Conducting Research with Large Administrative Health Care Databases: Challenges and Strategies

Lorene Nelson, PhD, MS
Associate Professor, Division of Epidemiology
Associate Director, Center for Population Health Sciences
Stanford University School of Medicine

RAC Meeting 8 August 2016 San Francisco, CA
Committee on Designing an Epidemiologic Study for Multiple Sclerosis and Other Neurologic Disorders in Veterans of the Persian Gulf and Post 9/11 Wars

Roderick J. Little, PhD (Chair)
University of Michigan

Babette Brumberg, PhD
University of Florida

Francesca Dominici, PhD
Harvard T.H. Chan School of Public Health

Elena Erosheva, PhD
University of Washington

Michael Goldberg, MD
Columbia University College of Physicians & Surgeons

Donald Hedeker, PhD
The University of Chicago Biological Sciences

Annette Langer-Gould, MD, PhD, MS
Kaiser Permanente Research

Lorene Nelson, PhD, MS
Stanford University School of Medicine

DeJuran Richardson, PhD
Lake Forest College

Ira Shoulson, MD
Georgetown University

Lawrence Steinman, MD
Stanford University

Barbara Vickrey, MD, MPH
Icahn School of Medicine at Mount Sinai

Christina Wolfson, PhD
McGill University

BOARD ON THE HEALTH OF SELECT POPULATIONS

The National Academies of Sciences • Engineering • Medicine
Public Law 1100-389 S.3023, enacted in 2008:

Directed the VA to contract with IOM to conduct an epidemiologic study to determine the incidence, prevalence and risk of developing multiple sclerosis (MS), and other neurologic diseases as a result of service in the 1990-1991 Gulf War or OEF / OIF / OND.

Other diseases the committee was to consider: Parkinson’s disease, brain cancers, migraine, and “central nervous system abnormalities that are difficult to precisely diagnose.”
Overview

- **Sources of bias by study type**
  - Selection bias
  - Confounding
  - Misclassification of health outcomes

- **Secondary Data Sources**
  - Healthcare utilization data (claims data, HMO data)
  - Electronic health record (EHR) data
  - Death certificates

- **Examples**
  - Estimating national prevalence of MS
  - Following Gulf War Veteran cohort for health outcomes

- **Conclusions**
Gulf War Cohorts – 1990-1991

621,901 GW veterans (deployed)
- exposed to oil well fires +/- nerve gas (~29%)
- not exposed to fires or nerve gas (~71%)

747,247 GW era veterans (nondeployed)
- not exposed to Persian Gulf region
Types of Study Designs: Surveys

**Surveys**

- National Health Survey of Persian GW Veterans
  - 11,441 GW vets (70%)
  - 9,476 GW era vets (64%)

- Longitudinal Health Study of Persian GW Veterans
  - 6,111 GW vets (41%)
  - 3,859 GW era vets (26%)
  - GW vets – 37% MSI
  - GW era vets – 12% MSI

- Follow-up Study of Gulf War and Gulf Era Veterans
  - 8,104 GW vets (57%)
  - 6,198 GW era vets (43%)
  - GW vets – 44% GWI
  - GW era vets – 20% GWI

621,901 GW veterans (deployed)

747,247 GW era vets (nondeployed)

- 1990-91
- 1993-95
- 2003-05
- 2012-13

[2016]

[Kang et al. JOEM, 2000] [Kang, et al. JOEM, 2009] [Dursa et al., JOEM 2016]
Types of Study Designs: Data Linkages

- 621,901 GW veterans (deployed)
- 747,247 GW era vets (nondeployed)

Data Linkages:
- Mortality data: 2004
- VHA healthcare data: 2002, 2011
- VA MS study: 2007
- Cancer registry: 2006
## Possible Biases by Type of Study Design

<table>
<thead>
<tr>
<th>Type of Study Bias</th>
<th>Surveys</th>
<th>Healthcare Utilization Data Linkages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selection bias</strong> - subjects are represented in a study in such a way that they do not represent their original cohort.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Volunteer bias</td>
<td>+++</td>
<td></td>
</tr>
<tr>
<td>Differences in VHA eligibility or coverage, or selective attrition by cohort</td>
<td></td>
<td>+++</td>
</tr>
</tbody>
</table>

| **Con founding** - a factor that is associated with the exposure and also with the disease, causing a spurious exposure-disease association |
| | ++ | +/− |

| **Measurement error** - Misclassification of the study subject into the wrong category (unexposed vs exposed, diseased vs diseased) |
| Recall bias | +++ | |
| Disease misclassification | ++ | ++ |
### Sources of Bias by Type of Study

<table>
<thead>
<tr>
<th>Type of Study Bias</th>
<th>Surveys</th>
<th>Mortality data</th>
<th>VHA Data Linkages</th>
<th>Disease Registries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection bias</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confounding</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Misclassification</td>
<td></td>
<td><strong>Strong</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Strength / Likelihood of Bias

- **Strong**
- **Moderate**
- **Weak**
### Misclassification of Health Conditions by Type of Study

<table>
<thead>
<tr>
<th>Health Condition</th>
<th>Surveys</th>
<th>Mortality data</th>
<th>VHA Data Linkages</th>
<th>Disease Registries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gulf war illness</td>
<td>+++</td>
<td>-</td>
<td>+/-</td>
<td>N/A</td>
</tr>
<tr>
<td>Migraine</td>
<td>+++</td>
<td>-</td>
<td>-</td>
<td>N/A</td>
</tr>
<tr>
<td>MS</td>
<td></td>
<td>-</td>
<td>++</td>
<td>N/A</td>
</tr>
<tr>
<td>PD</td>
<td></td>
<td>-</td>
<td>++</td>
<td>N/A</td>
</tr>
<tr>
<td>ALS</td>
<td></td>
<td>++</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Cancer</td>
<td></td>
<td>-</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Brain tumor</td>
<td></td>
<td>++</td>
<td>++</td>
<td>+++</td>
</tr>
</tbody>
</table>

*too rare in survey samples*

+++ best method          - weakest method
Overview

- Sources of bias by study type
  - Selection bias
  - Confounding
  - Misclassification of health outcomes

- Secondary Data Sources
  - Healthcare utilization data (claims data, HMO data)
  - Electronic health record (EHR) data
  - Death certificates

- Examples
  - Estimating national prevalence of MS
  - Following Gulf War Veteran cohort for health outcomes

- Conclusions
Types of “Secondary” Health Care Data

- Administrative Healthcare Data (Claims)
- Electronic Health Records
- Death Certificates
Strengths & Limitations

**Strengths:**
- Complete coverage of deaths
- Searchable for primary and underlying causes of death
- Better for conditions with high case fatality (ALS, brain tumor)

**Limitations:**
- Variable coding practices
- Many chronic conditions underascertained (MS, PD)
Unbiased Data Linkage for Death Certificates

- 621,901 GW veterans (deployed)
- 747,247 GW era vets (nondeployed)

1990-91

Mortality data

~ 100% linkage

2004

2011
Administrative healthcare data (claims)

- Health-care data collected for payment for medical services
  - hospital admissions,
  - outpatient visits,
  - diagnoses
  - tests,
  - procedures
  - drugs
- Represented as ICD-9 codes, CPT codes, HCPCS codes

**Strengths:**
- Provides largest populations (e.g., Medicare, commercial)
- Reasonably accurate data on: enrollment, medications, procedures, hospital outcomes

**Limitations:**
- ICD-9 diagnoses not always accurate
- Lab results usually not available
- No physical measures (BMI, BP, ..)
<table>
<thead>
<tr>
<th>Government</th>
<th>Population-based?</th>
<th>Hospital (inpatient)</th>
<th>Physician (outpatient)</th>
<th>Medications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medicare - national</td>
<td>Yes, &gt; 65 years</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Medicaid - national (by state)</td>
<td>No, low income &amp; disabilities</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>VHA - national</td>
<td>No, veterans</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Indian Health Service - national</td>
<td>No, Indian and Alaskan natives</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>National Hospital Discharge Survey</td>
<td>Yes (probability sample)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Private Organizations</th>
<th>Population-based?</th>
<th>Hospital (inpatient)</th>
<th>Physician (outpatient)</th>
<th>Medications</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMO organizations – select regions</td>
<td>No</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Commercial insurance claims – select regions</td>
<td>No</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Possible Biases in the VHA Care Subset?

621,901 GW veterans (deployed)

747,247 GW era vets (nondeployed)

VHA healthcare data 2002 - 2013

2002: 286,955 (46%)

2013: 269,635 (36%)
Strengths & Limitations

**Strengths:**
- Has lab results, physiologic measures
- Can search progress notes (NLP) for detailed clinical info

**Limitations:**
- Enrollment not always known (no denominator)
- Drug prescriptions, not fills
- Missing data
Overview

- Sources of bias by study type
  - Selection bias
  - Confounding
  - Misclassification of health outcomes

- Secondary Data Sources
  - Healthcare utilization data (claims data, HMO data)
  - Electronic health record (EHR) data
  - Death certificates

- Examples
  - Estimating national prevalence of MS
  - Following Gulf War Veteran cohort for health outcomes

- Conclusions
Estimating the National Prevalence of MS 2008-2010

- Other than SEER cancer registry (and now ALS National Registry), no nationwide registries for chronic diseases.

- Much of what we know about the descriptive epidemiology of MS comes from intensive studies in small populations.

- Very hard to get information on temporal trends, differences according to race/ethnicity.

- How do we do this in a fragmented U.S. health care system?

National MS Society MS Prevalence Working Group
U.S. Sources of Health Insurance Coverage, 2007

How do we access these data?

2009
Age < 65

Federal includes Medicaid, Medicare, CHIP and other means tested programs, VA, DoD, Tri-care.

Commercial claims databases (Optum, Truven, IMS Health)

State all-payer databases

Employer-sponsored 57%

Uninsured 18%

Federal [PERCENTAGE]

Self pay [PERCENTAGE]
MS misclassification when using claims data

<table>
<thead>
<tr>
<th></th>
<th>MS</th>
<th>Not MS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gold Standard</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Not MS</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

- **Claims data algorithm**

- **Sensitivity** ---\( \frac{a}{a+c} \) = 87%
- **Positive predictive value** ---\( \frac{a}{a+b} \) = 98%
- **Specificity** ---\( \frac{d}{b+d} \) = 83%

> 2 inpatient codes, or > 3 outpatient codes or > 1 disease modifying treatments
PD misclassification when using claims data

Gold Standard

<table>
<thead>
<tr>
<th></th>
<th>MS (a)</th>
<th>Not MS (b)</th>
<th>MS (c)</th>
<th>Not MS (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not MS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**positive predictive value**

\[
\frac{a}{a+b}
\]

- **MS**: 98%
- **PD**: ~80-85%

**Sensitivity**

\[
\frac{a}{a+c}
\]

- **MS**: 87%
- **PD**: ~70-73%

**Specificity**

\[
\frac{d}{b+d}
\]

- **PD**: ~80-85%

Claims data algorithm

- > 2 inpatient codes, or > 3 outpatient codes or > 1 disease modifying treatments

medical record or EMR review
Overview

- Sources of bias by study type
  - Selection bias
  - Confounding
  - Misclassification of health outcomes

- Secondary Data Sources
  - Healthcare utilization data (claims data, HMO data)
  - Electronic health record (EHR) data
  - Death certificates

- Examples
  - Estimating national prevalence of MS
    - Following Gulf War Veteran cohort for health outcomes

- Conclusions
Presence of One or More ICD-9 Code for 4 Diseases Among GW Veterans and GW Era Veterans (2002-2013)

<table>
<thead>
<tr>
<th>Disease</th>
<th>Gulf War Deployed (286,995)</th>
<th>Gulf War Nondeployed (269,635)</th>
<th>Unadjusted Odds Ratio (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migraines</td>
<td>16,327</td>
<td>14,115</td>
<td>1.09 [1.07-1.12]</td>
</tr>
<tr>
<td>Multiple sclerosis</td>
<td>1,040</td>
<td>1,089</td>
<td>0.90 [0.82-0.98]</td>
</tr>
<tr>
<td>Parkinson’s disease</td>
<td>403</td>
<td>487</td>
<td>0.78 [0.68-0.89]</td>
</tr>
<tr>
<td>Brain tumor</td>
<td>342</td>
<td>332</td>
<td>0.97 [0.83-1.13]</td>
</tr>
</tbody>
</table>
Conclusions

- Challenging area of research

- Possible future approaches:
  - Continue follow-up of original cohorts using existing methods at periodic intervals.
  - Use most sensitive and specific case-finding algorithms when identifying health outcomes in utilization data.
  - Link subset with survey data with VHA health care utilization data (better control for confounding variables)
  - Investigate electronic medical records sources of data to reduce misclassification (VINCI).
Thank You

National MS Prevalence Working Group (NMSS)

IOM committee