Electronic health records to study population health: opportunities and challenges

Caroline A. Thompson, PhD, MPH

Assistant Professor of Epidemiology

San Diego State University

Caroline.Thompson@mail.sdsu.edu

@epicaroline

ELECTRONIC HEALTH RECORDS (EHR)

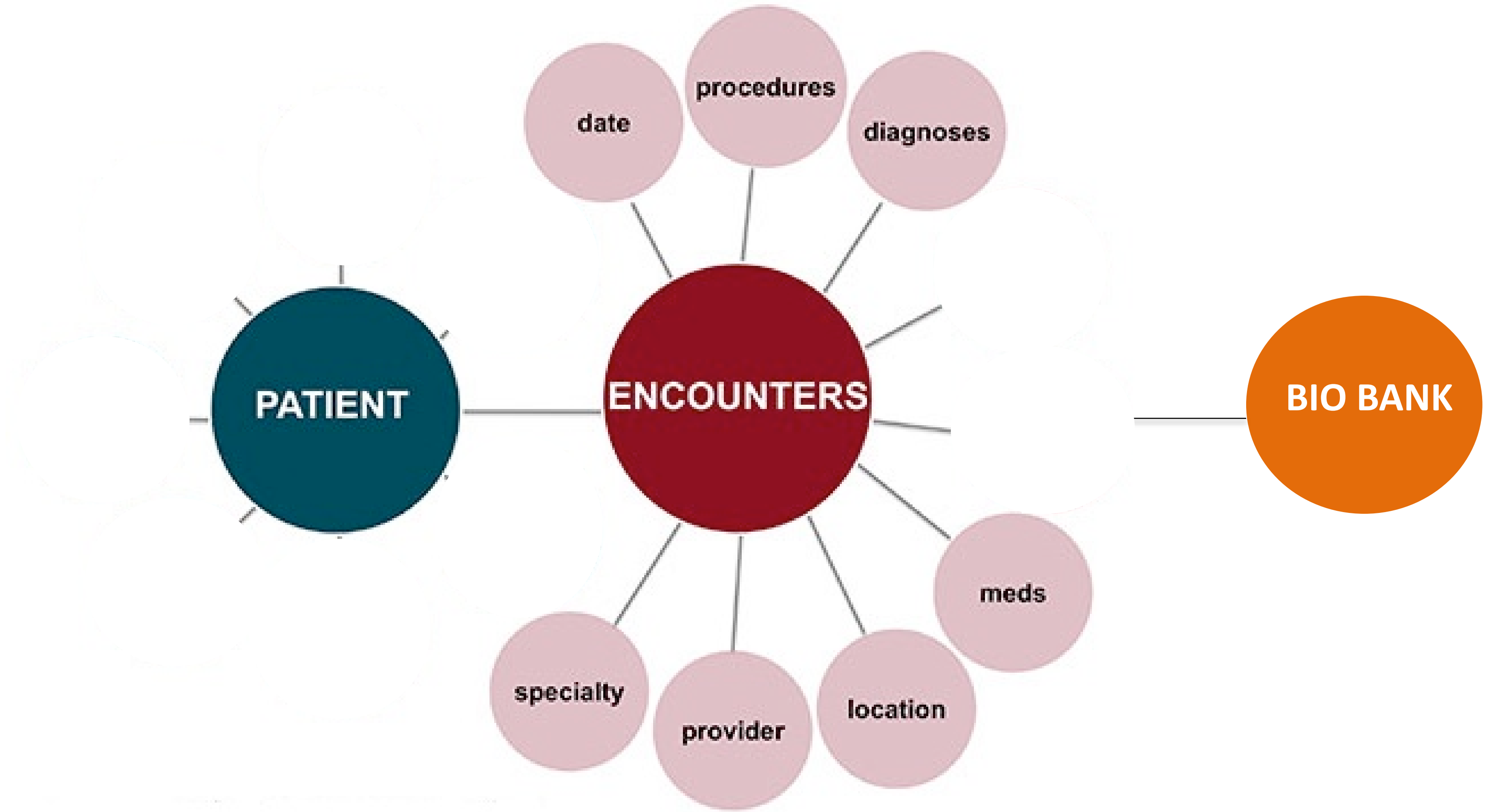
"Digital exhaust" - longitudinal electronic record of patient health information.

"Integrated care"

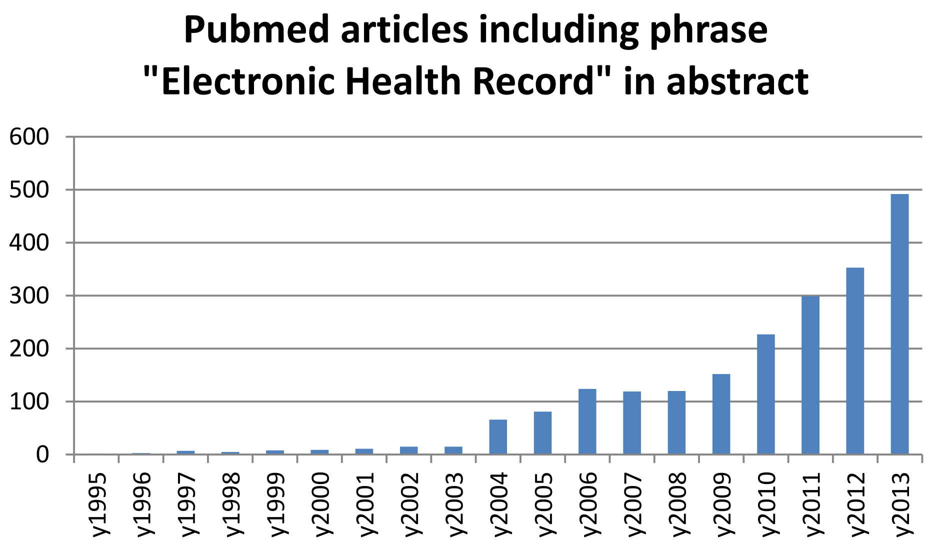
Meaningful Use

eHR
INCENTIVE PROGRAM

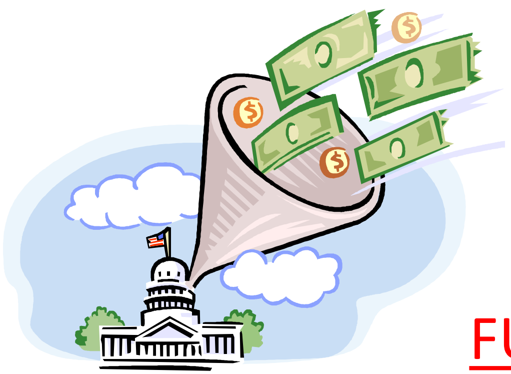
DATA: INSURANCE CLAIMS VS EHR



EHR FOR RESEARCH?



- Veterans Administration
- HMO Research Network
  - Kaiser Permanente
  - PAMF
  - Geisinger health system
- Universities:
  - Harvard + FDA Sentinel
  - Stanford + PAMFRI Oncoshare
  - UCSD CTRI



FUNDING CLIMATE

MAMMOGRAPHY SCREENING IN ASIANS

- Rates among Asians (N=11,268) were the same as NH whites (N=41,927), but screening compliance varied after disaggregating to Asian sub ethnicities.

Table 1: Mammography screening completion in Asians

	Up-to-date		Overdue or Refused	
Non-Hispanic Whites	**32105**	**(76.5%)**	**9822**	**(23.4%)**
All Asian patients	**8452**	**(75.0%)**	**2816**	**(24.9%)**
By Asian subgroup:				
Asian Indian	**1362**	**(67.0%)**	**668**	**(32.9%)**
Japanese	**903**	**(81.0%)**	**211**	**(18.9%)**
Chinese	**4001**	**(77.0%)**	**1195**	**(22.9%)**
Filipino	**1547**	**(74.3%)**	**535**	**(25.6%)**
Korean	**248**	**(73.1%)**	**91**	**(26.8%)**
Vietnamese	**258**	**(83.4%)**	**51**	**(16.5%)**
Native Hawaiian, Pacific Islander	**133**	**(67.1%)**	**65**	**(32.8%)**

Table 2: Predictors of Mammography screening completion (Asians only)

	OR (95% CI)
Race / Ethnicity	
Chinese	1.00
Asian Indian	0.53 (0.47, 0.60)
Filipino	1.27 (1.07, 1.50)
Japanese	0.72 (0.64, 0.82)
Korean	0.78 (0.60, 1.02)
Vietnamese	1.54 (1.12, 2.12)
Native Hawaiian, Pacific Islander	0.53 (0.39, 0.74)
Other Characteristics	
Enrolled in "My Health Online"	1.32 (1.20-1.46)
Primary language	
English	1.00
Not English, physician concordant	0.86 (0.64-1.15)
Not English, not physician concordant	0.81 (0.71-0.92)
Female provider	1.16 (1.00-1.34)
Primary care visits in the past 2 years (per yr)	1.22 (1.20-1.25)

Hierarchical multivariate logistic regression with random intercept for primary care provider, fixed effects at the provider level (sex, degree, specialty, language concordance) and fixed effects at the patient level (age, detailed race/ethnicity, enrolled in myHealthOnline, language concordance, primary care visits in the past two years)

Thompson CA, Gomez SL, Chan JK, et al. Routine cancer screening compliance in a diverse Asian American Population. *Cancer Epidemiology Biomarkers and Prevention*, 11(2208-17), 2014.

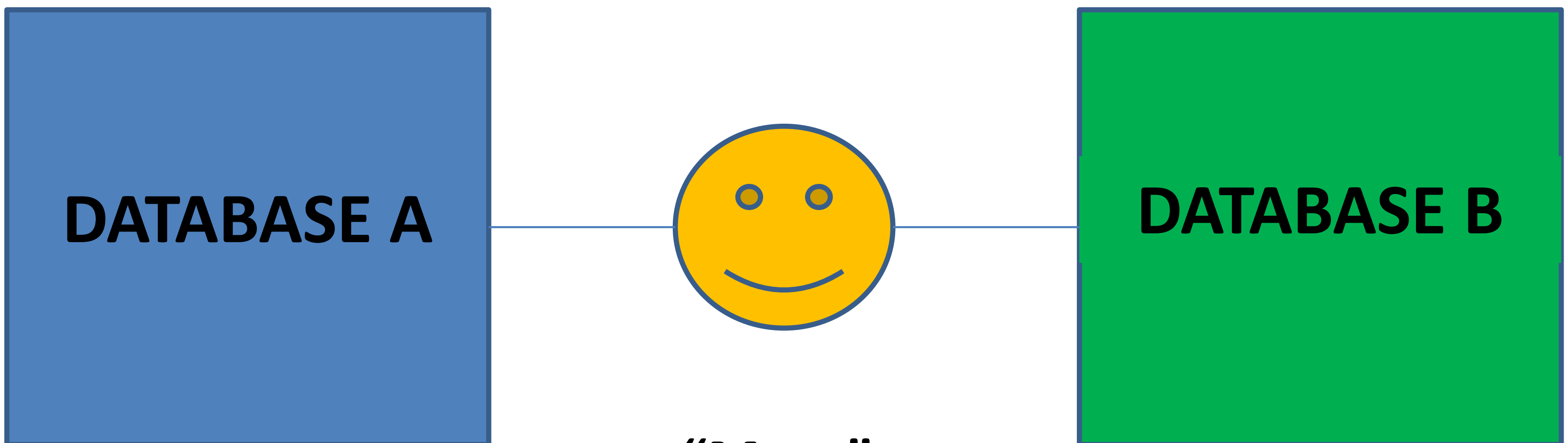
EHR FOR CANCER RESEARCH

CANCER CONTROL CONTINUUM

PREVENTION	EARLY DETECTION	DIAGNOSIS	TREATMENT	SURVIVORSHIP	END-OF-LIFE CARE
-Tobacco control -Diet -Physical activity -Sun exposure -Virus exposure -Alcohol use	-Colorectal cancer screening -Breast cancer screening -Cervical cancer screening	-Biopsy -Histological assessment -Pathology reporting -Tumor stage documented	-Chemotherapy -Hormone therapy -Radiation -Surgery	-Surveillance -Psychosocial care -Management of long-term effects	-Hospice -Palliation
?		?		?	?

ELECTRONIC HEALTH RECORDS

DATA LINKAGE



"Mary"
DOB: 01/23/1945
SSN: 123-45-678
Zip: 12345

CALIFORNIA CANCER REGISTRY (CCR)

Statewide population-based cancer surveillance system

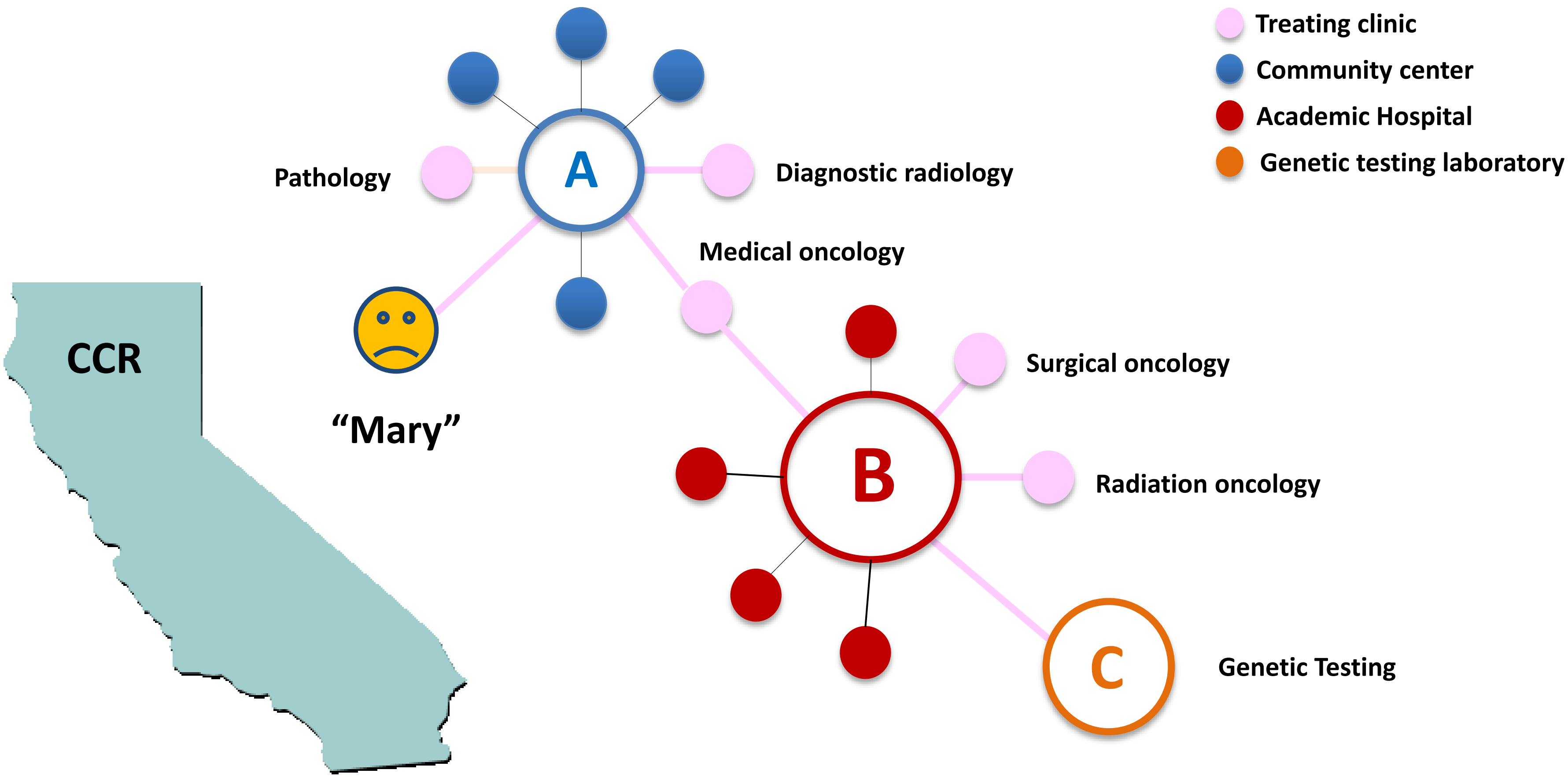
- Data available:
  - Tumor details
  - Initial treatment summaries
  - Survival, SES
- NOT available:
  - Detailed treatment history
  - Providers
  - Cancer recurrences
  - Genetic testing

BREAST CANCER CARE ACROSS HEALTHCARE SYSTEMS

- PAMF: Multispecialty community health care system
- Stanford: Tertiary academic medical center

Epic

LINKING DATA FOR BREAST CANCER RESEARCH



Thompson CA, Kurian AS, Luft HL. Linking Electronic Health Records to Better Understand Breast Cancer Patient Pathways Within and Between Two Health Systems. *eGEMs (Generating Evidence & Methods to improve patient outcomes)*: Vol. 3: Iss. 1, Article 5, 2015.

IS EHR-BASED RESEARCH VALID?

- Data collected for clinical and billing purposes, NOT for research
  - Every data point is subject to a unique selection mechanism, i.e., the reason the patient sought care
- Frequent in/out migration
  - Changes in jobs, insurance, geography
- Data ambiguities/errors/omissions
- Data "pooling" or linkage may introduce additional biases

THE MISSING DATA PROBLEM

X* = observed/measured value of exposure or confounding variable

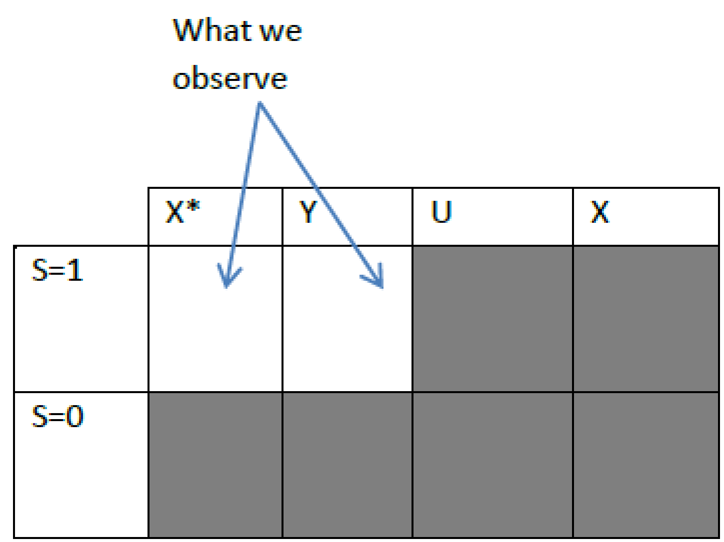
Y = cancer

U = unknown or unmeasured exposure or confounding variable

X = true value of exposure or confounding variable

S= selected / not lost to follow up

All major forms of bias can be thought of as special cases of missing data.



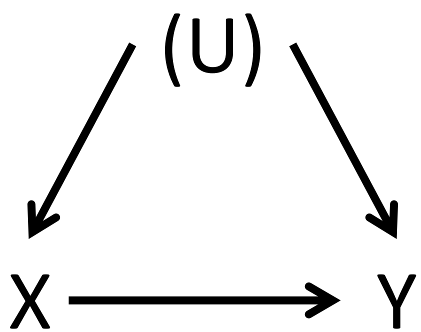
MISSING DATA IN EHR

- Data generated for research provides answers to Yes/No questions
- Data from the EHR provides Yes/**Blank** data

Is the absence of evidence the same as the evidence of absence?

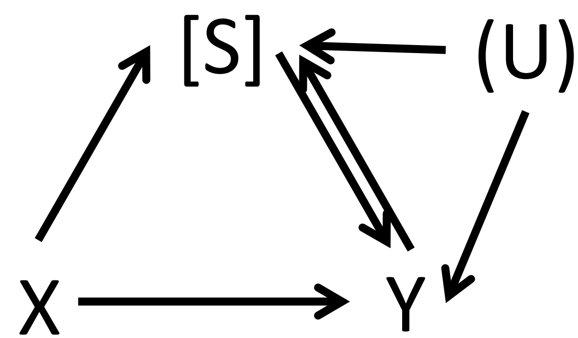
TYPES OF BIAS: DIRECTED ACYLIC GRAPHS

CONFOUNDING BIAS



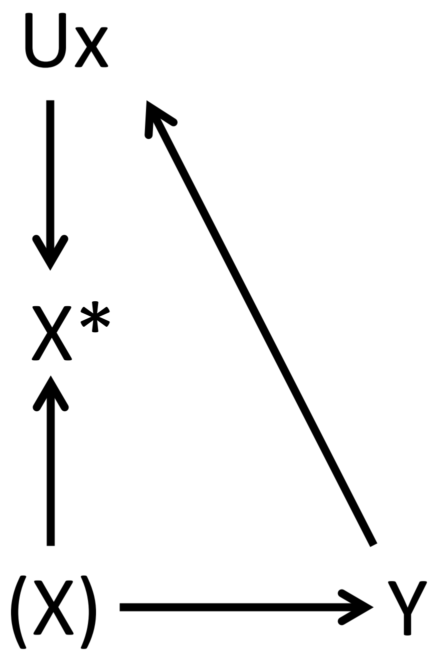
- Unmeasured variables (SES)
- Confounding by indication

SELECTION BIAS



- Inappropriate study design
- Loss to follow up (attrition)

TYPES OF BIAS



Misclassification

INFORMATION BIAS

- Coding errors
- ICD9-ICD10 conversion
- Patient withholding

EXAMPLE: CONFOUNDING BIAS

Published by Oxford University Press on behalf of the International Epidemiological Association
International Journal of Epidemiology 2006;**35**:337–344

Advance Access publication 20 December 2005
doi:10.1093/ije/dyi274

Evidence of bias in estimates of influenza vaccine effectiveness in seniors

Lisa A Jackson,[1,2]* Michael L Jackson,[1,2] Jennifer C Nelson,[1,3] Kathleen M Neuzil[4] and Noel S Weiss[2]

Afzal et al. BMC Medical Informatics and Decision Making 2013, 13:30
http://www.biomedcentral.com/1472-6947/13/30

BMC Medical Informatics & Decision Making

RESEARCH ARTICLE Open Access

Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records

Zubair Afzal*, Martijn J Schuemie, Jan C van Blijderveen, Elif F Sen, Miriam CJM Sturkenboom and Jan A Kors

Abstract

Background: Distinguishing cases from non-cases in free-text electronic medical records is an important initial step in observational epidemiological studies, but manual record validation is time-consuming and cumbersome. We compared different approaches to develop an automatic case identification system with high sensitivity to assist manual annotators.

Methods: We used four different machine-learning algorithms to build case identification systems for two data sets, one comprising hepatobiliary disease patients, the other acute renal failure patients. To improve the sensitivity of the systems, we varied the imbalance ratio between positive cases and negative cases using under- and over-sampling techniques, and applied cost-sensitive learning with various misclassification costs.

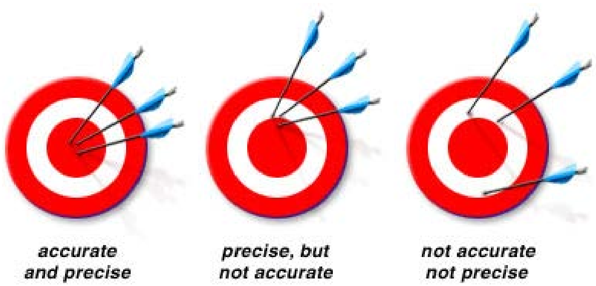
Results: For the hepatobiliary data set, we obtained a high sensitivity of 0.95 (on a par with manual annotators, as compared to 0.91 for a baseline classifier) with specificity 0.56. For the acute renal failure data set, sensitivity increased from 0.69 to 0.89, with specificity 0.59. Performance differences between the various machine-learning algorithms were not large. Classifiers performed best when trained on data sets with imbalance ratio below 10.

Conclusions: We were able to achieve high sensitivity with moderate specificity for automatic case identification on two data sets of electronic medical records. Such a high-sensitive case identification system can be used as a pre-filter to significantly reduce the burden of manual record validation.

Keywords: Class imbalance, Random sampling, Cost sensitive learning, Electronic health records, Improving sensitivity

EPIDEMIOLOGY FOR "BIG DATA"

ACCURACY VS. PRECISION



In large sample sizes, the impact of random error decreases, while that of systematic error becomes more pronounced.

MAINSTREAM MEDIA

HOW CAN EPIDEMIOLOGY HELP?

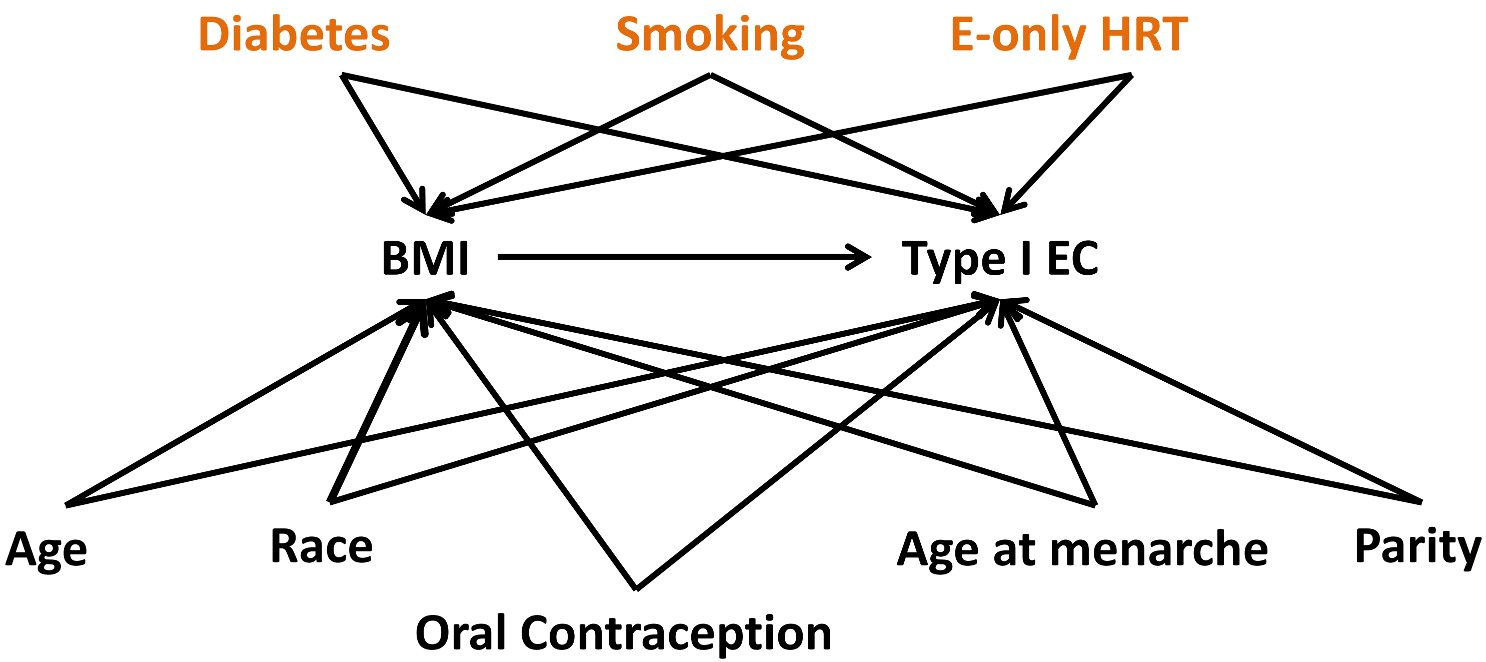
- Respect for the underlying data generating mechanisms
  - Directed Acyclic Graphs
- Careful study design
  - Definition of the study population
  - Anticipation of potential biases
- Validation studies for classification methods
- Bias analysis

➢ **Multidisciplinary teams, including experts from informatics, medicine and epidemiology will be required to make valid inference about population health using EHR data.**

WHAT IS BIAS ANALYSIS?

- <u>Quantitative</u> treatment of uncertainty in nonrandomized research
  - As opposed to qualitative treatment in the discussion section of the publication
- Estimate the magnitude and direction of systematic error
- To produce "adjusted" point estimates and confidence intervals that reflect systematic error as well as random error

VISUALIZING BIAS USING DAGS: BMI AND RISK OF ENDOMETRIAL CANCER



PARTIALLY MEASURED CONFOUNDING

EVIDENCE-BASED PRECISION MEDICINE?

- On demand cohort querying by doctors in the clinic
  - "Patients like mine", "Green button" projects
- Automated algorithms that can detect disease development from routine tests <u>before</u> clinical symptoms
- Treatments tailored to the specific genetics of the patient or disease (e.g., cancer)

THANK YOU

"And it was so typically brilliant of you to have invited an epidemiologist."