

# **Augmenting Publicly Available Social Determinants of Health Datasets with Clinical Data to Power Multi-Institutional Clinical Research**



Sajjad Abedian, MS  
Evan Sholle, MS  
Yongkang Zhang, PhD

11/14/2020

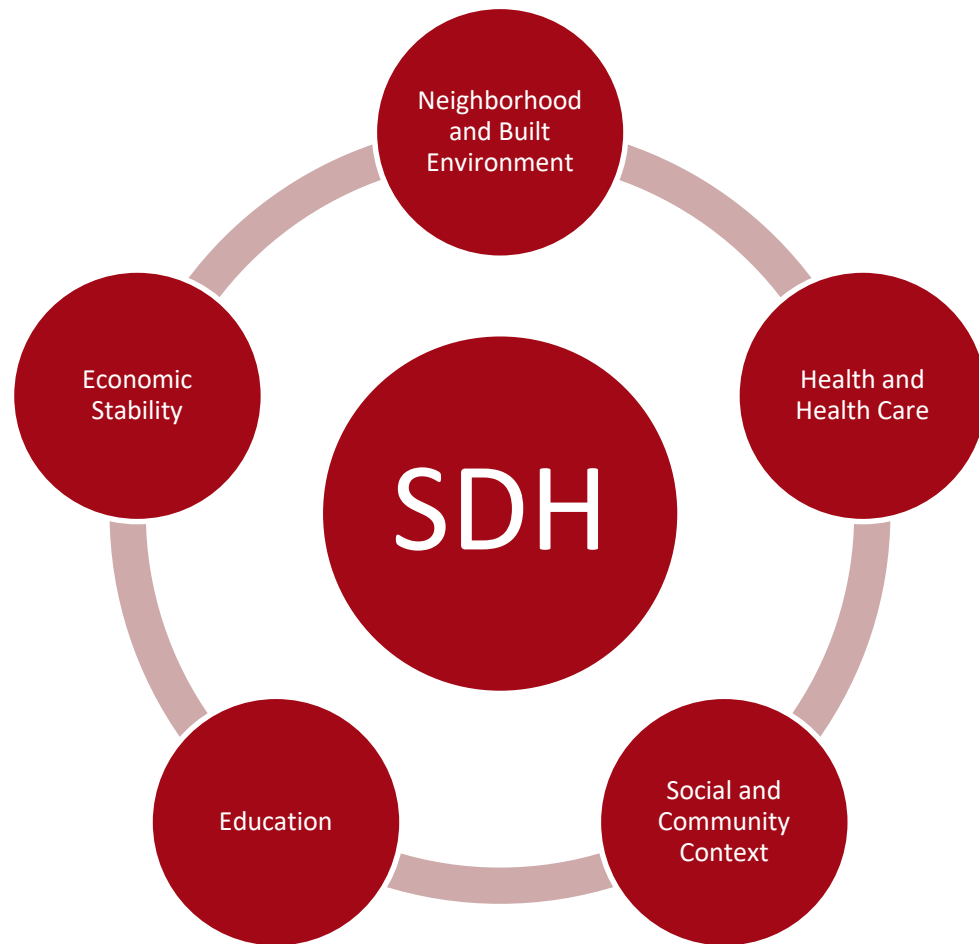
# Outline

- **SDH Data Section**
  - What is SDH?
  - How is it collected?
  - How to link it?
  - Interactive: how to create a SDH analytical file?
- **Clinical Data Section**
  - Interactive: how to create a patient-level analytical file combined with SDH variables?
  - Association Between Residential Neighborhood Social Conditions and Health Care Utilization and Costs

# What is SDH?

And why is it important?



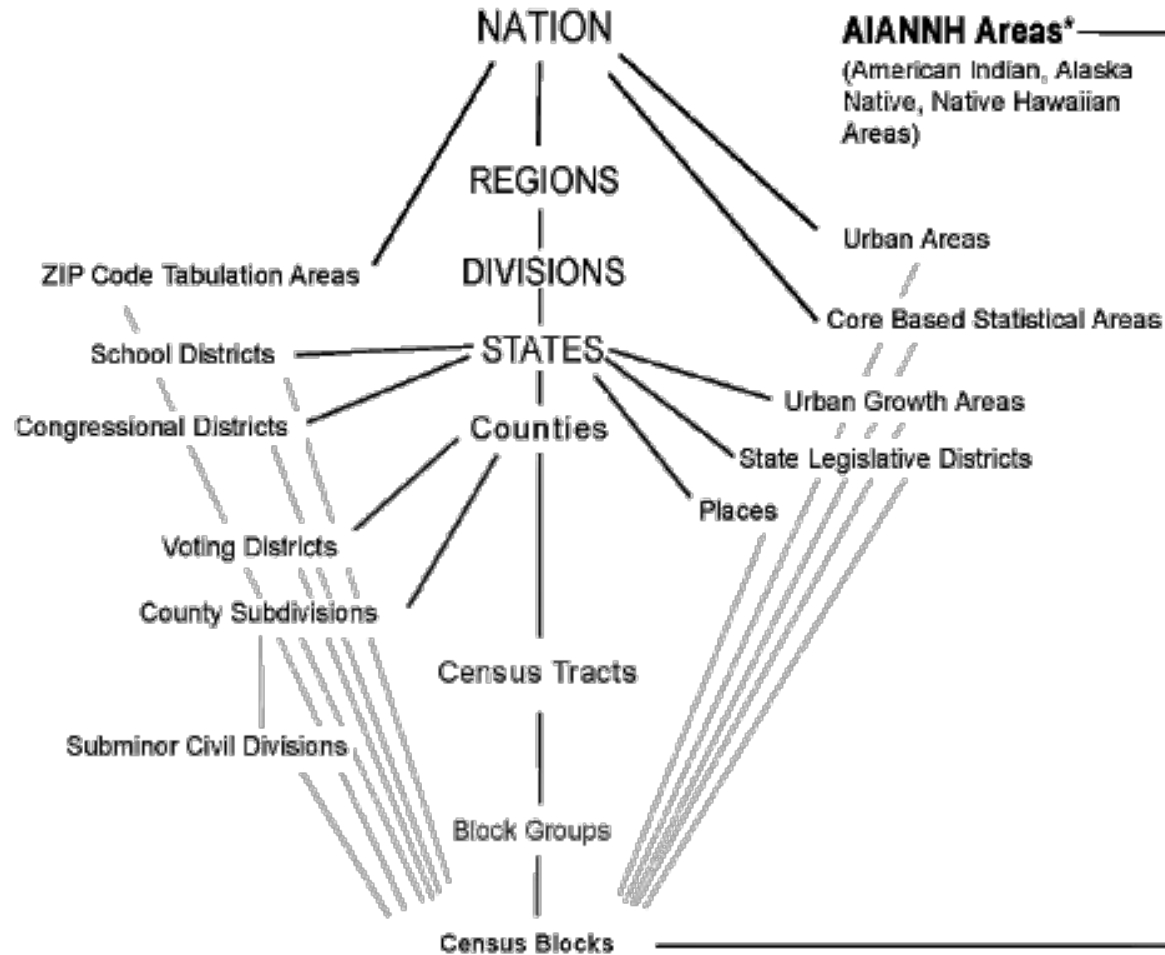


# How is it collected?



# Geographical Units

## Standard Hierarchy of Census Geographic Entities



# Census Tract vs. Block Group

## Census Summary Levels:

Blocks

Block Groups

Tracts\*

County Subdivisions\*

Counties\*

State\*



\*Included with Maptitude. Blocks and block groups available separately.

# Census Tract vs. Block Group

## Census Summary Levels:



06 – identifies California,

067 – identifies Sacramento County within California,

001101 – identifies Census Tract 11.01 within Sacramento County and

1085 – identifies Census Block 1085 within tract 11.01.

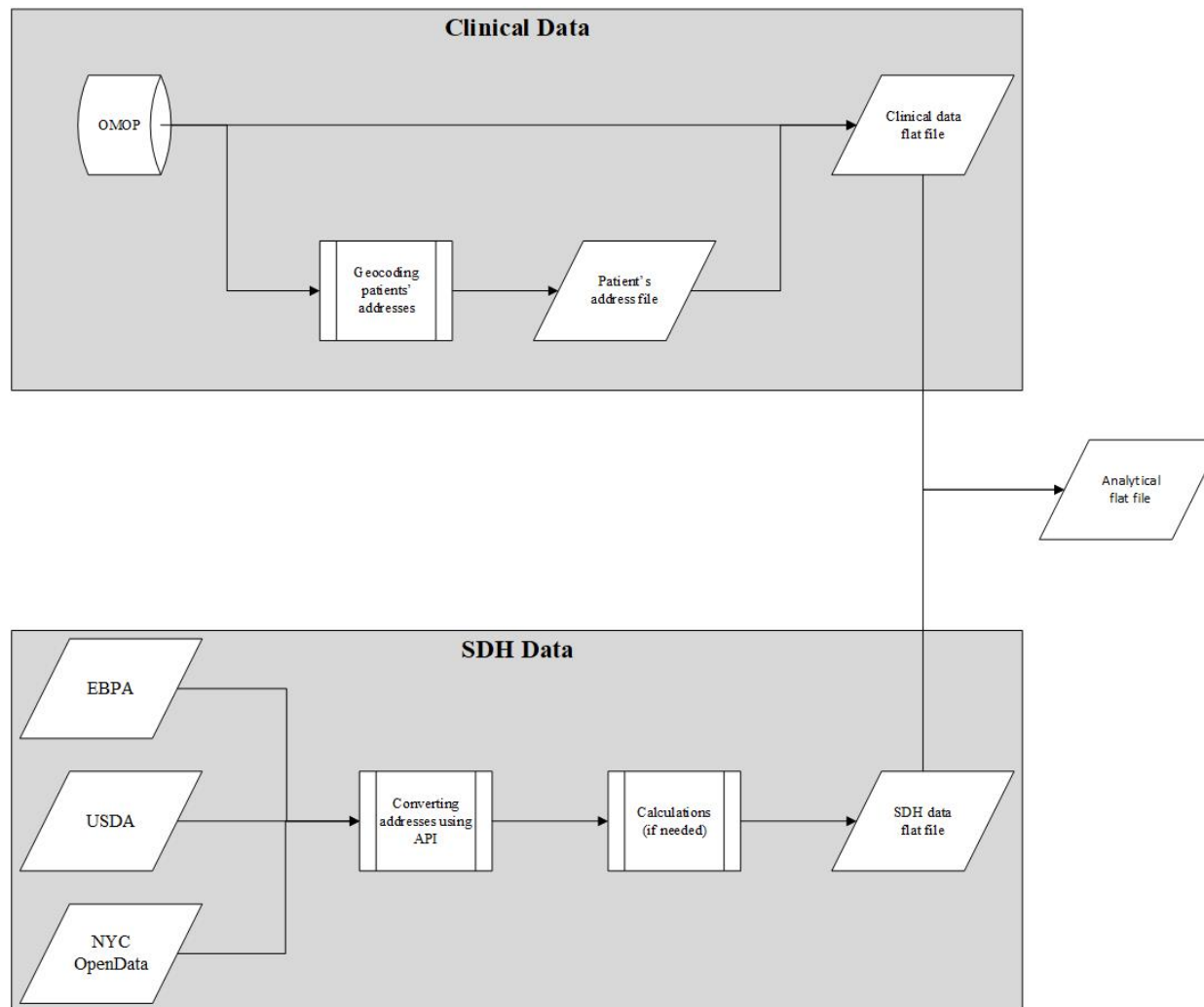
\*Included with Maptitude. Blocks and block groups available separately.



# Sources

- **Publicly available data sources**
  - American Community Survey (ACS)
  - EPA National Air Toxics Assessment
  - USDA
  - NYC OpenData
  - Factors Affecting Communities and Enabling Targeted Services (FACETS)
- **EHR**
  - Structured
  - Unstructured

# Process of Linking EHR to SDH



# VACCINe

Variables Affecting Care in Communities and  
Individuals in New York

[https://github.com/SajjadAbedian/AMIA\\_2020\\_VACCINE/blob/main/VACCINe\\_walkthrough.ipynb](https://github.com/SajjadAbedian/AMIA_2020_VACCINE/blob/main/VACCINe_walkthrough.ipynb)



# **Association Between Residential Neighborhood Social Conditions and Health Care Utilization and Costs**

Yongkang Zhang, PhD

11/14/2020



# Disclosure

I disclose that neither I nor my partner have relevant financial relationships with commercial interests.

# Learning Objectives

- Understand strategies to combine social determinants of health data, clinical data, and administrative claims data for health services research.
- Understand the associations between disadvantaged neighborhood social conditions and healthcare utilization and costs among Medicare patients.

# SDoH and Healthcare

- SDoH have strong associations with healthcare quality, utilization, and patient outcomes
  - Social factors account for 25%–60% of deaths in the United States
  - Neighborhood social disadvantage is associated with increased risk of 30-day readmission
  - Food-insecure children are twice as likely to report being in fair or poor health and at least 1.4 times more likely to have asthma, compared to food-secure children

# Efforts to Address Social Risk Factors

- A growing interest in addressing social risk factors to improve population health among policy makers, researchers, and healthcare providers
  - Comprehensive screening to identify unmet health-related social needs
  - Increase patients' awareness of community services
  - Navigation services to assist high-risk patients with accessing community services
  - Alignment between clinical and community services to ensure that community services are available and responsive to the needs of patients



# Gaps in Literature

- Many studies have found that disadvantaged social conditions are associated with poor clinical outcomes and hospital readmissions
- Little is known about how SDoH are associated with other types of potentially preventable healthcare utilization and costs associated with preventable utilization
- Limited evidence is available regarding how SDoH influence total healthcare utilization or costs

# Objectives of the Study

- To examine the associations of neighborhood social conditions with total annual and potentially preventable Medicare costs

# Methods: Population and Sample

- Medicare fee-for-service patients who had at least one clinical encounter in five major health systems in New York City
  - Patients with a valid 9-digit zip code in New York or New Jersey
  - Continuously enrolled in Medicare Parts A and B for 12-month in 2013 and 2014
  - Had no Medicare Advantage enrollment
  - 93,429 patients were included

# Methods: Healthcare Utilization and Costs

- Medicare claims data
  - Healthcare utilization across all care settings and different providers and organizations
  - Payment made by Medicare, patients, and other payers
  - Importance of using claims data to measure utilization and costs (Smith et al., JAMIA, 2019).
- Total Medicare costs
  - Standardized total annual Medicare costs across all service categories for 2013 and 2014, including physician services, outpatient services, inpatient services, durable medical equipment, Part D prescription, and post-acute care (e.g., skilled nursing facility and home health) and hospice services

# Methods: Healthcare Utilization and Costs

- Potentially preventable utilization
  - Preventable ED visits
    - Billings' algorithm
  - Preventable hospitalizations
    - AHRQ's PQI indicators
  - Unplanned 30-day hospital readmissions
    - CMS algorithm for pay-for-performance programs
- Potentially preventable costs
  - Standardized episodic costs of all potentially preventable utilization and all services delivered within 30 days after a preventable ED visit or hospitalization for each patient in 2013 and 2014.

# Methods: Demographics, Diagnoses, and Procedures

- Medicare claims data across all care settings
- Electronic health records data for the same patient cohort from multiple health systems in NYC.

# Methods: Social Determinants of Health

- Multiple choices of SDoH from different resources
  - Individual or community level?
    - Medicare claims and EHR data have very limited individual level SDoH: dual-eligible status
    - Comprehensive SDoH are available at community level
  - What level of community?
    - Medicare claims data contain state, county, and 5-digit zip codes of patient residential location
    - 9-digit zip codes are available from EHR data



# Methods: Social Determinants of Health

- Multiple choices of SDoH from different resources
  - Data available with different granularity

| Neighborhood              | Distinct Areas |
|---------------------------|----------------|
| State                     | 50             |
| County                    | 3,200          |
| Zip Code Tabulation Area  | 32,000         |
| 5-digit Zip Code          | 42,000         |
| Census Tract              | 74,000         |
| Census Block Group        | 221,000        |
| Census Block              | 11 millions    |
| Possible 9-digit Zip Code | 30 millions    |





# Methods: Social Determinants of Health

- Many policy and economic studies used state for country level SDoH data
- Health policies are developed and implemented at the state or county level
- Policy makers need to understand how state or county level SDoH are associated with uptake of health policies among healthcare organizations and providers
  - Health information technology
  - Novel payment and delivery models (e.g., ACO)
- State or country level SDoH data may be important confounders that need to be adjusted in policy analysis



# Methods: Social Determinants of Health

|                               | <b>ZIP<br/>(Zone Improvement Plan) code</b>  | <b>ZCTA<br/>(ZIP Code Tabulation Area) code</b>       |
|-------------------------------|--|---|
| <b>Developer</b>              | United States Postal Service (USPS)  | United States Census Bureau                           |
| <b>Year of Implementation</b> | 5-digit: 1963<br>9-digit zip+4: 1983   | Census 2000   |
| <b>Total Numbers</b>          | 42,000   | 32,000  |
| <b>Purpose</b>                | Introduced as as a method to improve mail delivery service   | Tabulate summary statistics                           |
| <b>Data availability</b>      | Limited  | 5-year American Community Survey                      |
| <b>Boundary</b>               | Could cross states   | In a few cases, ZCTAs can cross into bordering states |
| <b>Frequency of change</b>    | ZIP codes change frequently. Every year there are thousands of changes to the USPS ZIP Code records. | Could be stable for 10 years or more.                 |



# Methods: Social Determinants of Health

- Census block groups: small, relatively homogenous areas of 600-3,000 persons
- Census tracts: aggregates several block groups, and represent approximately 4,000 persons
- Many SDoH measures are available at both tract and block group level
  - American Community Survey

# Methods: Social Determinants of Health

- Multiple choices of SDoH from different resources
  - Composite score or individual measures?
    - Individual measures reflecting important neighborhood socioeconomic status: median household income, unemployment rate, % of people without high school degree etc.
    - Composite scores: a summary of neighborhood social conditions across different measures: Area Deprivation Index, Social Deprivation Index, Social Vulnerability Index etc.

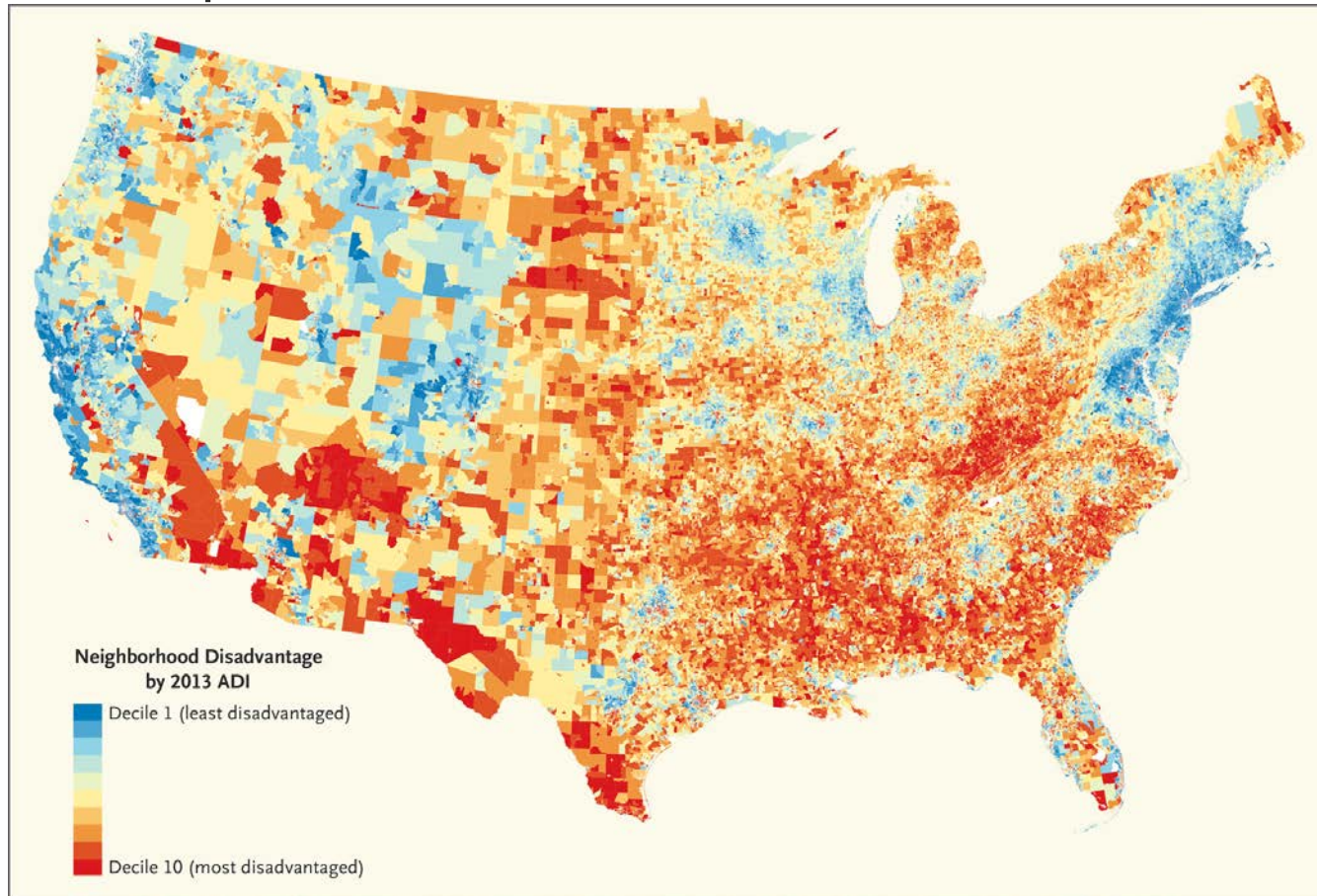
# Methods: Social Determinants of Health

- Multiple choices of SDoH from different resources
  - Area Deprivation Index

| Components for Area Deprivation Index  | Weight  |
|--|---------|
| Percentage of population aged $\geq 25$ with $< 9$ years of education  | 0.0849  |
| Percentage of population aged $\geq 25$ with at least a high school diploma  | -0.0970 |
| Percentage of employed persons aged $\geq 16$ in white collar occupations  | -0.0874 |
| Median family income   | -0.0977 |
| Income disparity (log of 100 * ratio of the number of households with $< \$10,000$ annual income to the number of households with $\geq \$50,000$ annual income) | 0.0936  |
| Median home value  | -0.0688 |
| Median gross rent  | -0.0781 |
| Median monthly mortgage  | -0.0770 |
| Percentage of owner-occupied housing units (home ownership rate)   | -0.0615 |
| Percentage of civilian labor force population aged $\geq 16$ unemployed (unemployment rate)  | 0.0806  |
| Percentage of families below the poverty level   | 0.0977  |
| Percentage of population below 150% of the poverty threshold   | 0.1037  |
| Percentage of single-parent households with children aged $< 18$   | 0.0719  |
| Percentage of occupied housing units without a motor vehicle   | 0.0694  |
| Percentage of occupied housing units without a telephone   | 0.0877  |
| Percentage of occupied housing units without complete plumbing   | 0.0510  |
| Percentage of occupied housing units with $>1$ person per room   | 0.0556  |

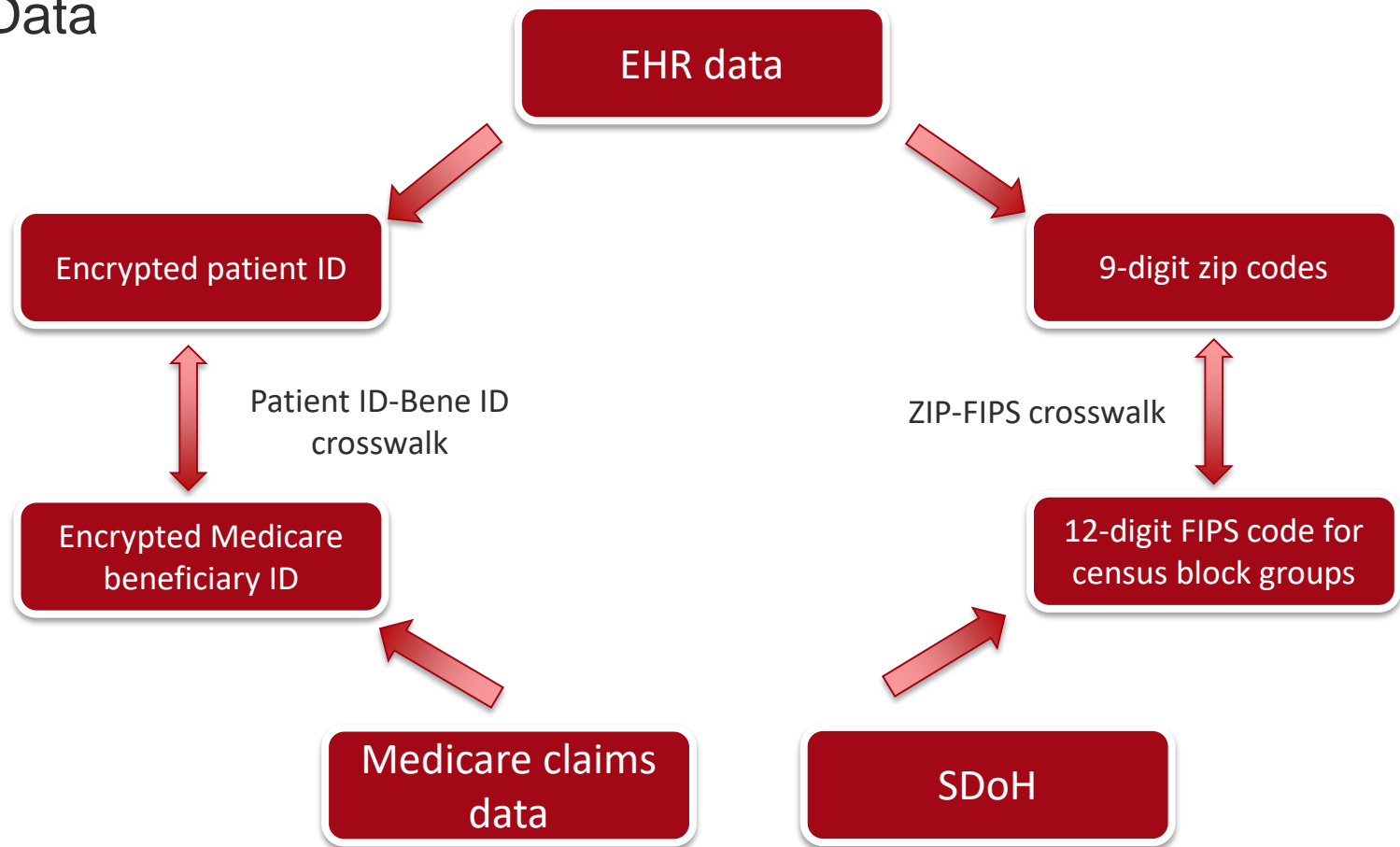
# Methods: Social Determinants of Health

- Multiple choices of SDoH from different resources
  - Area Deprivation Index



# Methods: Data Linkage

- Using EHR data as a bridge between SDoH and Medicare Data



# Analysis: Descriptive Analysis

- We first compared differences in demographics, comorbidities, and other patient characteristics across patients in different **ADI quintiles**
  - Demographics: age, sex, and race
  - Comorbidities: number of chronic conditions, frailty, serious mental illness, serious physical illness, and end-stage renal disease
  - CMS Hierarchical Condition Categories score for each patient to measure the.
  - Dual-eligible status and Medicare Part D coverage
- We used  $\chi^2$  tests for categorical variables and t-tests for continuous variables to test for differences in patient characteristics across ADI quintiles.



# Analysis: Multivariable Analysis

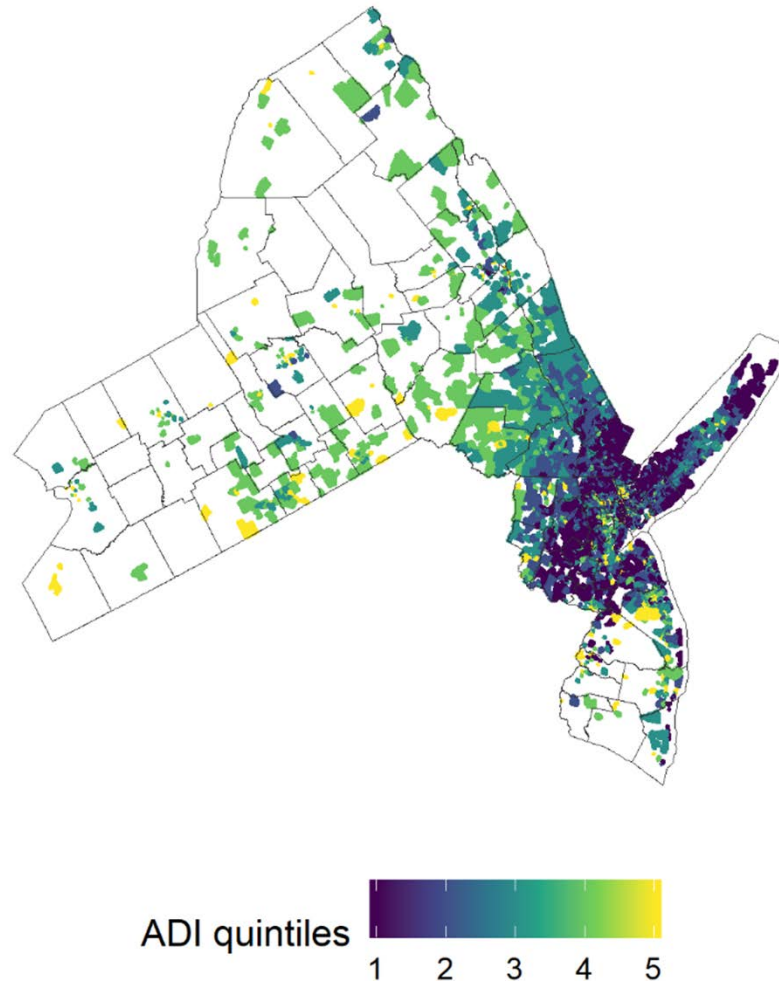
- Generalized linear models (GLM) to examine the association of ADI quintiles with (1) Total annual Medicare costs and (2) Cost associated with preventable utilization
- We also tested the association between ADI quintiles and the probability of having the three types of the preventable utilization by using logistic regressions

# Analysis: Multivariable Analysis

- We focused on the neighborhoods with the least disadvantaged conditions (quintile 1) and the most disadvantaged conditions (quintile 5), as compared with the neighborhoods with the intermediate level conditions (quintile 3)
- All regressions were adjusted for demographic, comorbidities, and other patient characteristics of each year
- We additionally added year-fixed effects to control for secular changes in outcomes and indicators for hospital referral regions to account for supply side variations across regions

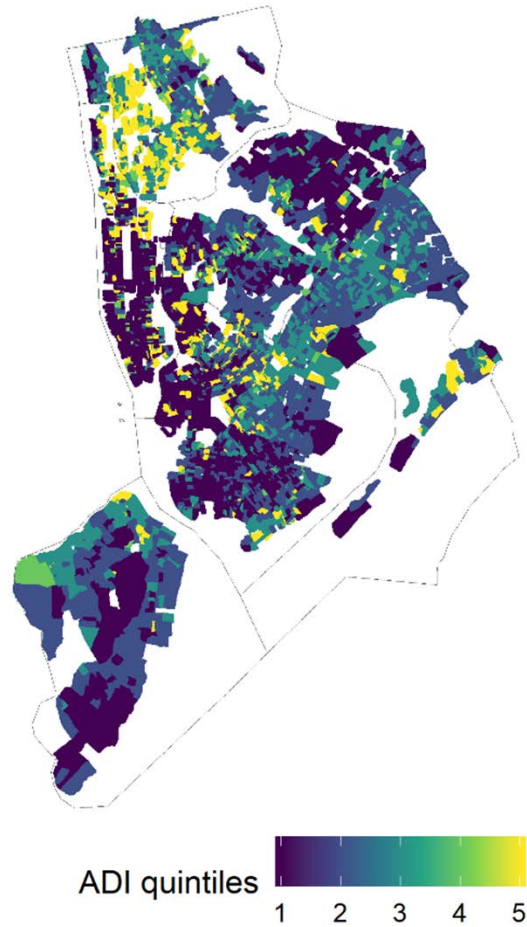
# Results: Descriptive Analysis

- Geographic location and social conditions



# Results: Descriptive Analysis

- Geographic location and social conditions



# Results: Descriptive Analysis

- Patients with different social conditions have different characteristics.

| Patient and Neighborhood Characteristics                                   | Quintiles of ADI at Census Block Group Level |                            |                            |                            |                          |                          |
|--|--|----------------------------|----------------------------|----------------------------|--------------------------|--------------------------|
|  | All<br>(N = 93,429)                          | Quintile 1<br>(N = 48,974) | Quintile 2<br>(N = 17,838) | Quintile 3<br>(N = 13,480) | Quintile 4<br>(N = 5010) | Quintile 5<br>(N = 8127) |
| Neighborhood characteristics   |  |                            |                            |                            |                          |                          |
| No. census block groups included in this study                             | 12,420                                       | 3849                       | 3516                       | 2690                       | 1102                     | 1263                     |
| Average number of Medicare or dual-eligible residents in each block group* | 160.0  | 156.0                      | 159.4                      | 163.0                      | 167.3                    | 152.1                    |
| Median family income (\$)*   | 82,197                                       | 133,891                    | 91,750                     | 77,917                     | 66,008                   | 41,197                   |
| Patient characteristics  |  |                            |                            |                            |                          |                          |
| Age (%)  |  |                            |                            |                            |                          |                          |
| < 65   | 12.4   | 7.2                        | 16.0                       | 18.0                       | 16.9                     | 24.1                     |
| 65–74  | 40.1   | 42.2                       | 40.8                       | 37.5                       | 37.2                     | 32.0                     |
| 75–84  | 32.3   | 34.5                       | 29.9                       | 29.2                       | 30.1                     | 29.5                     |
| ≥ 85   | 15.2   | 15.9                       | 13.4                       | 15.3                       | 15.8                     | 14.4                     |
| Sex (%)  |  |                            |                            |                            |                          |                          |
| Male   | 39.4   | 40.1                       | 40.6                       | 38.8                       | 37.5                     | 35.1                     |
| Female   | 60.6   | 60.0                       | 59.4                       | 61.2                       | 62.5                     | 64.9                     |
| Race (%)   |  |                            |                            |                            |                          |                          |
| White  | 79.6   | 88.3                       | 76.6                       | 73.2                       | 73.1                     | 48.9                     |
| Black  | 9.7  | 4.1                        | 12.0                       | 14.6                       | 14.9                     | 27.5                     |
| Hispanic   | 3.2  | 1.2                        | 3.2                        | 4.5                        | 4.8                      | 11.7                     |
| Other  | 7.5  | 6.5                        | 8.3                        | 7.7                        | 7.2                      | 11.9                     |
| Dual eligible (%)  | 23.6   | 13.5                       | 25.7                       | 32.5                       | 27.9                     | 61.8                     |
| Part D coverage (%)  | 76.8   | 74.4                       | 77.1                       | 78.7                       | 77.5                     | 87.1                     |
| ESRD   | 1.9  | 1.0                        | 2.6                        | 2.8                        | 2.0                      | 4.6                      |
| Originally enrolled in Medicare because of disability                      | 19.9   | 12.8                       | 24.9                       | 27.3                       | 25.5                     | 35.8                     |
| No. chronic conditions (%)   |  |                            |                            |                            |                          |                          |
| 0  | 3.5  | 3.4                        | 3.3                        | 3.2                        | 4.0                      | 4.1                      |
| 1–2  | 10.6   | 11.6                       | 9.2                        | 9.5                        | 10.8                     | 9.3                      |
| 3–5  | 33.3   | 36.1                       | 30.8                       | 29.8                       | 32.5                     | 28.9                     |
| 6–8  | 32.7   | 31.9                       | 34.2                       | 33.8                       | 31.9                     | 32.4                     |
| ≥ 9  | 20.0   | 17.0                       | 22.5                       | 23.8                       | 20.9                     | 25.3                     |
| Frailty (%)  | 18.6   | 16.9                       | 19.4                       | 20.9                       | 20.0                     | 22.9                     |
| Mental illness (%)   | 21.2   | 18.8                       | 20.8                       | 24.7                       | 23.9                     | 28.7                     |
| Seriously ill (%)  | 17.1   | 15.2                       | 18.7                       | 19.1                       | 18.5                     | 20.5                     |
| HCC score, mean  | 1.5  | 1.3                        | 1.6                        | 1.6                        | 1.5                      | 1.7                      |

We compared the differences in patient characteristics between quintile 1 and each other quintile (quintiles 2–5) using *t* test and the difference are all statistically significant ( $P < 0.001$ ).

\*Data are from American Community Survey to represent the overall characteristics of these communities.

ADI indicates Area Deprivation Index; ESRD, end-stage renal disease; HCC, Hierarchical Condition Categories.

# Results: Descriptive Analysis

- Patients with disadvantaged social conditions have higher unadjusted Medicare spending

|                 | Average total annual Medicare costs (\$) | Physician (\$) | Outpatient (\$) | Inpatient (\$) | Post-acute care and hospice (\$) | DME (\$) | Part D (\$) |
|-----------------|--|----------------|-----------------|----------------|----------------------------------|----------|-------------|
| ADI, quintile 1 | 17,661                                   | 5,928          | 2,460           | 3,744          | 1,732                            | 230      | 3,566       |
| ADI, quintile 2 | 22,965                                   | 6,880          | 3,210           | 5,260          | 2,400                            | 325      | 4,890       |
| ADI, quintile 3 | 23,841                                   | 6,867          | 3,311           | 5,305          | 2,665                            | 336      | 5,356       |
| ADI, quintile 4 | 21,564                                   | 6,174          | 2,965           | 4,830          | 2,567                            | 312      | 4,716       |
| ADI, quintile 5 | 26,815                                   | 6,174          | 4,154           | 6,001          | 2,969                            | 432      | 6,860       |
| Average         | 20,670                                   | 6,398          | 2,901           | 4,514          | 2,147                            | 285      | 4,425       |

# Results: Descriptive Analysis

- Patients with disadvantaged social conditions have higher unadjusted preventable spending and utilization

|                 | Potentially preventable Medicare costs (\$) | Preventable ED visits (%) | Preventable hospitalizations (%) | 30-day readmissions (%) * |
|-----------------|---|---------------------------|----------------------------------|---------------------------|
| ADI, quintile 1 | 1,343                                       | 8.7                       | 3.7                              | 13.7                      |
| ADI, quintile 2 | 2,260                                       | 11.9                      | 5.4                              | 16.8                      |
| ADI, quintile 3 | 2,568                                       | 13.3                      | 6.1                              | 18.3                      |
| ADI, quintile 4 | 2,232                                       | 13.5                      | 5.6                              | 18.9                      |
| ADI, quintile 5 | 3,361                                       | 18.5                      | 8.1                              | 20.0                      |
| Average         | 1,918                                       | 11.1                      | 4.8                              | 16.3                      |

*\* Analysis only using hospitalized patients*

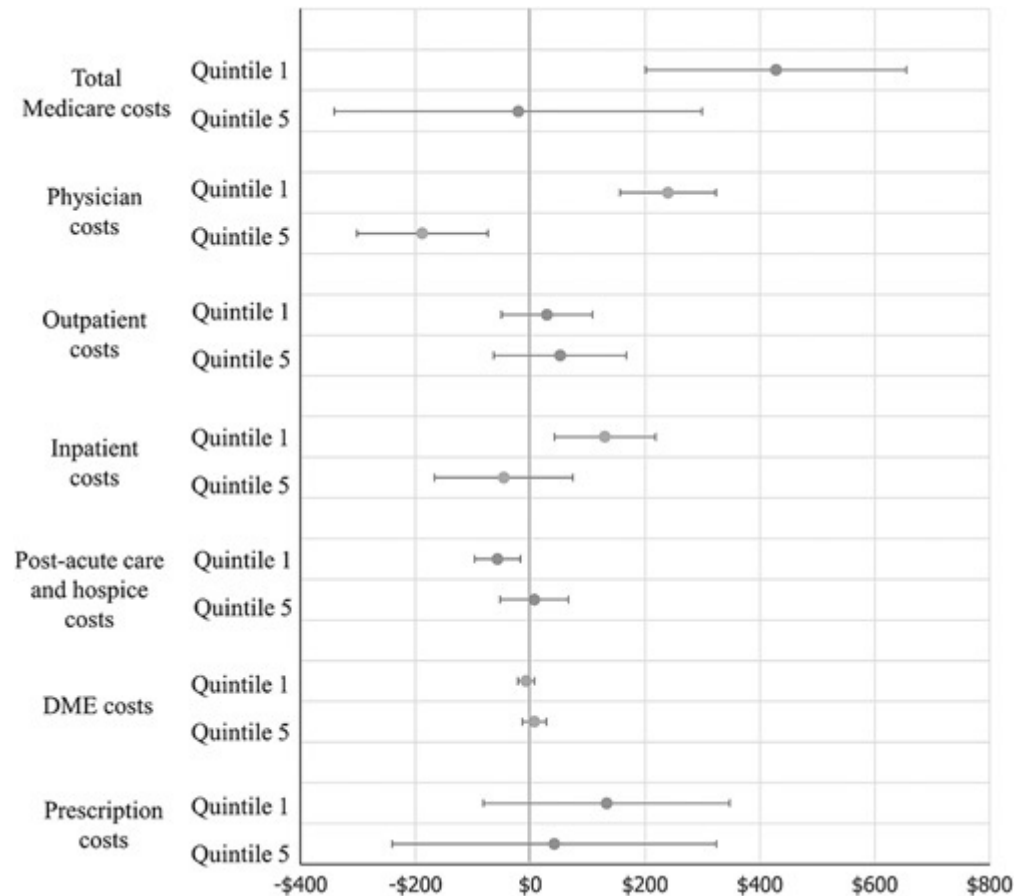
# Results: Multivariable Analysis

- Compared with the neighborhoods with intermediate level social conditions (quintile 3), the least disadvantaged neighborhoods (quintile 1) was associated with higher total Medicare costs (\$427 or 3.2% higher,  $P < 0.001$ )
- The most disadvantaged neighborhoods (quintile 5) had similar total Medicare costs (\$22 or 0.2% lower,  $P = 0.89$ ) after adjusting for patient demographics, comorbidities, and other characteristics



# Results: Multivariable Analysis

- The association between social disadvantage and spending varies by care setting

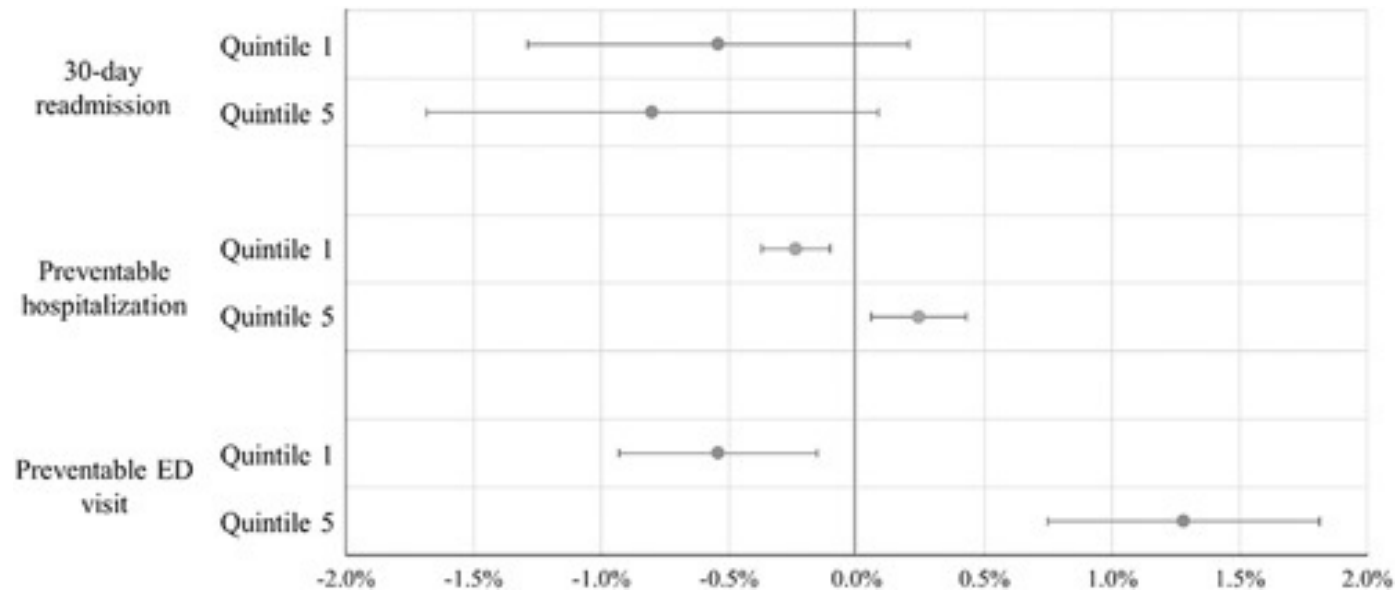


# Results: Multivariable Analysis

- Compared with the neighborhoods with intermediate level social conditions (quintile 3), the least disadvantaged neighborhoods (quintile 1) had similar potentially preventable costs (−\$23; 95% CI, −\$56 to \$10)
- The most disadvantaged neighborhoods (quintile 5) had higher potentially preventable costs (\$53; 95% CI, \$1–\$104) than patients with the intermediate level social conditions

# Results: Multivariable Analysis

- The association between social disadvantage and preventable healthcare utilization



# Discussions

- Disadvantaged neighborhood conditions are associated with lower total annual Medicare costs but higher potentially preventable costs after controlling for demographic, medical, and other patient characteristics.
- Socioeconomic barriers may limit access and use of primary care and disease management services, resulting in a higher proportion of their health care costs going to potentially preventable care.

# Discussions

- Results of this study do not necessarily indicate causal relationships between social conditions and health care utilization and costs
- This study was not able to incorporate individual level SDoH in the analysis and therefore cannot examine if the associations of community SDoH are still significant after controlling for individual level SDoH
- The 3 types of preventable utilization may not all be truly preventable; alternatively these services may not represent the totality of preventable utilization

# Acknowledgement

- Patient-Centered Outcomes Research Institute Health Systems PCORnet (Patient-Centered Clinical Research Network) Demonstration Project (HSD-1604-35187, PI: Kaushal)
- National Institute on Aging Pathway to Independence Award (K99AG064030, PI: Zhang)

# References

1. Amarasingham R, Moore BJ, Tabak YP, et al. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med Care* 2010;48:981–8.
2. Meddings J, Reichert H, Smith SN et al. The impact of disability and social determinants of health on condition-specific readmissions beyond Medicare risk adjustments: A cohort study. *J Gen Intern Med* 2017;32:71–80.
3. Blum AB, Egorova NN, Sosunov EA, Gelijns AC, DuPree E, Moskowitz AJ, Federman AD, Ascheim DD, Keyhani S. Impact of socioeconomic status measures on hospital profiling in New York City. *Circ Cardiovasc Qual Outcomes* 2014;7:391–397.
4. Keyhani S, Myers LJ, Cheng E, et al. Effect of clinical and social risk factors on hospital profiling for stroke readmission: a cohort study. *Ann Intern Med* 2014;161:775–84. doi:10.7326/M14-0361.
5. Jamei M, Nisnevich A, Wetchler E, Sudat S, Liu E. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. *PLoS One*. 2017;12(7):e0181173.
6. Leveraging Data-Driven Insights to Support Development Of Targeted Supplemental Benefits Under Newly Expanded Flexibilities In Medicare Advantage To Improve Outcomes In High-Cost, High-Need Beneficiaries. *AcademyHealth*, 2019.
7. Donze J, Aujesky D, Williams D, Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Intern Med* 2013;173:632-8.

# Thank you!

Yongkang Zhang, PhD  
Research Associate  
Department of Healthcare Research & Policy  
Division of Health Policy and Economics  
[yoz2009@med.cornell.edu](mailto:yoz2009@med.cornell.edu)



**Weill Cornell**  
**Medicine**