Overview

1. Types of data available from health systems
2. The Geisinger Health System, the Environmental Health Institute (EHI), the HMO Research Network
3. Examples of our ongoing research
   - Use of patient data
   - Use of other secondary data sources for environmental characterization
4. A bit more detail
   - MRSA and animal feeding operations
   - Type 2 diabetes and coal abandoned mine lands
5. Challenges and parting thoughts
6. Questions
EHR Data in Epidemiology

• EHR: longitudinal digital record of patient health information generated by clinical encounters in a care delivery setting; data generated daily
• ARRA 2009 committed substantial funds to increase use of EHR and claims data to improve practice & research
• Relatively low cost for large sample size, longitudinal data
• Prospective studies: broader range of clinical outcomes
• Genomic studies: for patient phenotyping
• Surveillance projects: near real-time data
• Comparative effectiveness research: compare clinical interventions in shorter time and with fewer costs than in prospective clinical trials
  – So useful that some conclude that continued CER success is now largely dependent on health information technology

Claims Data in Epidemiology

• Claims data are created by payers from bills generated by providers seeking payment for services rendered
• Sources: private insurers, Medicare, Medicaid, DoD, Dept. of Veterans Affairs
• Access is becoming easier
• Data on all inpatient & outpatient services while enrolled in a health plan, but not the outcome of these services
• Unlike EHR data (data limited to care received at one health system), claims include data on all covered services received regardless of provider
• Common use in pharmacoepidemiology and cost studies
Doing Epidemiology in Health Systems – Some Examples

• Secondary data analysis
  – Observational epidemiology, pharmacoepidemiology, health services research, clinical trials – common in health systems
  – Environmental epidemiology – Geisinger EHI unique?

• Use the EHR to recruit patients
  – By sociodemographic variables, by disease status

• Primary data collection
  – Any new data that gets IRB approval
  – To improve specificity of diagnosis, obtain additional tests
  – Tissue for genetic or epigenetic measurements

• Biobanking – link to EHR data for a variety of studies
  – MyCode study – tissue biobanking linked to EHR data
  – eMERGE network (electronic Medical Records & Genomics)

Doing “BIG” Epidemiology in MULTIPLE Health Systems

• HMO Research Network
  – Pioneered common data model; 18 members, EHRs & research enterprises; existing Cancer & CVD Research Networks

• FDA’s Sentinel Initiative and Mini-Sentinel Pilot Program
  – Pharmaceutical post-market risk identification & analysis system
  – “…highly functional network…includes…[over] 26 collaborating institutions…125M patients; establishment of distributed data set…conforms to common data model…” (Psaty JAMA 2013)

• Common Fund’s HC Systems Research Collaboratory
  – “…to strengthen the national capacity to implement cost-effective large-scale research studies that engage health care delivery organizations as research partners.”
  – Coordinating Center at Duke University; pragmatic clinical trials

• Contrast with CDC’s NEPHTN – fewer successes?
The Geisinger Clinic

- 40+ community practice clinics and 6+ hospitals
- 400,000+ primary care patients representative of the general population in the region
  - 2M+ specialty care patients
- EHR since 2001, > 11 years of data
- Across a large, varied geography (40+ counties)
- Patients can use the health system with any health insurance
- Recent partnerships with the Guthrie and Susquehanna Health Systems
- 30% of primary care patients have Geisinger Health Plan insurance – can get claims data
Geisinger EHR – Epic Software

- Primary care and specialty patients
- Inpatient, outpatient, emergency, and telephone encounters
- Socio-demographics, health insurance (surrogate for SES)
- Vital signs, doctor orders, problem list
- Laboratory tests, medications
- Procedures, imaging
  - Results may be in secondary databases
- ICD-9 codes accompany encounters, labs, procedures, medications, and orders

Overview of Environmental Epidemiologic Studies to Date

<table>
<thead>
<tr>
<th>Health Outcome</th>
<th>Environmental Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Marcellus shale ¹</td>
</tr>
<tr>
<td>Asthma</td>
<td>✓</td>
</tr>
<tr>
<td>Cardiovascular outcomes</td>
<td>✓</td>
</tr>
<tr>
<td>Chronic rhinosinusitis</td>
<td>✓</td>
</tr>
<tr>
<td>Diabetes</td>
<td>✓</td>
</tr>
<tr>
<td>MRSA</td>
<td>✓</td>
</tr>
<tr>
<td>Obesity, child</td>
<td>✓ ³</td>
</tr>
</tbody>
</table>

¹ Also efforts on injuries, adverse pregnancy outcomes
² AFO = animal feeding operation; AEU at farm & crop field application of manure
³ Food, physical activity, and land use environments

NIH = NIH-funded
Methods Common to All Studies

- Obtain patient data from EHR
- Geocode patients – automated and manual
- Consider how environment contributes to disease burden
  - Define exposure of interest – INDIVIDUAL or CONTEXTUAL measure (if latter, define relevant context)
- Use geographic information systems (GIS) to create exposure metrics
  - Get maps of points, lines, polygons, & metadata, or geocode data as needed
  - Create commonly used metrics: density, diversity, design, accessibility (distance, gravity), clustering
- Link exposure and patient measures
- Perform biostatistical analysis – person, place, time
The Food Environment

- Close up of a few of our 40 counties
- Two commercial data sources
  - D & B
  - InfoUSA
- Geocoded
- Used to create metrics
- Have available for four years, 1997-2010
A New MRSA: Community Associated (CA-)

- Since mid-1990s, large increase in MRSA infections in persons lacking prior contact with the healthcare system
- Shortly after, were recognized to be new MRSA strains
- Were rapidly disseminated among US general population, now affect patients with and without contact with healthcare system
- These new strains cause different clinical syndromes, particularly skin and soft tissue infections (SSTIs)
  - Incidence of SSTIs in US has been increasing
- These new strains now account for the majority of MRSA infections
- Large reservoirs of MRSA isolates now exist outside healthcare facilities

Identification of CA and HA MRSA Cases Using Electronic Health Records

PhD research of Joan A. Casey

Obtained Nutrient Management Plans for CAFOs, CAOs, VAOs, for swine & dairy/veal operations
n = 147 home + 271 importing fields with address; geocoded → circular buffer

n = 180 home operations with aerial photos; Google Earth → Shapefile → ArcMap

n = 131 importing fields, only township; randomly select point on appropriate land use type → circular buffer

Three methods for crop fields

Associations of Seasonal Crop Field Manure Exposure with HA-MRSA, CA-MRSA and SSTI (full multilevel modela)

<table>
<thead>
<tr>
<th></th>
<th>HA-MRSA</th>
<th>CA-MRSA</th>
<th>SSTI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adjusted OR (95% CI)</td>
<td>Adjusted OR (95% CI)</td>
<td>Adjusted OR (95% CI)</td>
</tr>
<tr>
<td><strong>Swine</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Q2</td>
<td>1.19 (0.97-1.46)</td>
<td>1.08 (0.89-1.31)</td>
<td>1.03 (0.88-1.20)</td>
</tr>
<tr>
<td>Q3</td>
<td>1.26 (1.03-1.55)</td>
<td>1.25 (1.04-1.52)</td>
<td>1.22 (1.05-1.41)</td>
</tr>
<tr>
<td>Q4</td>
<td>1.29 (1.04-1.60)</td>
<td>1.38 (1.13-1.68)</td>
<td>1.37 (1.18-1.60)</td>
</tr>
<tr>
<td>p</td>
<td>0.01</td>
<td>&lt; 0.001</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td><strong>Dairy/veal</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Q2</td>
<td>0.83 (0.68-1.03)</td>
<td>0.97 (0.80-1.18)</td>
<td>0.91 (0.78-1.06)</td>
</tr>
<tr>
<td>Q3</td>
<td>0.93 (0.76-1.13)</td>
<td>0.91 (0.75-1.10)</td>
<td>0.85 (0.73-0.99)</td>
</tr>
<tr>
<td>Q4</td>
<td>0.77 (0.62-0.97)</td>
<td>1.25 (1.02-1.53)</td>
<td>1.02 (0.87-1.19)</td>
</tr>
<tr>
<td>p</td>
<td>0.06</td>
<td>0.06</td>
<td>0.95</td>
</tr>
</tbody>
</table>

CA-MRSA = community-associated methicillin-resistant S. aureus; HA-MRSA = healthcare-associated MRSA; SSTI = skin and soft tissue infection

a Controlled for sex, age, race/ethnicity, ever-smoking status, antibiotic prescription in prior 2 years, community type, and community socioeconomic deprivation

b p-value for linear trend

In press, JAMA Internal Medicine
Coal Abandoned Mine Lands – the Contextual Effects of Chronic Environmental Contamination


Community Stress, Psychosocial Hazards, and EPA Decision-Making in Communities Impacted by Chronic Technological Disasters

Psychosocial stress has Stephen R. Couch, PhD, and Chariton J. Coles, PhD

Critical Biological Pathways for Chronic Psychosocial Stress and Research Opportunities to Advance the Consideration of Stress in Chemical Risk Assessment

Emerging evidence sug-

Cumulative Risk Assessment for Combined Health Effects From Chemical and Nonchemical Stressors

Cumulative risk assess-

What are the contextual effects of living in this community?

Pennsylvania has 5,000 abandoned coal mines, > 70% of all in U.S.

PhD research of Ann Y. Liu
Abandoned Coal Mine Lands (AML) in Pennsylvania

Abandoned structure

Acid mine drainage

Waste pile

Subsidence

Photos from PA DEP

When Does Diabetes Start?

Healthy → Pre-diabetes → Diabetes → Complications

100 mg/dL ≤ FBS ≤ 125 mg/dL → HbA1c (screening) → ICD-9 code → HbA1c (monitoring) → Treatment

mean duration = 159 days → mean duration = 117 days → mean duration = 1534 days

HbA1c (pre-therapeutic)

n = 7337, mean = 7.51%

1st ICD-9 diabetes code

n = 17,959, mean = 7.64%

HbA1c (post-ICD-9)

mean duration = 1732 days

HbA1c (last-ever)

The contextual influence of coal abandoned mine lands in communities and type 2 diabetes in Pennsylvania

Ann Y. Liu, Frank C. Curriero, Thomas A. Glass, Walter F. Stewart, Brian S. Schwartz

Health & Place 22 (2013) 115–122
EHR Challenges

- Patient must seek care
- Cannot exactly determine if patient is under observation and is well or has left care
- Persons enter and exit cohort at any time
- As a secondary data source, data are not perfect
  - ICD-9s have known problems; some data in text
- Large learning curve for use; much processing
- Many variables desired for analysis are not available (e.g., certain test results, income)
- Can be difficult to link patients (e.g., mother-child)
- No measures of environmental exposures
- “Should have validated against medical records …”

Parting Thoughts

- Big data from health care is now here; volumes will continue to grow with increasing migration to EHRs
- Tremendous opportunities going forward
  - Greatly expand the capacity to generate new knowledge
  - Helps with knowledge dissemination, including through messaging to physicians with data-driven information
  - May help with translating personalized medicine initiatives into clinical practice (e.g., eMERGE)
  - Empower patients to play a more active role in healthcare, direct data collecting and messaging
- We must increase efforts to use data for public health purposes as well
Thank you for listening

If you want copy of slides email: bschwart@jhsph.edu