Incorporating External Data into Existing Longitudinal Studies

Public Workshop: Addressing Inadequate Information on Important Health Factors in Pharmacoepidemiology Studies Relying on Healthcare Databases

Jacqueline Major, PhD
Division of Epidemiology-1
OPE/OSE/CDER/FDA
May 4, 2015

Disclosure Statement

There are no conflicts of interest to disclose.

The views expressed in this talk are that of the presenter and do not necessarily represent the views of the US FDA

Goal

- Discuss a method for bringing external data or contextual measures into an existing cohort study

Outline of presentation

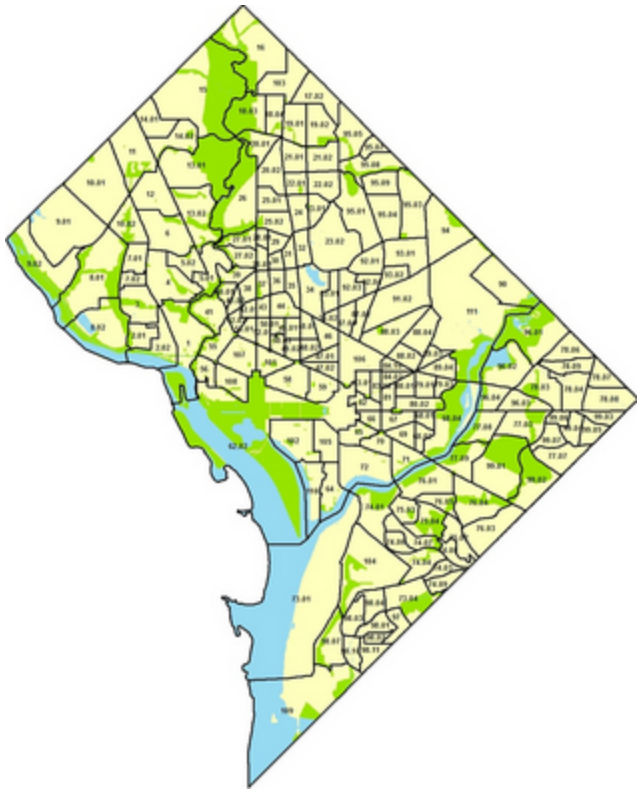
- Various levels of information
- Spatial units
- Illustrative example

Levels of Information

- Patient-level data
  - Incomplete or missing covariate information
  - Income, education, lifestyle, environment
- Aggregate data
  - External data sources
  - SES, smoking/drinking, ambient air quality, and healthcare infrastructure

Spatial Units

- Defining our physical environment
  - Latitudes/longitudes
  - Street level
  - Census tract
  - County
  - Etc.



Washington DC census tracts

Illustrative Example – Chronic Liver Disease Mortality

- **Examine risk in NIH-AARP Diet and Health Study by:**
  - Neighborhood socioeconomic deprivation index
  - Healthcare resources (Insured, acculturation, hospitals, physicians)
  - Alcohol Outlet Density

Major JM, Sargent JD, Graubard BI, Carlos HA, Hollenbeck AR, Altekruse SF, Freedman ND, McGlynn KA (2013)
Local geographic variation in chronic liver disease and hepatocellular carcinoma: contributions of socioeconomic deprivation, Alcohol retail outlets, and lifestyle.

Existing Cohort with Person-level Measures

Self report: Questionnaire

- Age
- Sex
- Race
- Education
- Marital status
- Body mass index
- Health status
- Smoking status
- Alcohol intake
- Physical activity
- Meat/veg/fruit intake
- Calories
- Saturated fat intake

… and accounted for in the analyses

HCC incidence: state cancer registries

CLD mortality: National Death Index

Aggregate Data - SES

- Participant addresses
  - HIPPA compliance
  - Third party geocoded residential addresses
  - 11-digit Census Tract Identifier
    - First 2 digits represent State
    - Next 3 digits represent County
    - Last 6 digits represent the Tract ID
  - Approx. 90% exact match; remaining intersection/centroid
- > 18,500 unique tracts
- Linked participant record to 2000 US Census on tract id

Aggregate Data - SES

USA➔ State➔ County➔ Tract➔ Block group➔ Block

- Data is presented in Summary Files (SF1-SF4)
- All 50 states and US territories
- Each SF differs in data resolution and information it contains (SF3 has 813 tables)

Aggregate Data – Healthcare Availability

- **Area Health Resource File (AHRF)**
  - Health Resources and Services Administration (HRSA), DHHS
- **Aggregated data at the county level**
  - Number of hospitals and physicians
  - Population size for each county

Aggregate Data – Alcohol Outlet Density

- **North American Industry Classification System codes**
  - Retail addresses
- **Kernel Density Estimation (KDE)**
  - Adaptive bandwidth
  - Assigns a density value to each location
    - Irrespective of arbitrary administrative boundaries
    - Smooth, continuous surface
    - Account for the population density (background population)

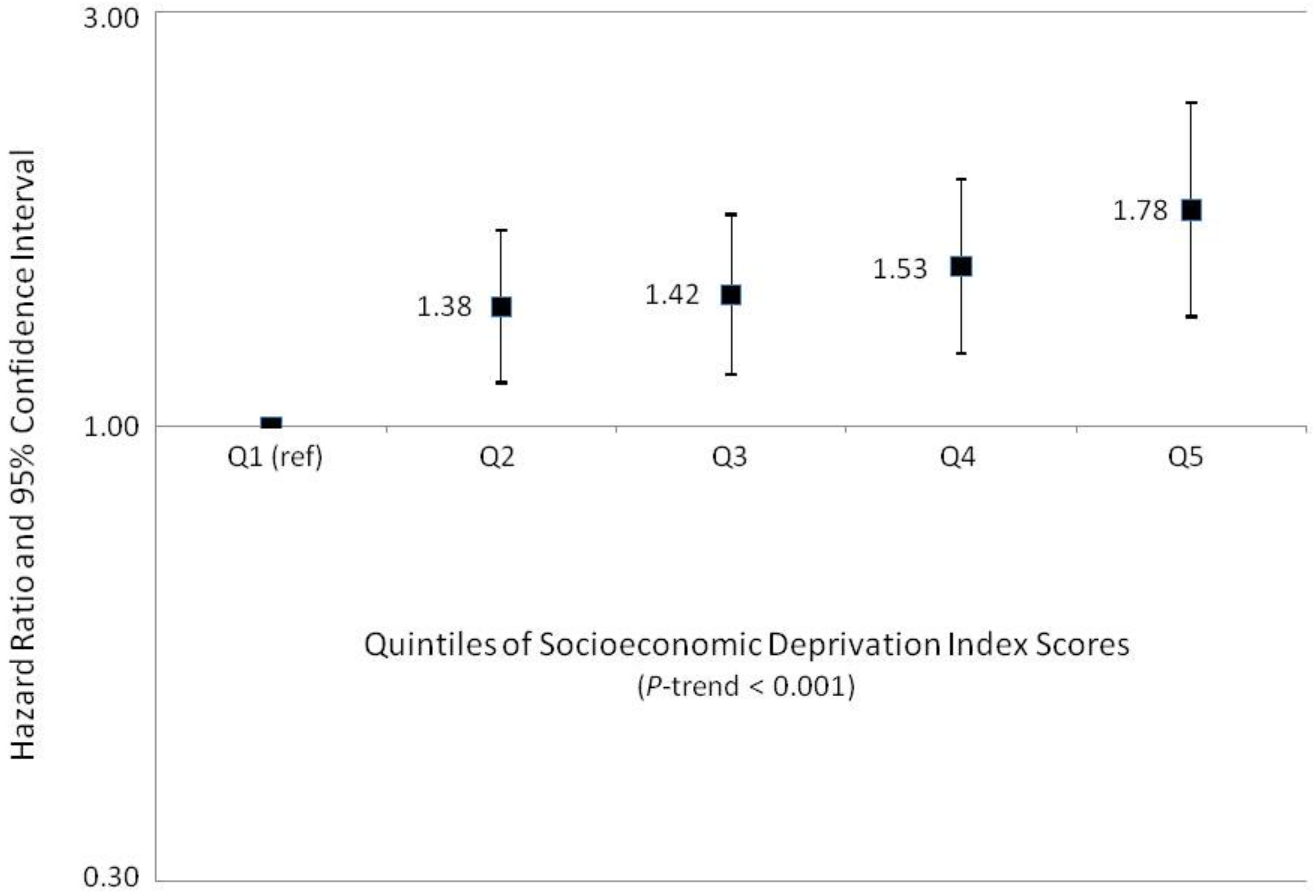
SES Indicators, Healthcare Factors, CLD mortality

	CLD Deaths	
	Minimally adj	Fully adj
Socioeconomic measures		
Most vs. least deprived	**2.36 (1.79, 3.11)**	**1.78 (1.34, 2.36)**
Healthcare Infrastructure		
Fewer hospitals	0.85 (0.66, 1.11)	0.86 (0.67, 1.10)
Fewer physicians	0.90 (0.67, 1.21)	0.88 (0.65, 1.17)
More alcohol outlets	**1.41 (1.22, 1.62)**	**1.26 (1.09, 1.45)**

Minimally adjusted models account for age, sex, and race.

HRs (95% CIs) reported; CLD deaths = 805.

Neighborhood SES and CLD Mortality



Fully adjusted model

Summary of study findings

- **Chronic liver disease mortality**
  - Associations with neighborhood deprivation and alcohol outlet density
  - Residing in deprived areas may increase risk of CLD mortality beyond that explained by individual SES and lifestyle

Limitations

- Predominantly older, middle class
  - Variation in SES may not be as wide as other study populations
  - Associations may be stronger in more socioeconomically-diverse studies
- Relatively few existing prospective studies geocoded
  - Geographically-diverse cohort studies
- HIPPA compliance, third party geocodes, costs

Strengths

- Large prospective study, NIH-AARP (>500,000 participants)
- Geographically diverse (to an extent)
- Multiple layers of data
  - Linkage of the 2000 US Census and other external data sources to the NIH-AARP allowed us to examine these associations, while accounting for important health risk factors.

Thank You!