INFORMATICS PROFESSIONALS. LEADING THE WAY.

De-identifying Socioeconomic Data at the Census Tract Level for Medical Research Through Constraint-based Clustering

Data Set Management

S85

Yongtai Liu, Douglas Conway, **Zhiyu Wan**, Murat Kantarcioglu, Yevgeniy Vorobeychik, Bradley A. Malin
Vanderbilt University
Twitter: @liu_yongtai, @zhiyuwan
#AMIA2021

Disclosure

I and my spouse/partner have no relevant relationships with commercial interests to disclose.

Adding Socioeconomic Status (SES) into Clinical Records

- Synthetic Derivative Database (SD)
  - De-identified clinical data derived from VUMC[1]'s electronic medical records
  - Rich research resource
  - No socioeconomic information
- There is a need to add SES data into SD Database
  - Person's health status is associated with his/her SES
  - Adding SES can promote more meaningful investigations and findings
- How to acquire SES data?
  - Single patient's SES info is not readily available
  - Estimate a person's SES based on the neighborhood he/she lived in

1. VUMC: Vanderbilt University Medical Center (VUMC)

Census Tract Level SES Data

- 5-year estimates of the American Community Survey (ACS)
  - social, economic, demographic, and housing characteristics of the U.S. population
  - provides data every year, **publicly** available
  - Census Tract-Level Data (census tract: neighborhood of 2,000 – 8,000 people)

- Selected Features: Six ACS Features (based on Brokamp et al[1])
  - ACS data contains over 1,000 variables
  - Select six features covering economics, education, insurance and housing.

- One Additional Feature: Deprivation Index
  - Use one feature to capture the “community deprivation”
  - The first component of the principal components analysis of the six ACS features.

1. Brokamp C, et al. Material community deprivation and hospital utilization during the first year of life: an urban population-based cohort study. Ann Epidemiol. 2019

Example of Census Tract SES Features

Tract ID

Six ACS Features

Scaled, 0 to 1

Census Tract FIPS	Population	Fraction Assisted Income	Fraction High School Education	Median Income	Fraction No Health Insurance	Fraction Poverty	Fraction Vacant Housing	Deprivation Index
10010201 00	1845	12.0689	90.6	67826	9.1	10.6775	1.4379	0.2811
10010202 00	2172	24.1379	82	41287	8.8	22.4137	10.820	0.3995

Problems of Publishing Census Tract Level Data

- VUMC Synthetic Derivative database was de-identified according to Safe Harbor.
  - Added SES data must comply with Safe Harbor

- HIPAA's Safe Harbor (18 rules)
  - Geographic location: all geographic subdivisions smaller than a state must be removed
  - Except for the initial three digits of the ZIP code if combining all ZIP codes with the same three initial digits contains more than **20,000 people**

Problems of Publishing Census Tract Level Data

Patient ID	Weight	Height	…	Medium Income	High School Education
12345	97lbs	5ft		**67826**	**90.6**

Census Tract FIPS	Population	Fraction Assisted Income	Fraction High School Education	Median Income	Fraction No Health Insurance	Fraction Poverty	Fraction Vacant Housing	Deprivation Index
100102	1845	12.0689	90.6	67826	9.1	10.6775	1.4379	0.2811
100102 0200	2172	24.1379	82	41287	8.8	22.4137	10.820	0.3995
100102 0300	3385	12.9007	86.3	46806	5.9	14.6528	11.3652	0.3261

Patient lives in an area with less than 20,000 people, against HIPAA Safe Harbor, NOT De-identified!!!

Solution: Constraint-Based Clustering

- Group similar tracts together into one cluster, while satisfying:
  - Constraint 1. Each cluster must contain at least two census tracts. (*MinTract ≥ 2)*
  - Constraint 2. Each cluster must cover at least 20,000 individuals. (*MinPop ≥ 20,000)*

Group with Six ACS Features

Census Tract FIPS	Population	Fraction Assisted Income	Fraction High School Education	Median Income	Fraction NO Health Insurance	Fraction Poverty	Fraction Vacant Housing	Deprivation Index
19089960100	3,884	12.12	91.4	$47,000	4.4	13.32	7.46	0.30
19179961100	3,025	11.16	91.5	$50,321	4.7	11.93	7.97	0.28
39061026101	8,344	11.27	90.9	$48,733	3.7	12.40	7.95	0.29
39143961300	4,069	12.64	92.5	$50,553	4.6	13.77	7.27	0.29
42037050800	3,513	12.85	92.1	$44,917	4.0	12.13	5.66	0.29
After De-id	22,835	12.01	91.6	$48,304	4.3	12.71	7.26	0.29

Tracts ≥ 2

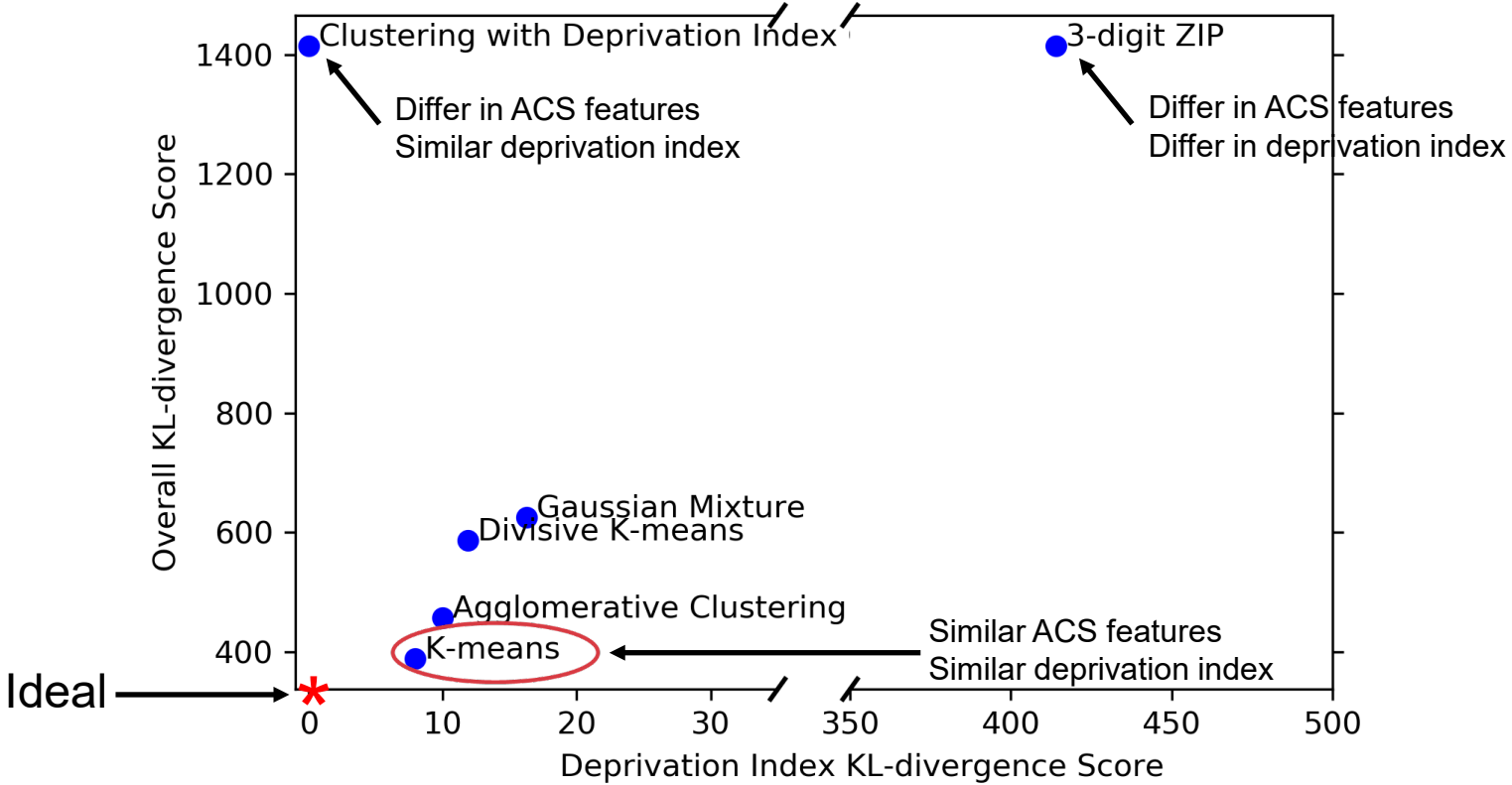
Pop ≥ 20,000

Evaluation. Data Utility

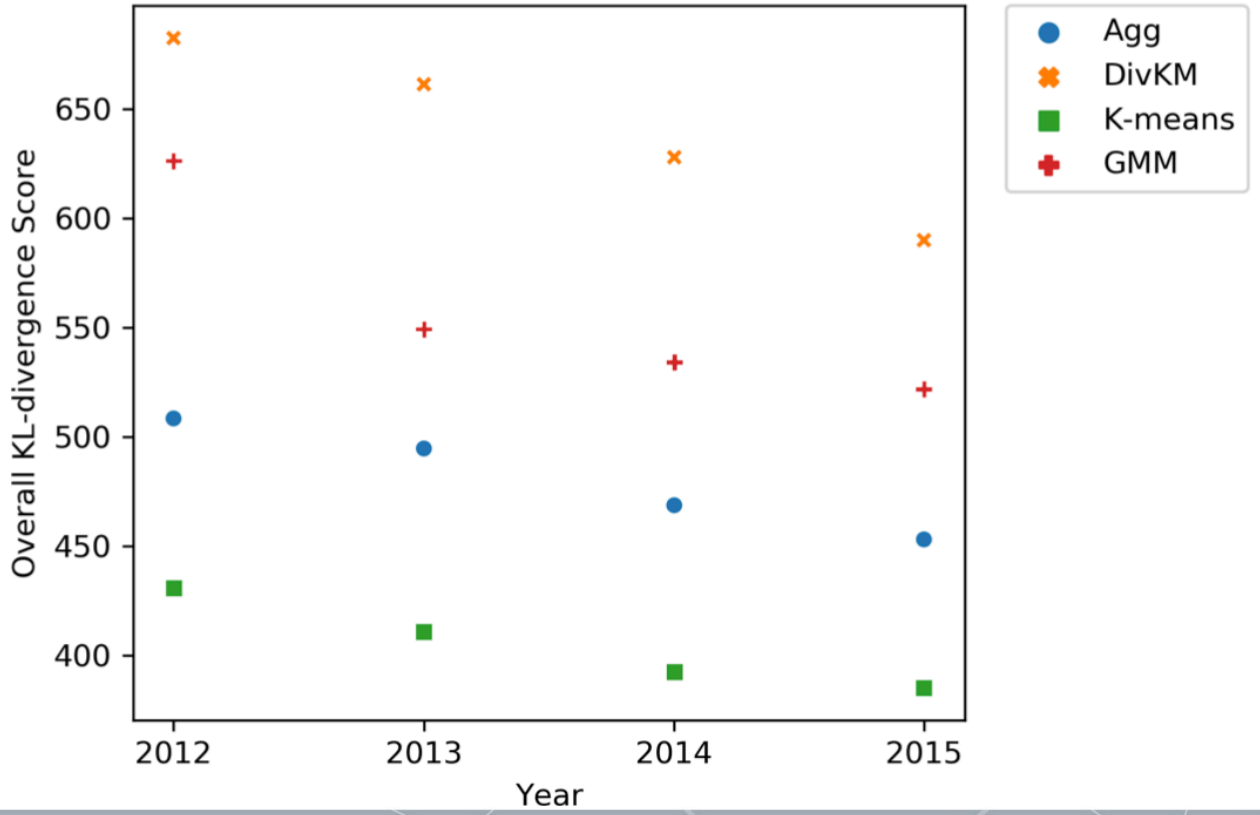
- Clustering Approaches
  - K-means, Hierarchical, DBSCAN etc.
  - HIPAA Recommended 3-digit ZIP approach
    - Group all tracts with same initial 3-digit zip (e.g., 372**) into one cluster, with *MinPop ≥ 20,000*

- Measuring Clustered Data Utility with ***KL-Divergence***
  - $D_{KL}(Clustered||Original)$
  - a higher score suggests a larger divergence from clustered data to the original

Result 1. Clustering 2017 ACS Data



Result 2. Multiple Years Comparison



Result 3. 3-digit ZIP v.s. Clustering (CBC)

Census Tract FIPS	De-id method	Population	Percentage Assisted Income	Percentage High School Education	Median Household Income	Percentage Lacking Health Insurance	Percentage Poverty	Percentage Vacant Housing	Deprivation Index
19089-960100		3,884	12.12%	91.4%	$47,000	4.4%	13.32%	7.46%	0.30
	3-digit ZIP (521**)	58,083	8.5%	92.0%	$51,608	4.2%	10.70%	8.33%	0.30
	CBC	22,835	12.01%	91.6%	$48,304	4.3%	12.71%	7.26%	0.29

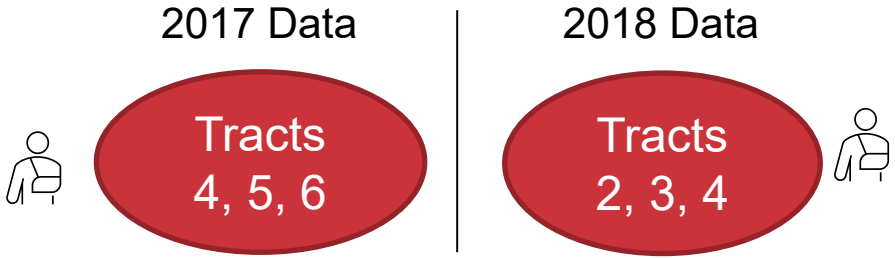
Summary

- Introduced a constraint-based clustering approach to generate census tract-level socioeconomic data that is de-identification compliant.
- Our approach achieved a substantially better data utility than the Safe Harbor recommended 3-digit ZIP method.
- Our approach has been applied to VUMC Synthetic Derivative.

Limitation & Next Steps

- Data Utility Evaluation
  - KL-Divergence only indicates the similarity between distributions of data
  - True utility should be better measured by real-world studies

- Multiple Releases over time
  - Identify unique census tract through linking multiple releases over years



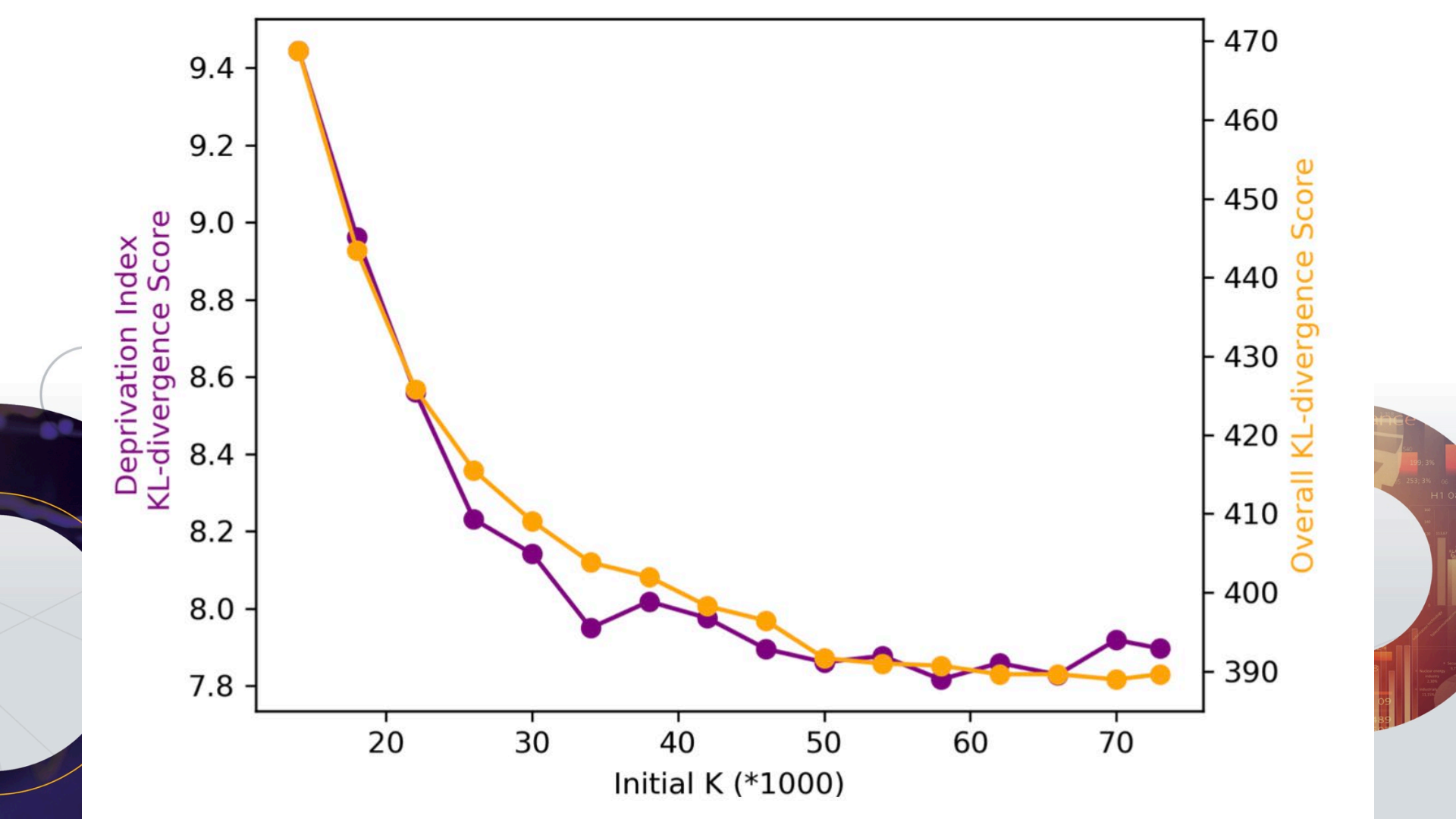
Patient in Tract No.4

Thank you!

Email us at:

Yongtai Liu: yongtai.liu@vanderbilt.edu

Zhiyu Wan: zhiyu.wan@vanderbilt.edu


Deprivation Index
KL-divergence Score
Overall KL-divergence Score
Initial K (*1000)
9.4
9.2
9.0
8.8
8.6
8.4
8.2
8.0
7.8
470
460
450
440
430
420
410
400
390
20
30
40
50
60
70

Learning Objectives

After participating in this session the learner should be better able to:

- Recognize the benefit of adding census tract-level socioeconomic data into clinical records.
- Realize the privacy risk of adding census tract-level socioeconomic data directly into clinical records.
- Understand the purpose and principles of the constraint-based clustering protection approach.