Electronic Medical Record Mining

Prafulla Dawadi

School of Electrical Engineering and Computer Science

Introduction

- "An electronic health record is a systematic collection of electronic health information about an individual patient or population."

- Big data in term of complexity, sheer volume, diversity and timeliness.

- Sources
    1. Electronic health data
    2. Ancillary clinical data
    3. Clinical text
    4. Medical imaging data
    5. Epidemiology and Behavioral data (mobility sensor data and social network data)

- Very broad research topic!

Electronic medical records

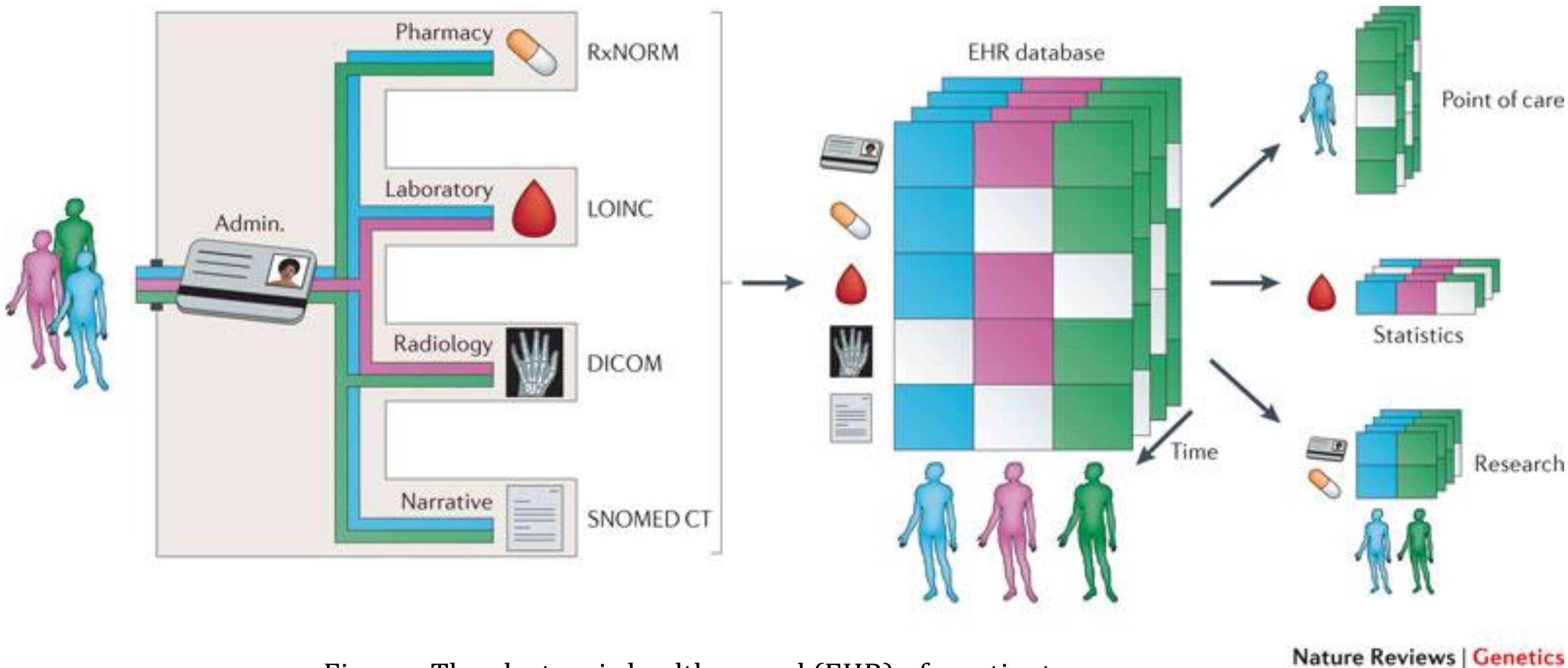


Figure : The electronic health record (EHR) of a patient.

Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." *Nature Reviews Genetics* 13.6 (2012): 395-405.

EMR mining

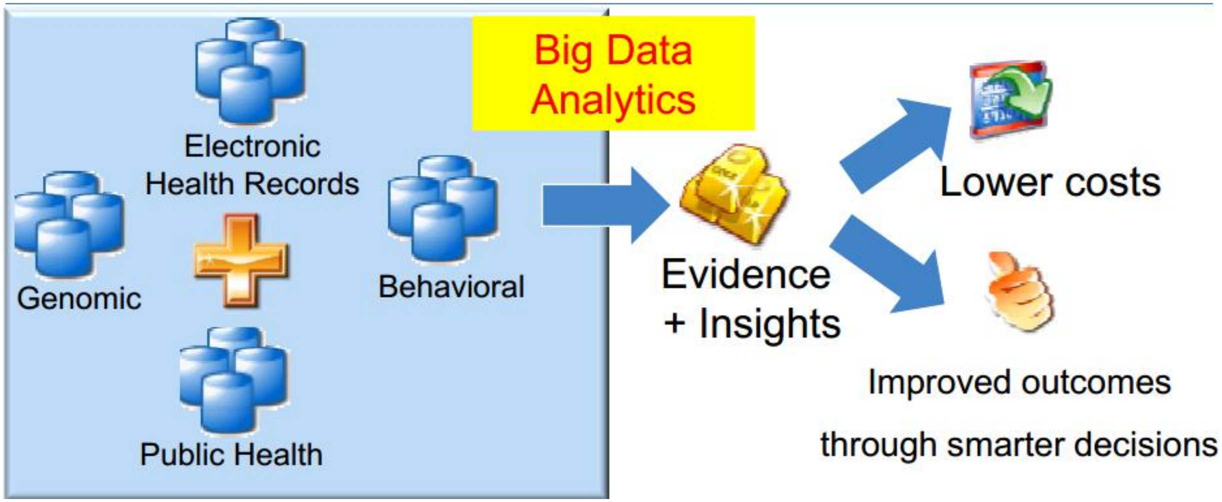


Fig: EMR mining

Example 1

- Identify high-risk patient and ensure they get the treatment they need
- Develop algorithms to predict the number of days a patient will spend in a hospital

Example 2

- Identify high rates of readmissions among patients with heart failure, heart attack, and pneumonia

extracted from Big Data Analytics for Healthcare, Jimeng Sun and Chandan K. Reddy, SDM 2013, Austin, Texas

Electronic medical record mining

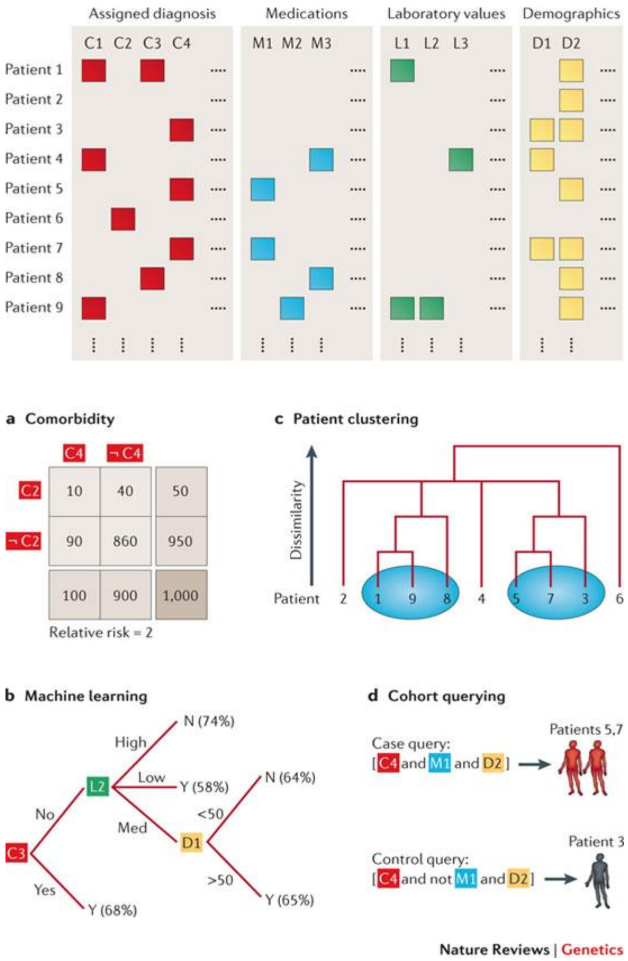


Figure: A simplified illustration of an electronic health record (EHR) research database and some of the data-driven methods.

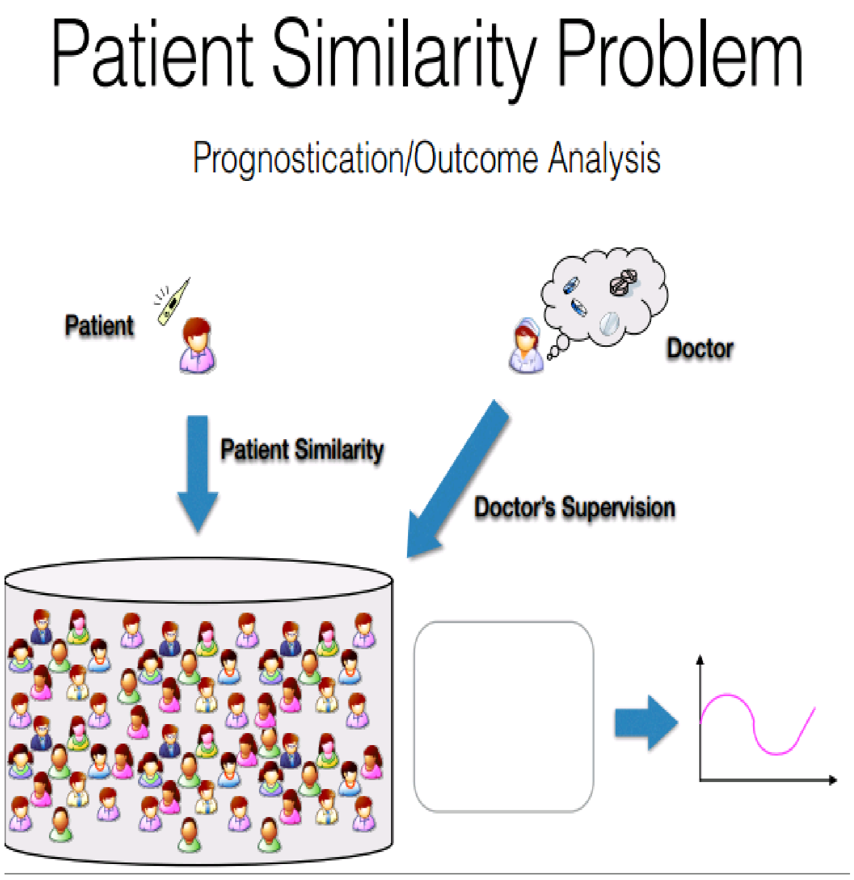
Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." *Nature Reviews Genetics* 13.6 (2012): 395-405.

Research questions

- Patient similarity analytics
- Disease progression modeling
- Personalized medication
- Integrating genetics
- Predictive modeling

Research questions

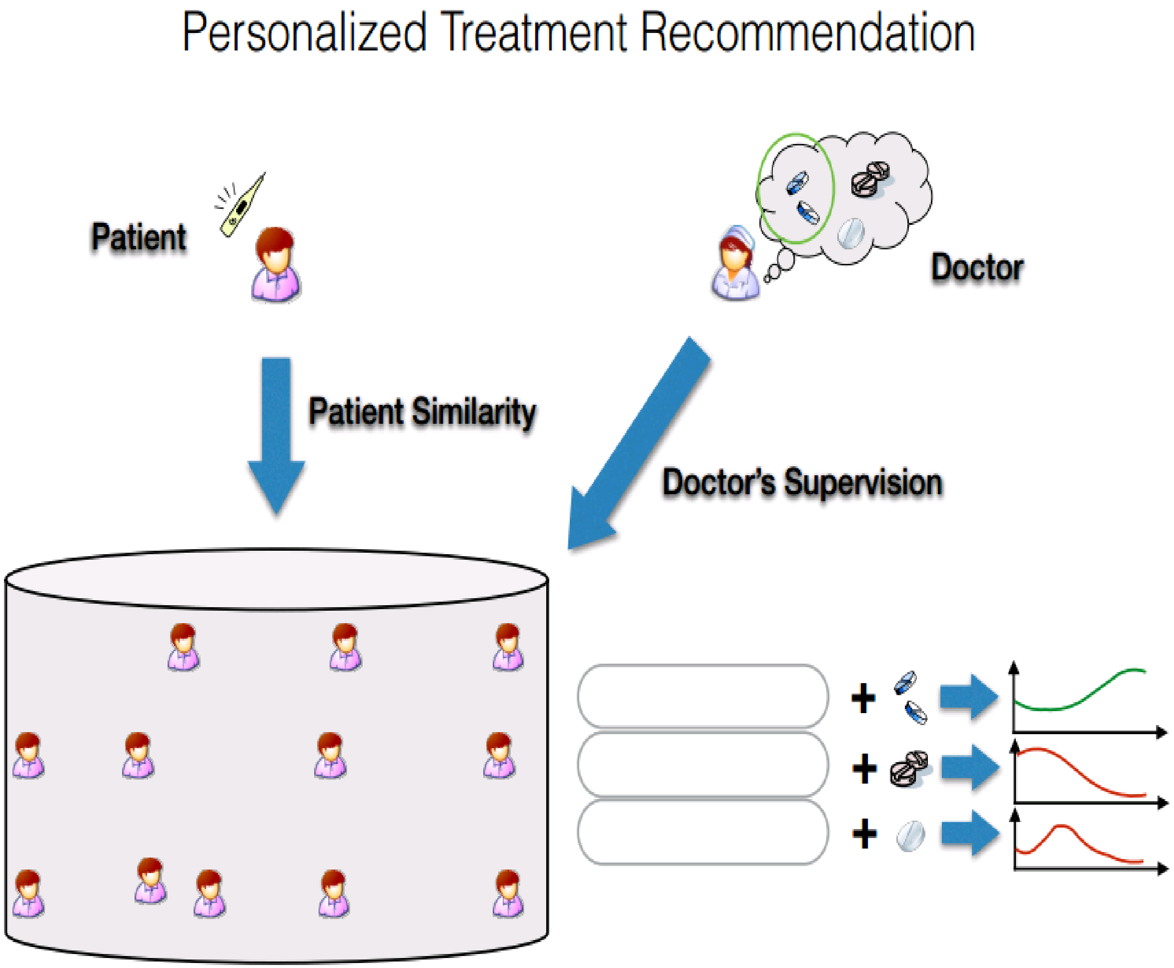
- Patient similarity analytics



Data Analytics in Healthcare: Problems, Challenges and Future Directions, Fei Wang, CIKM 2014, Shanghai, China

Research questions

- Personalized medication

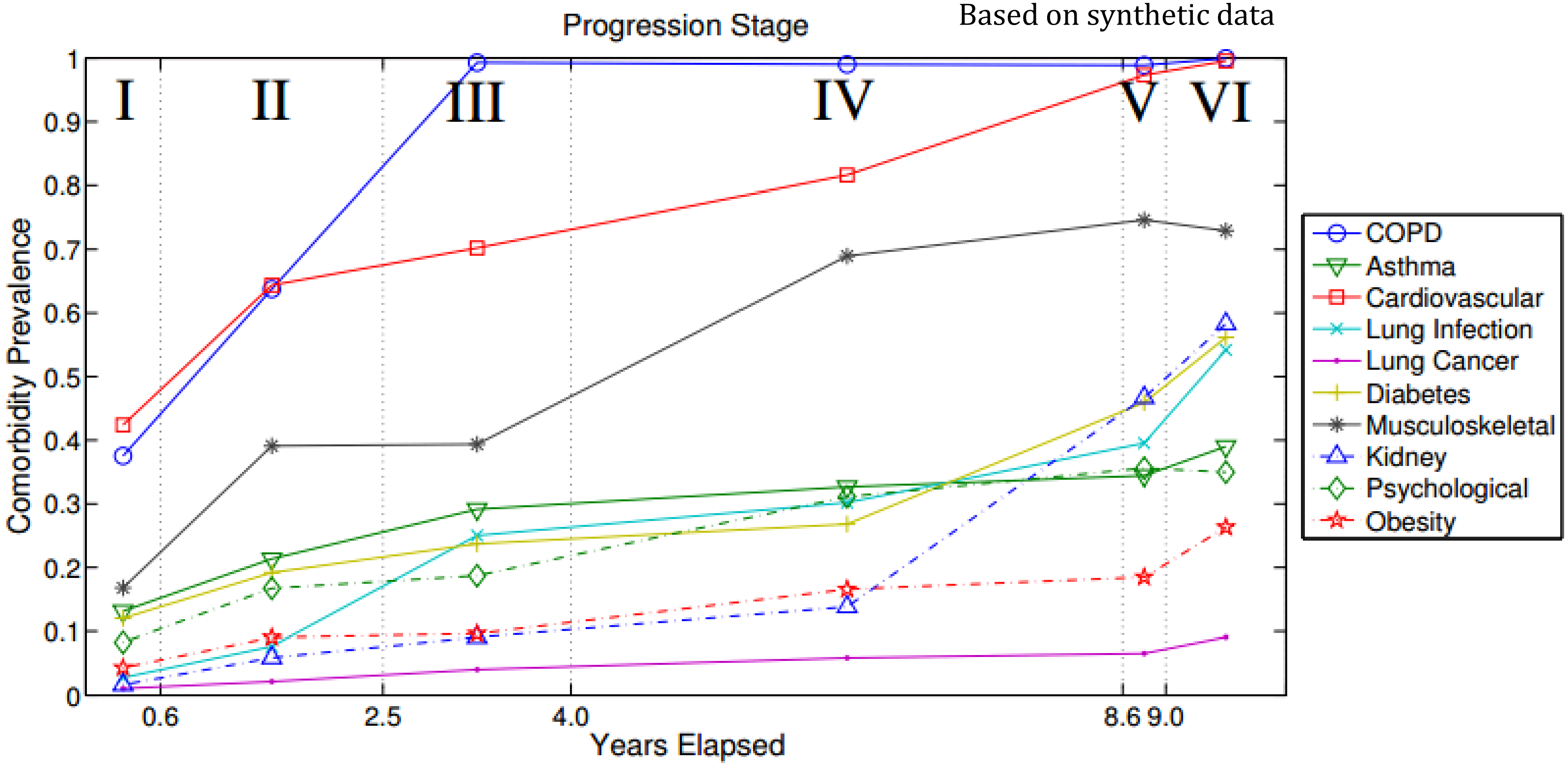


Data Analytics in Healthcare: Problems, Challenges and Future Directions, Fei Wang, CIKM 2014, Shanghai, China
Jimeng Sun, Fei Wang, Jianying Hu, Shahram Edabollahi: Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 14(1): 16-24 (2012)

Research questions

- Disease progression modeling

Based on synthetic data



Unsupervised Learning of Disease Progression Models, Xiang Wang, David Sontag, Fei Wang , KDD 2014

Research questions

- Integrating genetics
  - Identify common genetics factors that influence health and disease
  - Compare genes for people with disease and without the disease (controls)
  - **Objective** : To better understand the biological mechanisms underlying the disease

Research questions

- Predictive modeling

Marzyeh Ghassemi
Tristan Naumann
Finale Doshi-Velez
Nicole Brimmer
Rohit Joshi
Anna Rumshisky
Peter Szolovits
MIT

UNFOLDING PHYSIOLOGICAL STATE: MORTALITY MODELING IN INTENSIVE CARE UNITS

Introduction

- Use electronic health care records to identify the factors that influence patient outcomes in ICU setting.

- **Objective** : Mortality prediction in the intensive care unit.
  - Patients severity of illness is constantly evolving.
  - Data from many measurement devices.
  - Free text and clinical notes and reports → Focus of this paper.

Related work

- Clinical literature: Clinical decision rules for predicting mortality
- Use several hundred structured clinical variables to create a real time ICU acuity score that reported an ACU of 0.88 using first 24 hours of data
- Clinical notes + physiological data + discharge summaries to predict patient outcome
- Used Hierarchical Drichlet Process to nursing notes from first 24 hours for ICU patient risk stratification

Dataset

- Dataset : Multiparameter Intelligent Monitoring in Intensive Care (MIMIC) II 2.6 database
  - EMR Record : 26,870 ICU patients from 2001-2008
  - Patients age, sex, SAPS-II scores, International Classification of Diseases Ninth Revision (ICD -9) diagnosis, Elixhauser scores for 30 comorbidities as calculated from ICD-9 scores

- Target outcome : Patient mortality outcomes

- Clinical Notes:
  - All clinical notes recorded prior to the patients first discharge
    - Notes from nursing, physicians, labs, and radiology
    - Exclude discharge summaries because they state the patients outcome

Notes

- ICD stands for International Classification of Diseases (365.04 Ocular Hypertension)
- SAPS Stands for Simplified Acute Physiology Score is a type of ICU scoring systems

Vocabulary

- Tokenize the free text and remove stop words
- Use TF-IDF to find the 500 most informative words in each patients notes
- Final vocabulary was union of each patient vocabulary
  - 1 million to 285,840 words
- Exclusion criteria
    - Fewer than 100 non stop words
    - Under the age of 18

- The final training set:
  - 19,308 patients with 473,764 (24 notes per each patient)
  - 30% as a test set , 70% were training set

Features

- Structured features
  - Age, Gender, SAPS II score on admission, Elixhauser scores for 30 EH comorbidities as calculated from ICD-9 codes

- Features from topic inferences
  - Clinical notes from 12 hours windows.
    - Set of all of notes that occurred in a particular time period as features for that period.
    - Three peaks in the not times distribution for any given day in a patients stay (6:00, 18:00, and 24:00)
  - Use Latent Drichlet Allocation to generate topics for each notes
  - Derive features using topic vectors
  - Use an enrichment where the probability of mortality for each topic is calculate as

$$\theta_k = \frac{\sum_{n=1}^{N} q_{n,k} * y_k}{\sum_{n=1}^{N} q_{n,k}}$$

where y is the noted mortality out come.

Overall flow

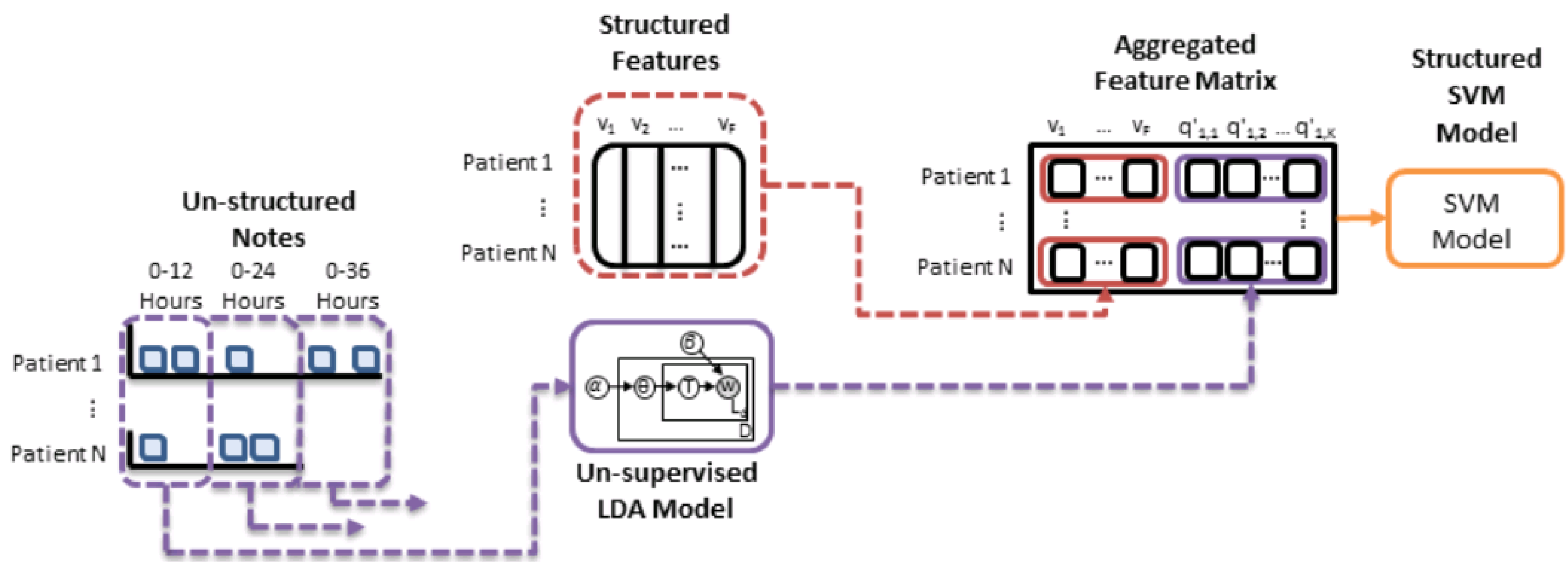


Figure: Overall flow of the experiment.

Topics

Table A.3: Top ten most probable words for all topics.

Topic Number	Top Ten Words
1	cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp
2	ccu, cath, mg, am, sp, groin, bp, cardiac, hr, cont
3	picc, line, name, procedure, catheter, vein, tip, placement, clip, access
4	biliary, mass, duct, metastatic, bile, cancer, left, ca, tumor, clip
5	liver, renal, hepatic, ascites, dialysis, failure, flow, transplant, portal, ultrasound
6	ct, contrast, pelvis, abdomen, fluid, bowel, clip, free, wcontrast, iv
7	thick, secretions, vent, trach, resp, tf, tube, coarse, cont, suctioned
8	chest, pneumothorax, tube, reason, clip, sp, ap, left, portable, ptx
9	remains, family, gtt, line, map, cont, levophed, cvp, bp, levo
10	name, neo, gtt, stitle, dr, sbp, resp, cont, wean, aware
11	remains, increased, temp, hr, pt, cc, ativan, cont, mg, continues
12	micu, code, stool, hr, bp, social, note, id, received, ccbr
13	chest, pulmonary, bilateral, edema, portable, clip, reason, ap, pleural, effusions
14	resp, cough, sats, mask, sob, wheezes, nc, status, mg, neb
15	intubated, vent, ett, secretions, propofol, abg, respiratory, resp, care, sedated
16	gtt, insulin, bs, lasix, endo, monitor, mg, am, plan, iv
17	drainage, pain, abd, fluid, draining, drain, incision, sp, intact, pt
18	heparin, afib, ptt, am, gtt, mg, rate, hr, pvcs, iv
19	name, pacer, namepattern, placement, heart, pacemaker, ventricular, av, rate, chest
20	left, lung, effusion, lobe, pleural, lower, chest, upper, ct, opacity
21	skin, noted, care, left, applied, changed, draining, coccyx, wound, edema
22	tube, placement, tip, line, portable, ap, reason, position, chest, ng
23	noted, shift, name, pt, patent, patient, foley, agitated, soft, mg
24	hct, pt, gi, blood, bleeding, am, stable, unit, bleed, noted
25	name, am, mg, able, bp, time, night, times, doctor, confused
26	pain, co, denies, oriented, neuro, plan, diet, po, pt, floor
27	name, family, neuro, care, noted, status, plan, stitle, dr, remains
28	clip, reason, ro, medical, examination, evidence, impression, underlying, condition, normal
29	neuro, sbp, bp, commands, iv, cough, soft, status, lopressor, swallow
30	skin, stable, social, family, intact, tsicu, id, note, support, endo
31	woman, female, husband, name, pain, patient, pm, am, hospital, noted
32	diagnosis, admitting, name, reason, please, examination, yearold, eval, findings, underlying
33	name, neck, soft, patient, noted, anterior, epidural, level, posterior, namepattern
34	ct, contrast, chest, lymph, optiray, images, lesions, iv, nodes, lobe
35	left, stenosis, disease, clip, reason, carotid, severe, report, radiology, final
36	femoral, foot, left, leg, iliac, groin, lower, patent, graft, extremity
37	acute, reason, head, clip, evidence, eval, name, wo, status, ct
38	aortic, aorta, cta, wwo, dissection, recons, contrast, left, aneurysm, chest
39	left, ivc, filter, vein, pulmonary, veins, dvt, clip, inferior, upper
40	left, fracture, ap, views, reason, clip, hip, distal, lat, report
41	spine, cervical, spinal, clip, thoracic, fall, lumbar, vertebral, contrast, reason
42	hemorrhage, head, ct, left, frontal, contrast, subdural, hematoma, clip, bleed
43	ct, trauma, contrast, injury, fracture, fractures, pelvis, clip, wcontrast, sp
44	contrast, brain, head, left, mri, images, mra, stroke, clip, cerebral
45	catheter, name, procedure, contrast, wire, french, placed, needle, advanced, clip
46	artery, left, common, distal, catheter, internal, branches, flow, name, middle
47	vein, stent, catheter, name, mm, portal, tips, balloon, venous, sheath
48	service, distinct, procedural, artery, sel, carotid, left, cath, name, clip
49	catheter, name, performed, embolization, contrast, bleeding, procedure, mesenteric, extravasation, clip
50	artery, carotid, left, aneurysm, injection, vertebral, internal, evidence, clip, cerebral

Topics

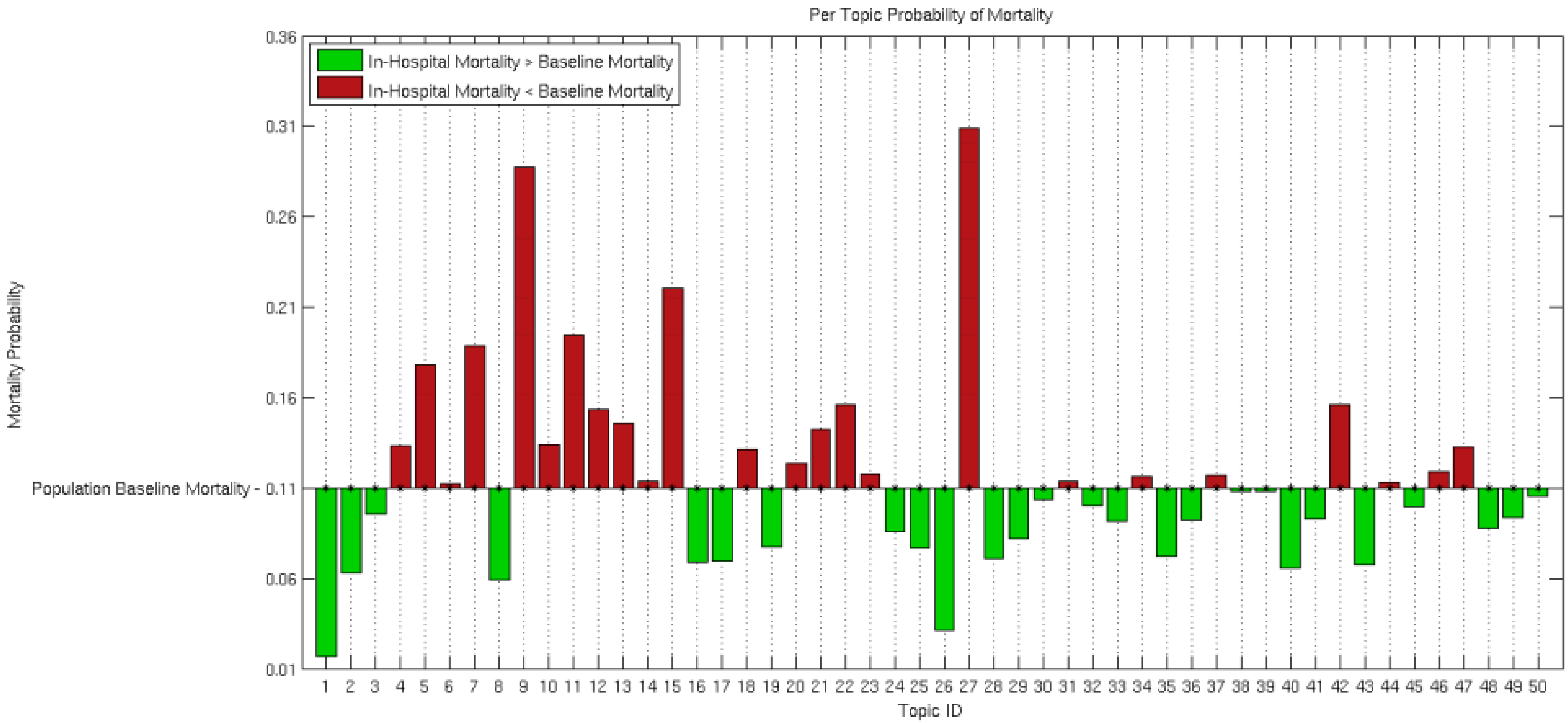


Figure: The relative distributions of the in-hospital mortality probabilities for each of the 50 topics.

The sets of topics that predict in-hospital mortality is different than 1-year post discharge mortality.

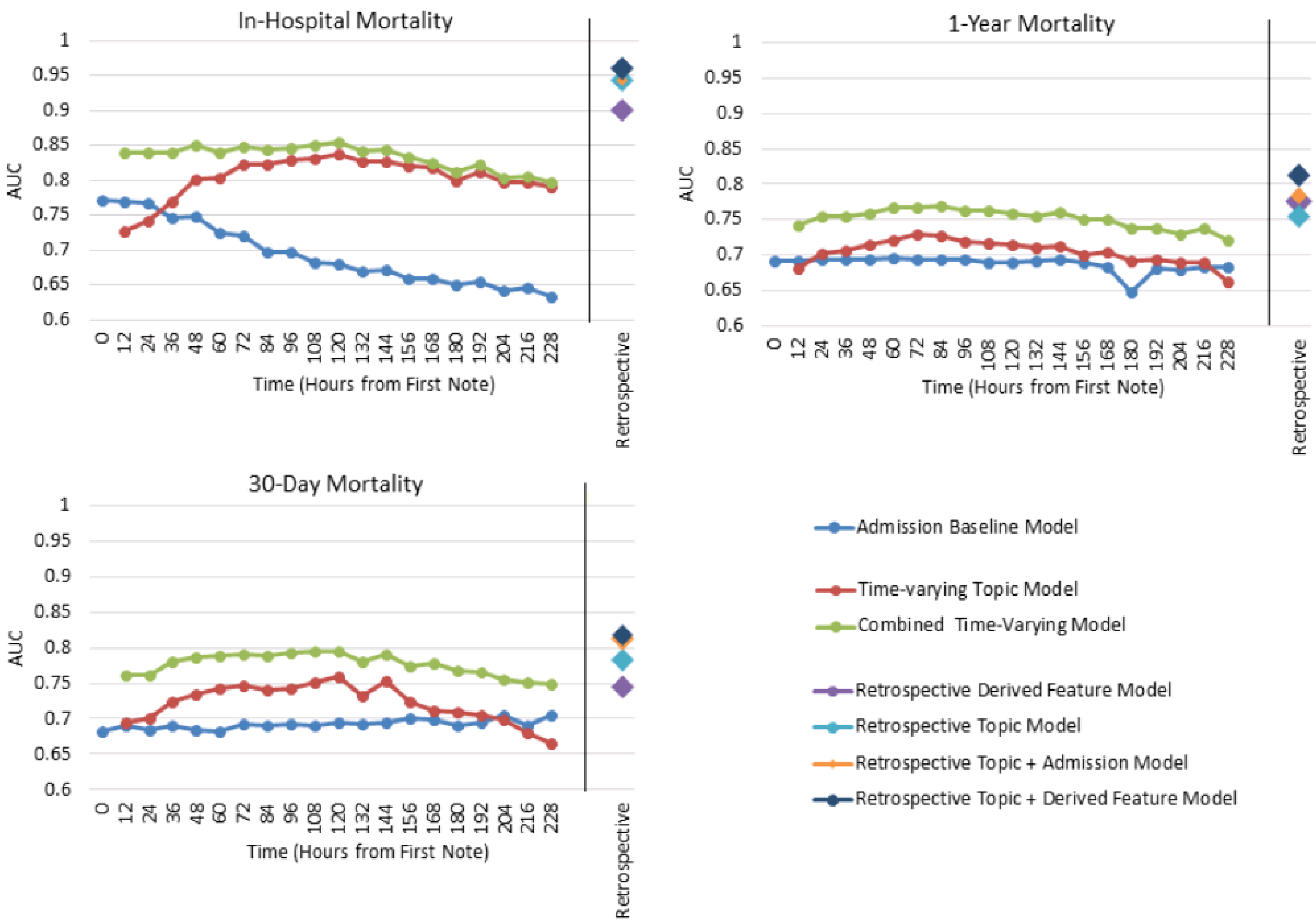
Prediction

- *Baseline prediction*: Structured features present at admission.
- *Dynamic outcome prediction*: Include larger set of patient notes in step-wise manner.
- *Retrospective outcome prediction*: Include all possible features.

Prediction settings

Models	Features
Admission baseline model	Use the structured features of age, gender, and the SAPS II score at admission.
Time varying model	Include notes in a step-wise fashion, extending the period of consideration forward by 12 hours at each step.
Combined time varying model	Time-varying Topic Model + the static structured features from Admission Baseline Model
Retrospective derived features model	A retrospective model + structured features
Retrospective topic model	A retrospective model + from all notes written during a patient's stay in the ICU.
Retrospective topic + admission model	A retrospective model combining structured features from Admission Baseline Model (gender, age, admitting SAPS scores) + latent topic features from Retrospective Topic Model. (53 features total)
Retrospective topic + Derived features	A retrospective model combining structured features from Retrospective Derived Features Model (gender, age, admitting/min/max/final SAPS scores, EH comorbidities) with latent topic features from Retrospective Topic Model. (86 features total)

Prediction results



Discussions

- Models that incorporated latent topic features were generally more predictive that those using only structured features and the combination performed the best.
- Results agree with previous set of results.
  - The first 24 hours of notes were highly relevant for the prediction.
- Predicting in-hospital mortality using admission baseline model becomes much less valuable to predict mortality as patients stay longer.
- The prediction performance of time varying models trends upward until 120 hours and then trended down.
- The predictive power of each topic changed depending on the   target outcome (1-day mortality, 30-day mortality and 1 year).

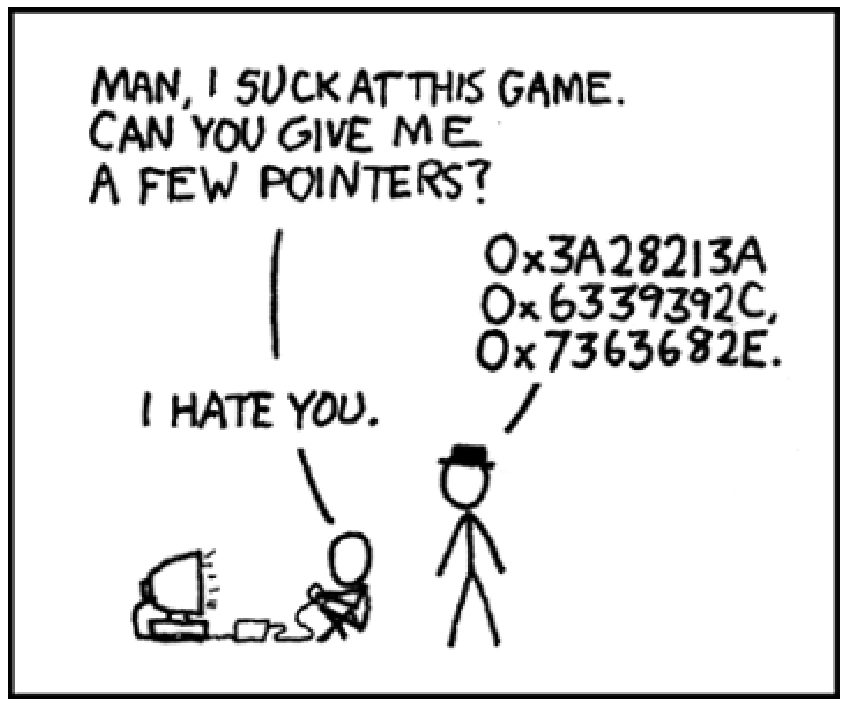
Discussions

- Dynamics of ICU patient into consideration.
- Noises in the clinical notes
- Predicting 1 year post discharge mortality
- Discussion of relationship between mortality and topics
- Age effect

Conclusions

- Augment standard clinical features with textual information in the form of topic-based features.
  - Increased performance in-hospital mortality prediction, 30-day mortality prediction and 1-year mortality prediction.
- The first 24 hours of patient information are often the most predictive of hospital mortality.

- Thank you!
- Questions



Xkcd.com

References

1. Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." *Nature Reviews Genetics* 13.6 (2012): 395-405.
2. Data Analytics in Healthcare: Problems, Challenges and Future Directions, Fei Wang, CIKM 2014, Shanghai, China
3. Big Data Analytics for Healthcare, Jimeng Sun and Chandan K. Reddy, SDM 2013, Austin, Texas
4. Jimeng Sun, Fei Wang, Jianying Hu, Shahram Edabollahi: Supervised patient similarity measure of heterogeneous patient records. SIGKDD Explorations 14(1): 16-24 (2012)
5. Unsupervised Learning of Disease Progression Models, Xiang Wang, David Sontag, Fei Wang , KDD 2014