - Trillions of dollars spent
 - Loss in life expectancy
 - Loss in quality of life
- Chronic disease management
 - Stabilize disease progression
 - Avoid exacerbation
 - Reduce symptoms and comorbidities
- Disease progression modeling is instrumental
 - Comprehensive characterization of the progression stages
 - Identify the progression trajectory of individual patients
 - Provide decision support for early intervention
- Chronic Obstructive Pulmonary Disease (COPD)
 - Impacts low-income population
 - Key risk factors: smoking and air pollution
 - Causes systematic illness

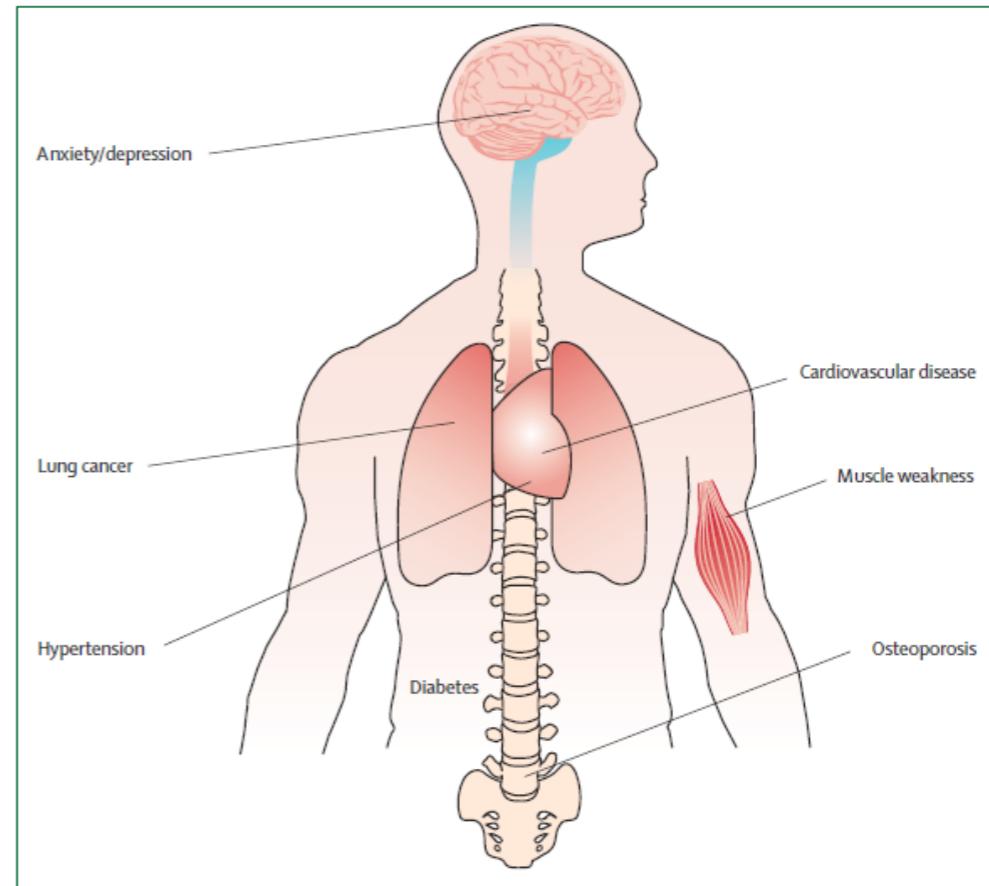
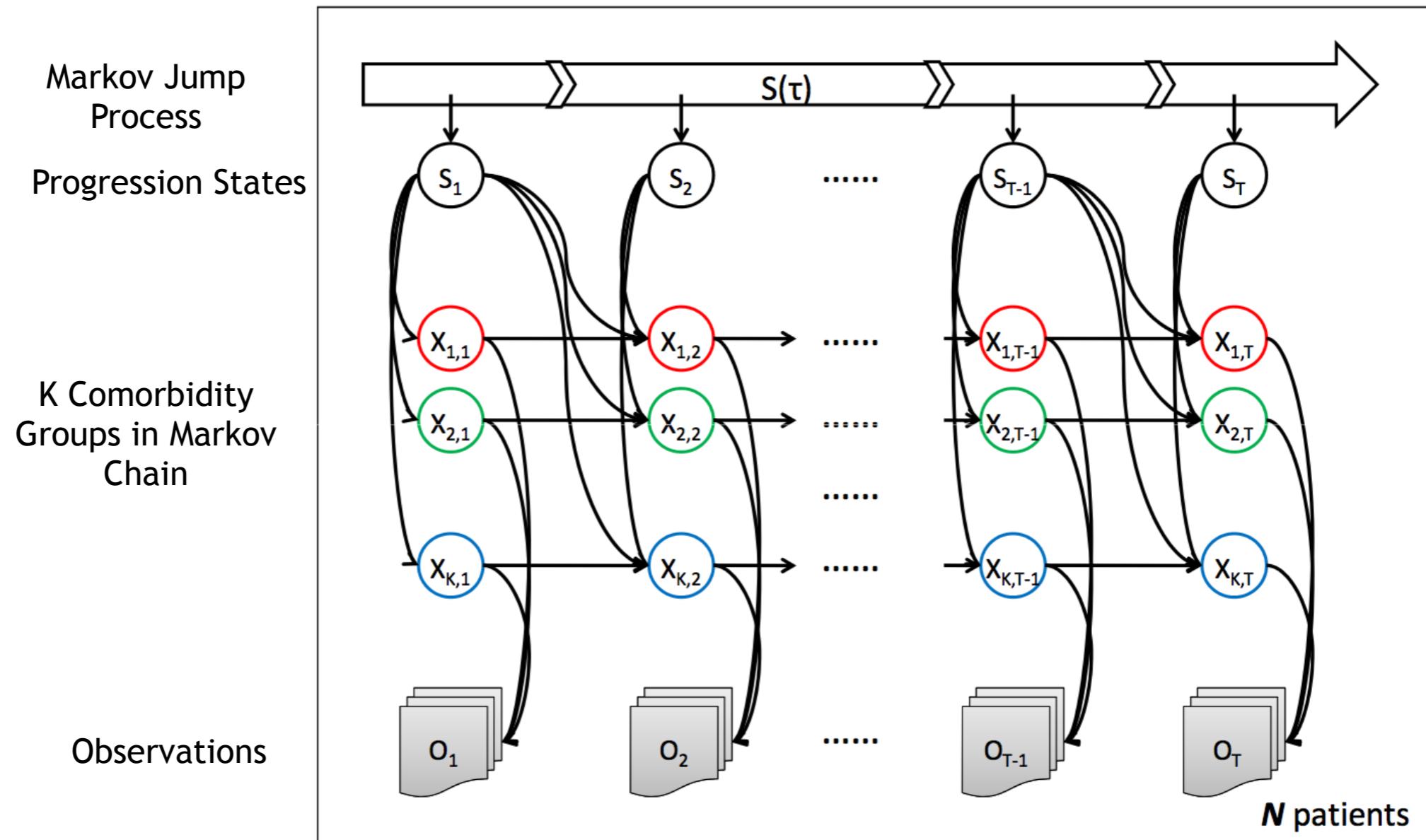


Figure 4: Comorbidities of chronic obstructive pulmonary disease

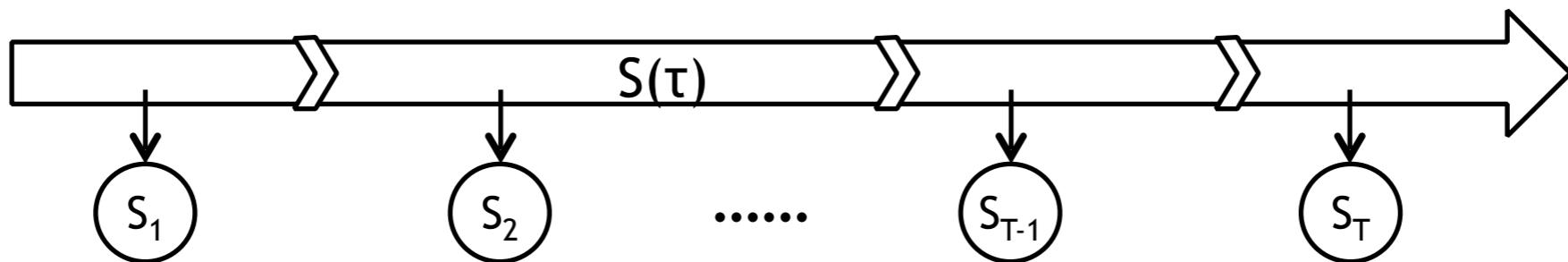
Challenges on Disease Progression Modeling

- Multiple covariates
- Progression heterogeneity
 - Disease subtypes, e.g. asthma vs. non-asthma
- Missing data
 - Doctors only document the relevant clinical context
- Incomplete records
 - Use censored records to capture lengthy progression path
- Irregular visits
 - Continuous-time model is needed
- Limited supervision
 - No ground truth regarding the current stage of progression

Our Proposed Model



Markov Jump Process

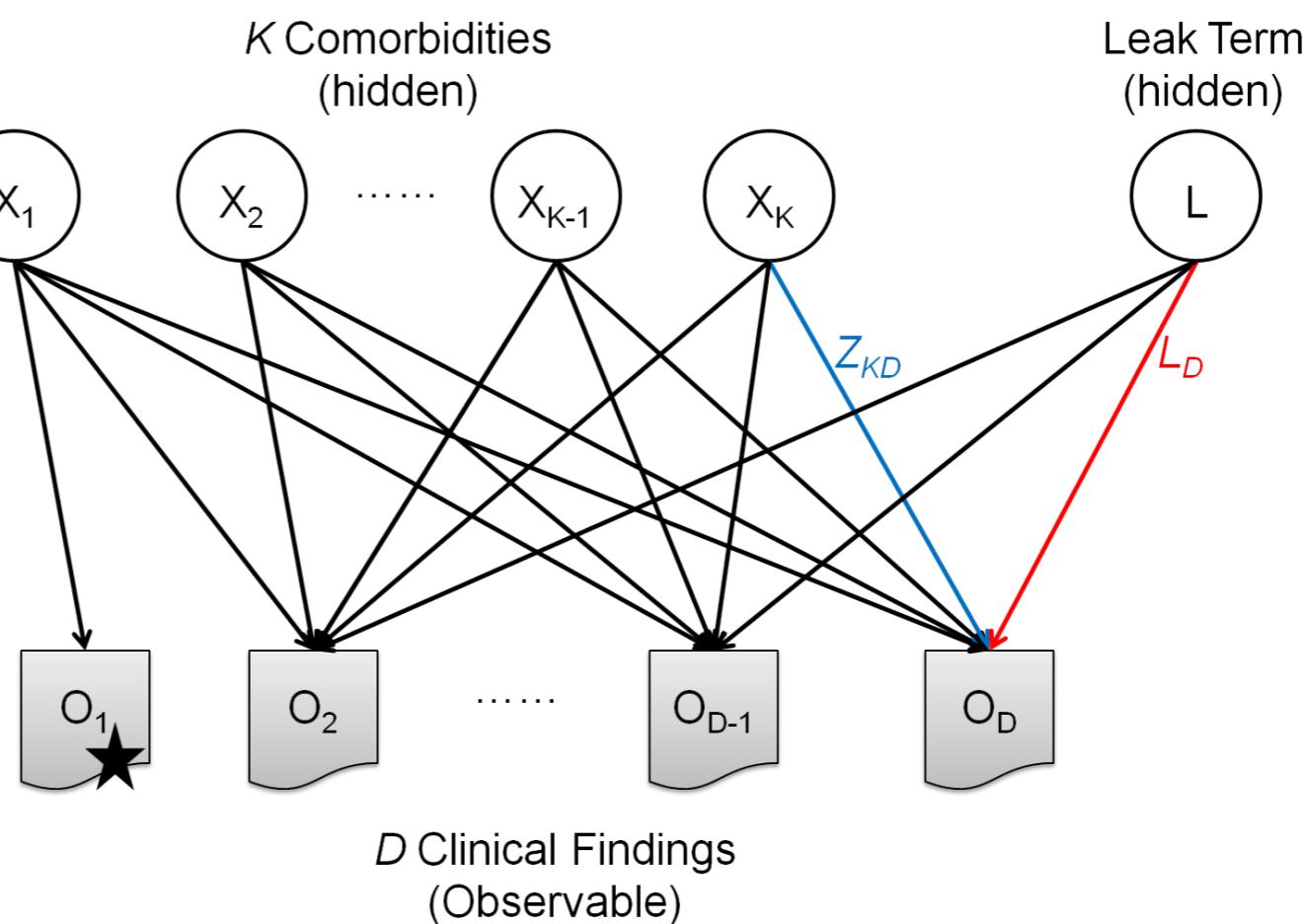
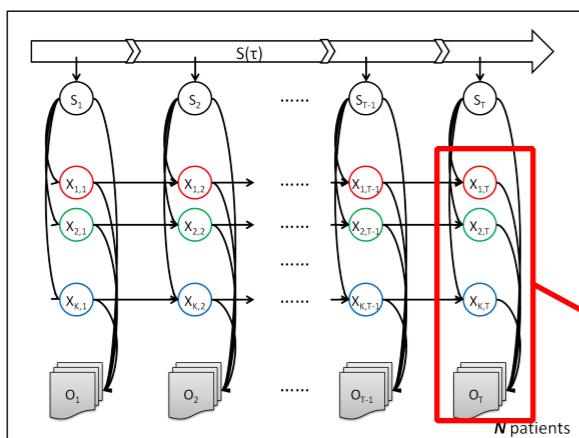


- A continuous-time Markov process with irregular discrete-time observations
- The transition probability is defined by a intensity matrix and the time interval:

$$\begin{aligned} A_{ij}(\Delta) &\triangleq P(S_t = j | S_{t-1} = i, \tau_t - \tau_{t-1} = \Delta; Q) \\ &= \text{expm}(\Delta Q)_{ij}, \end{aligned}$$

The Noisy-Or Network

Also known as QMR-DT network:
Quick Medical Reference, Decision
Theoretic



Anchored Noisy-Or Network

Use anchors findings to enable injection of domain expertise

An anchor is a medical finding that signifies the presence of a specific comorbidity

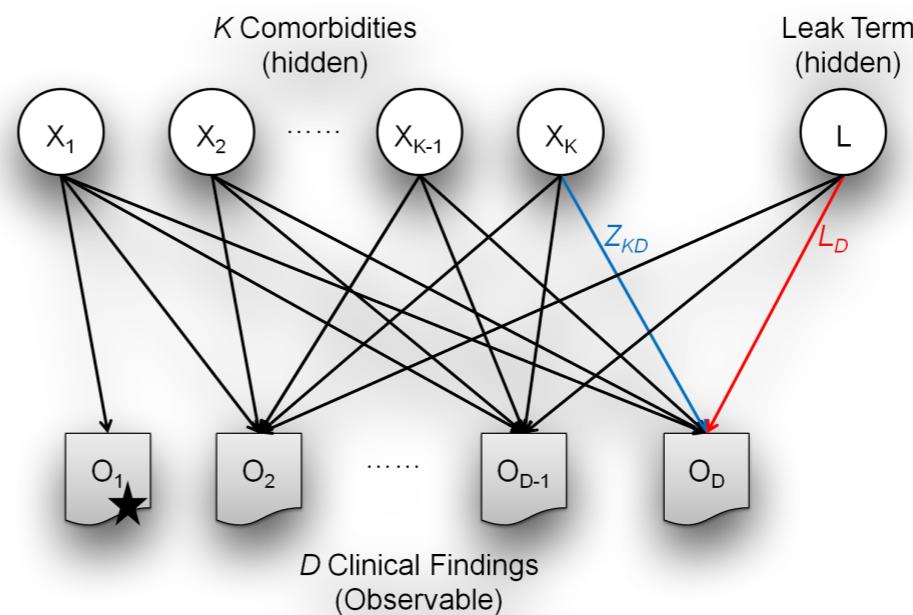


Table 2: COPD comorbidities, associated medical conditions, and anchor ICD-9 diagnosis codes used.

Comorbidity	Representative Conditions (Anchor ICD-9 Codes)
COPD	Chronic Bronchitis (491), Emphysema (492, 518), Chronic Airway Obstruction (496)
Asthma	Asthma (493)
Cardiovascular	Hypertension (401), Congestive Heart Failure (428), Arrhythmia (427), Ischemic Heart Disease (414)
Lung Infection	Pneumonia (481, 485, 486)
Lung Cancer	Malignant Neoplasm of Upper/Lower Lobe, Bronchus or Lung (162)
Diabetes	Diabetes with Different Types and Complications (250)
Musculoskeletal	Spinal Disorders (724), Soft Tissue Disorders (729), Osteoporosis (733)
Kidney	Acute Kidney Failure (584), Chronic Kidney Disease (585), Renal Failure (586)
Psychological	Anxiety (300), Depression (296, 311)
Obesity	Morbid Obesity (278)

Formalization & Implementation

Variables

Disease States (hidden): $S_{n,t} \in \{1, \dots, M\}$

Comorbidities (hidden): $X_{k,n,t} \in \{0, 1\}$

Clinical Findings (observable): $O_{d,n,t} \in \{0, 1\}$

Initial state probability

$$\pi = Pr(S_0)$$

State transition probability

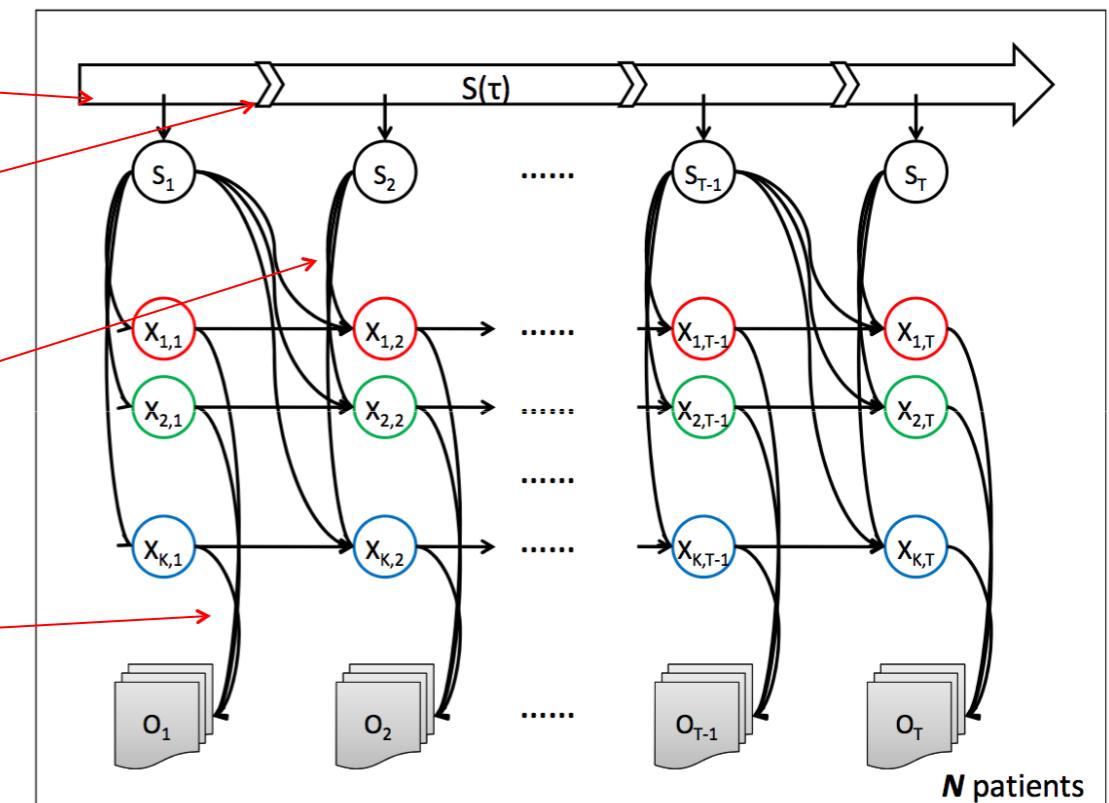
$$A(\Delta) = Pr(S_t | S_{t-1}, \tau_t - \tau_{t-1} = \Delta, Q) = \text{expm}(\Delta Q)$$

Comorbidity Onset Probability

$$B = Pr(X_t | X_{t-1}, S_t, S_{t-1}), \rho = Pr(X_0 | S_0)$$

Noisy-or Network

$$Pr(O_{d,n,t} = 1 | S, X) = 1 - (1 - L_d) \prod_k (1 - X_{k,n,t} Z_{k,d})$$



COPD Cohort Information

- Identify patients with COPD
 - At least one COPD-related diagnosis code
 - At least one COPD-related drug
 - OP_DATE is earliest occurrence of the COPD-related diagnosis
- Removed patients with too few records
- Removed ICD-9 codes that only occurred to a small number of patients
- Combined visits into 3-month time window
- 3,705 patients, 264 ICD-9 codes
- 34,976 visits, 189,815 observations

Comorbidity Groups

Asthma

Asthma*
Allergic Rhinitis
Cough
Acute Bronchitis
Acute Upper Respiratory Infections

Cardiovascular

Benign Essential Hypertension*
Unspecified Essential Hypertension*
Atrial Fibrillation*
Congestive Heart Failure*
Hyperlipidemia

Diabetes

Type II Diabetes without Complication*
Type II Diabetes without Complication, Uncontrolled*
Hyperlipidemia
Pure Hypercholesterolemia
Type II Diabetes with Renal Manifestations*

Kidney

Chronic Kidney Disease, Moderate* .
Anemia
Chronic Kidney Disease, Unspecified*
Urinary Tract Infection
Chronic Kidney Disease, Severe

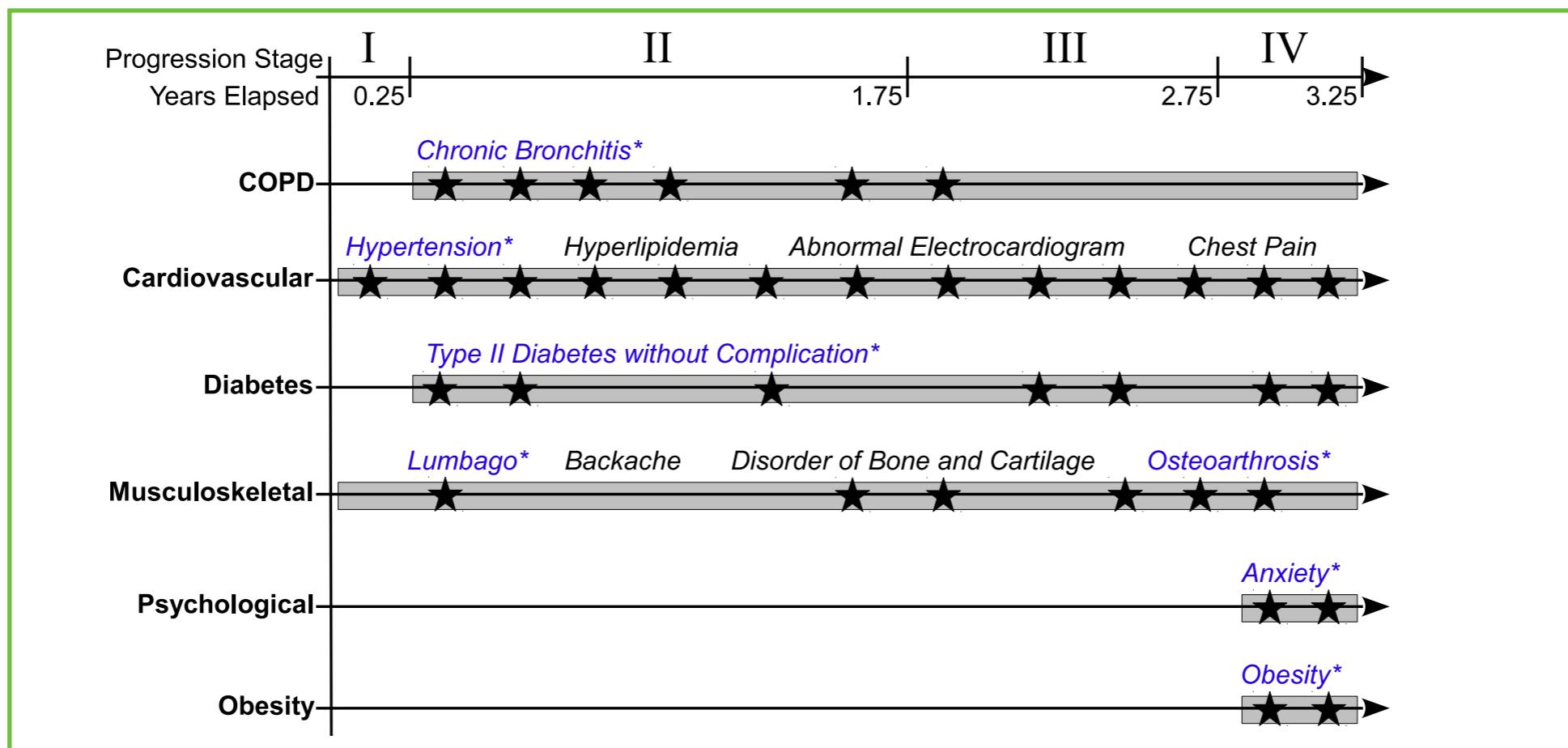
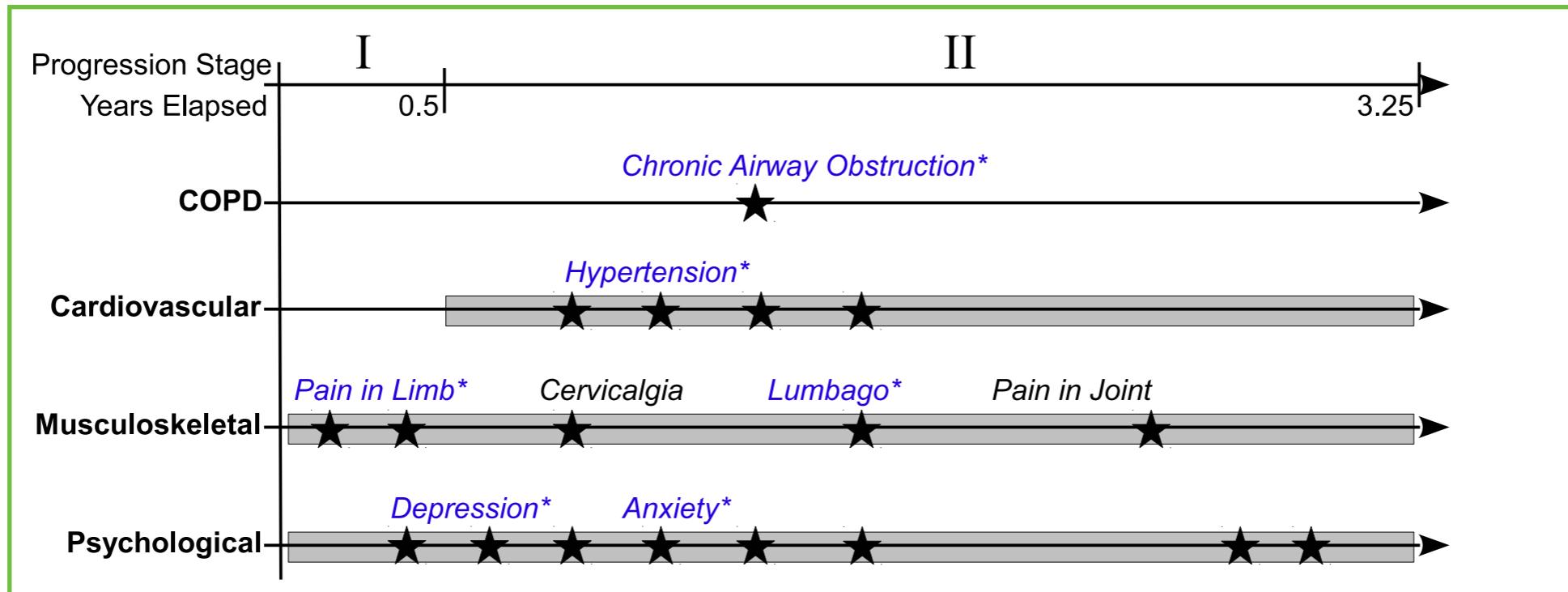
Lung Cancer

Cancer of Bronchus and Lung, Unspecified*
Other Diseases of Lung
Cancer of Other Parts of Bronchus or Lung*
Cancer of Upper Lobe, Bronchus or Lung*
Swelling, Mass, or Lump in Chest

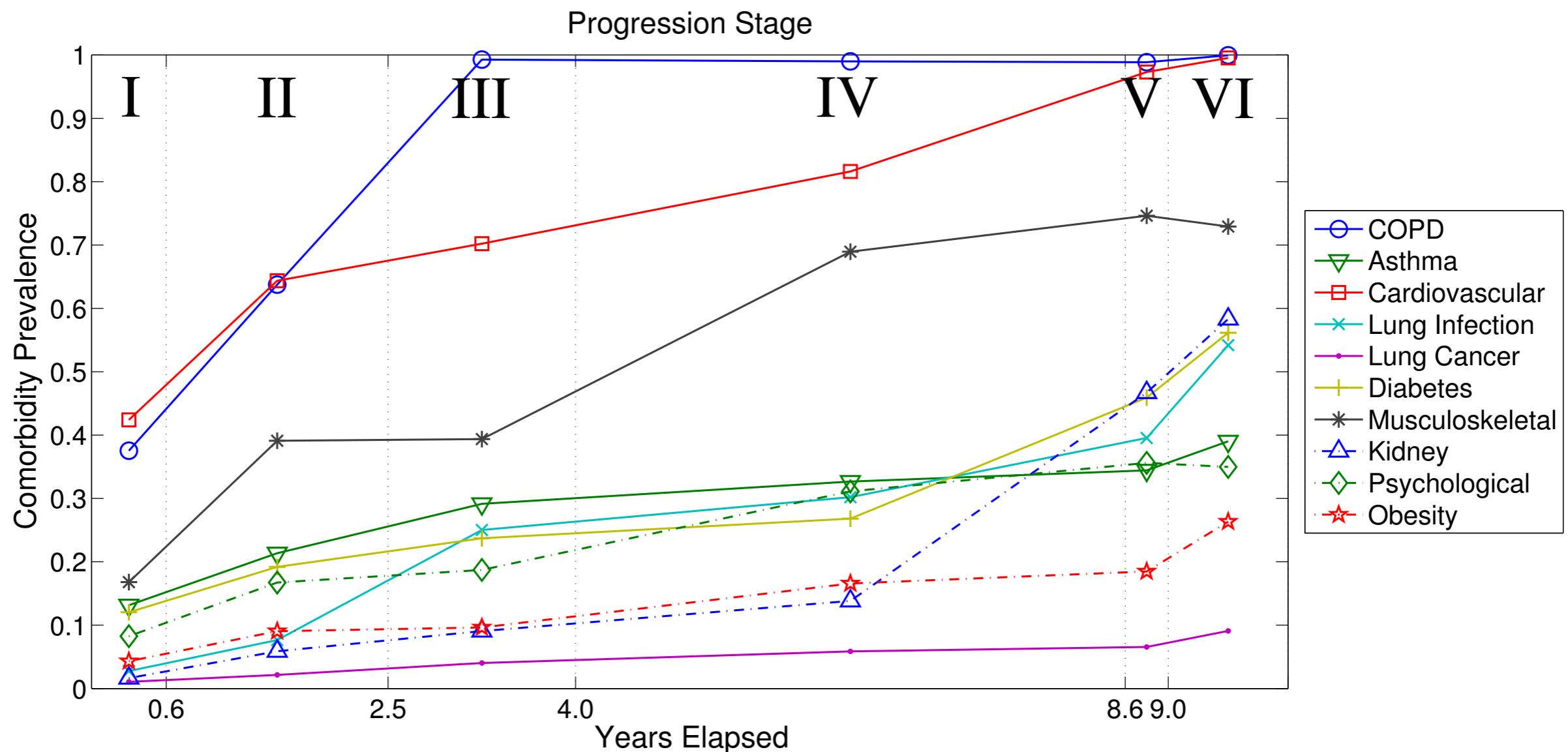
Obesity

Obesity, Unspecified*
Morbid Obesity*
Pure Hypercholesterolemia
Edema
Sleep Apnea

Inference of Individual Progression Path



Progression Rate



Roadmap

- Background
- Healthcare Data
- Patient Similarity Analytics
- EMR Densification
- Clinical Pathway Analysis
- Disease Progression Modeling
- Conclusions and Future Works

Future Directions

- Integration of heterogeneous data
 - Medical: EHR, Drug
 - Chemistry: Drug
 - Biological: Genotype
 - SocialPsycoBehavioral: Social media, Wearable device, Mobile ...
 - Environmental

Activities

- AMIA 2014 Workshop on “Data Mining in Medical Informatics: Electronic Phenotyping”. Nov.15. Full day workshop
- AAAI 2015 Tutorial on “Data Analytics with Electronic Health Records”. Jan. 25. 4 hours
- SDM 2015 The 4th Workshop on Data Mining for Medicine and Healthcare. *Pending* ...
- SDM 2015 Tutorial on Temporal Data Mining for Medical Informatics. *Pending* ...
- ICHI 2015: IEEE International Conference on Health Informatics. Oct. 2015. Dallas, Texas. Stay tuned ...

Acknowledgement

- IBM:
 - Shahram Ebadollahi
 - Jianying Hu
 - Xiang Wang
 - Ping Zhang
 - Robert Sorrentino
 - Nan Cao
 - Zhaonan Sun
 - Xinxin Zhu
 - Daby Sow
 - Harry Stavropoulos
 - Kenney Ng
 - Adam Perer
- Georgia Tech: Jimeng Sun
- NYU: David Sontag
- Stanford: Nigam Shah
- ASU: Hanghang Tong
- UMich: Jieping Ye
- UPitts: Yu-Ru Lin
- Rutgers: Hui Xiong
- UC Davis: Ian Davidson
- Temple: Zoran Obradovic
- SUNY Buffalo: Marianthi Markatou
- Columbia: Robert Moskovitch
- UTHealth: Hua Xu
- University of Maribor: Gregor Stiglic

Thank You

email: feiwang03@gmail.com