# Health Informatics

Lecture 5

Samantha Kleinberg
samantha.kleinberg@stevens.edu

- Next week: Midterm!

# Uses for medical data

- Finding long-term risk factors for disease
- Drug-drug interactions, side effects (post-market)
- Population health

and more!

# Uniqueness of Medical Data Mining

- Can't just apply methods from data mining/ ML
- Main points
  - Heterogeneity
  - Ethical, legal, social issues
  - Statistical philosophy
  - Special status of medicine

# Heterogeneity of data

- Structured/unstructured, Imaging
  - Many data mining methods can handle only one of these
- How to combine qualitative/quantitative information?

# Heterogeneity of data

- Importance of interpretation

- If we see CHF in record – what did clinician mean?
  - Could be suspected CHF
  - Hypothesis explaining symptoms
  - Past or family history…
- If we don't see indicator for CHF does that mean patient doesn't have it?
  - May not be billed for, may not be treated (if more pressing problems)

# Heterogeneity of data

- Standardization
  - Of data (recall ICD9 for example)
  - Of outcomes
- Example:
  - One group evaluates glucose control system by calculating how often glucose is within 70-150, another group uses 80-140. How to compare?

# Heterogeneity of data

- Difficulty applying precise labels
- Uncertainty
  - Does a particular billing code mean patient definitely has illness?
  - When did the illness start?
- **Test vs diagnosis**
  - We see imperfect indicators for a disease, not the disease itself
  - For prediction, target event may be diagnosis NOT onset of disease

# Ethics, legal and social factors

- Access to data (researchers, patients)
- Concern about liability
  - Affects what data can be collected, which tests can be done
- Privacy
  - Can we use all the data we want?
  - Can we analyze in Amazon cloud? Give text to mechanical turkers? Scrape data from password protected websites?

# HIPAA

Recapping your CITI training…

Health Insurance Portability and Accountability Act

Includes definitions for protected health information (PHI), and what can be shared and with whom

- Who's covered by HIPAA?
  - Ex: healthcare provider, researchers working with PHI from hospital
- What's required?
  - Usually need consent, unless waiver from IRB or meet certain other criteria
- Key component: data de-identified and you don't have reason to believe that they can be re-identified. If this satisfied, no longer covered by HIPAA

# 18 HIPAA identifiers

1. Names
2. Certain geographic information*
3. Dates other than year; all ages over 89 and all elements of dates (including year) indicative of such age, except when aggregated into a category of age 90 or older;
4. Phone numbers;
5. Fax numbers;
6. Electronic mail addresses;
7. Social Security numbers;
8. Medical record numbers;
9. Health plan beneficiary numbers;
10. Account numbers;
11. Certificate/license numbers;
12. Vehicle identifiers and serial numbers, including license plate numbers;
13. Device identifiers and serial numbers;
14. URLs
15. IP addresses
16. Biometric identifiers, including finger and voice prints;
17. Full face photographic images and any comparable images
18. Any other unique identifying number, characteristic, or code

- No longer covered
  - Remove 18 HIPAA identifiers, and no knowledge that people can be re-identified
  - Or guarantee “small” chance of re-identification
- Limited dataset
  - 16 of 18 removed (can keep dates)
  - No consent needed
  - Re-identification prohibited

# De-identifying data

- Basic: replace all IDs with research ID, change dates, remove HIPAA identifiers
- Different shift/function for each patient

- Structured data
  - Add noise
- Text
- Genome
  - Generalize sequence
- Images
  - Facial blur

# Statistical methods

- Do the assumptions of computational methods hold? Do we need new domain-specific ones?
  - E.g. Clinical NLP as distinct subset of NLP
- Importance of domain knowledge
- Error

# Statistical methods

The data are not static:

- EHR systems change
- Terminology changes
- New tests developed

# Nonstationarity

## Risk Calculator for Cholesterol Appears Flawed

By GINA KOLATA
Published: November 17, 2013 | 794 Comments

Last week, the nation's leading heart organizations released a sweeping new set of guidelines for lowering cholesterol, along with an online calculator meant to help doctors assess risks and treatment options. But, in a major embarrassment to the health groups, the calculator appears to greatly overestimate risk, so much so that it could mistakenly suggest that millions more people are candidates for statin drugs.

FACEBOOK
TWITTER
GOOGLE+
SAVE
EMAIL
SHARE
PRINT
SINGLE PAGE
REPRINTS


Enlarge This Image



Mark Graham for The New York Times

Dr. Nancy Cook and Dr. Paul M. Ridker of Harvard Medical School found that a new online calculator used to assess heart treatment options

The apparent problem prompted one leading cardiologist, a past president of the American College of Cardiology, to call on Sunday for a halt to the implementation of the new guidelines.

"It's stunning," said the cardiologist, Dr. Steven Nissen, chief of cardiovascular medicine at the Cleveland Clinic. "We need a pause to further evaluate this approach before it is implemented on a widespread basis."

# Other ways data may be nonstationary

- Changes in record keeping
- Changes in patient population over time
- Changes in terminology (gallop vs third heart sound; crackles vs rales)
- More accurate tests
- New diagnoses/risk factors

- Note diff between physiology and our observation of it!

# Statistical methods

Missing data

Very common and the reason for it changes interpretation

-Chronic vs acute conditions

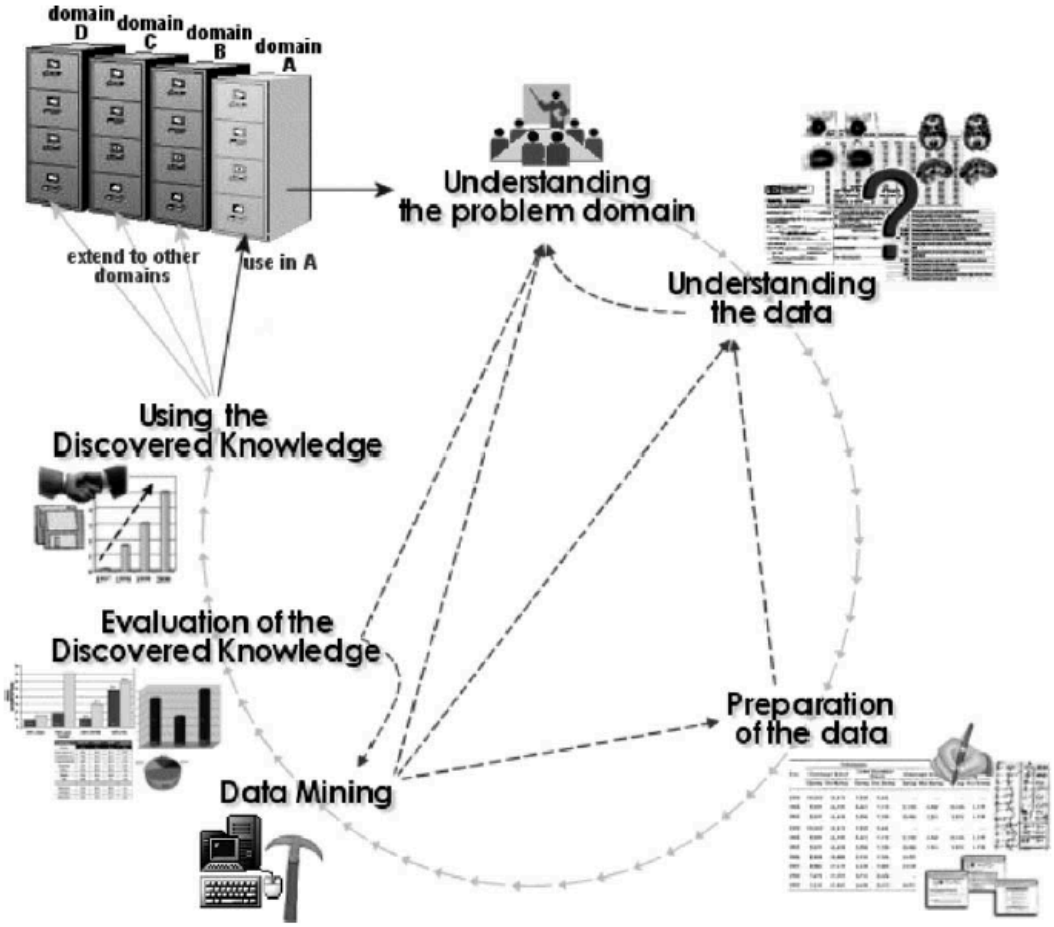
-expensive test (e.g. could be missing due to insurance coverage)

# Statistical methods

Redundancy, inconsistent data

-conflicting lab results

-note says patient is Male, later says Female




Cios, K.J. & William Moore, G., 2002, Uniqueness of medical data mining, *Artificial intelligence in medicine*, 26(1-2), pp. 1-24.

# Special status of medicine

- Impact/risk of false findings: good enough for a paper vs good enough to treat a patient
- Cannot do any conceivable test

# Today’s paper

# More on challenges

- Observational Data in general
  - Nonstationarity
  - Selection bias
  - Choosing variables
  - Seeing vs doing
- Biomedical Data
  - Biased approximation of truth
  - Sample size
  - Censoring (left, right)
  - Institutional differences
  - Fragmentation of data