

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 Brumback, 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

Public Law 110-389 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."

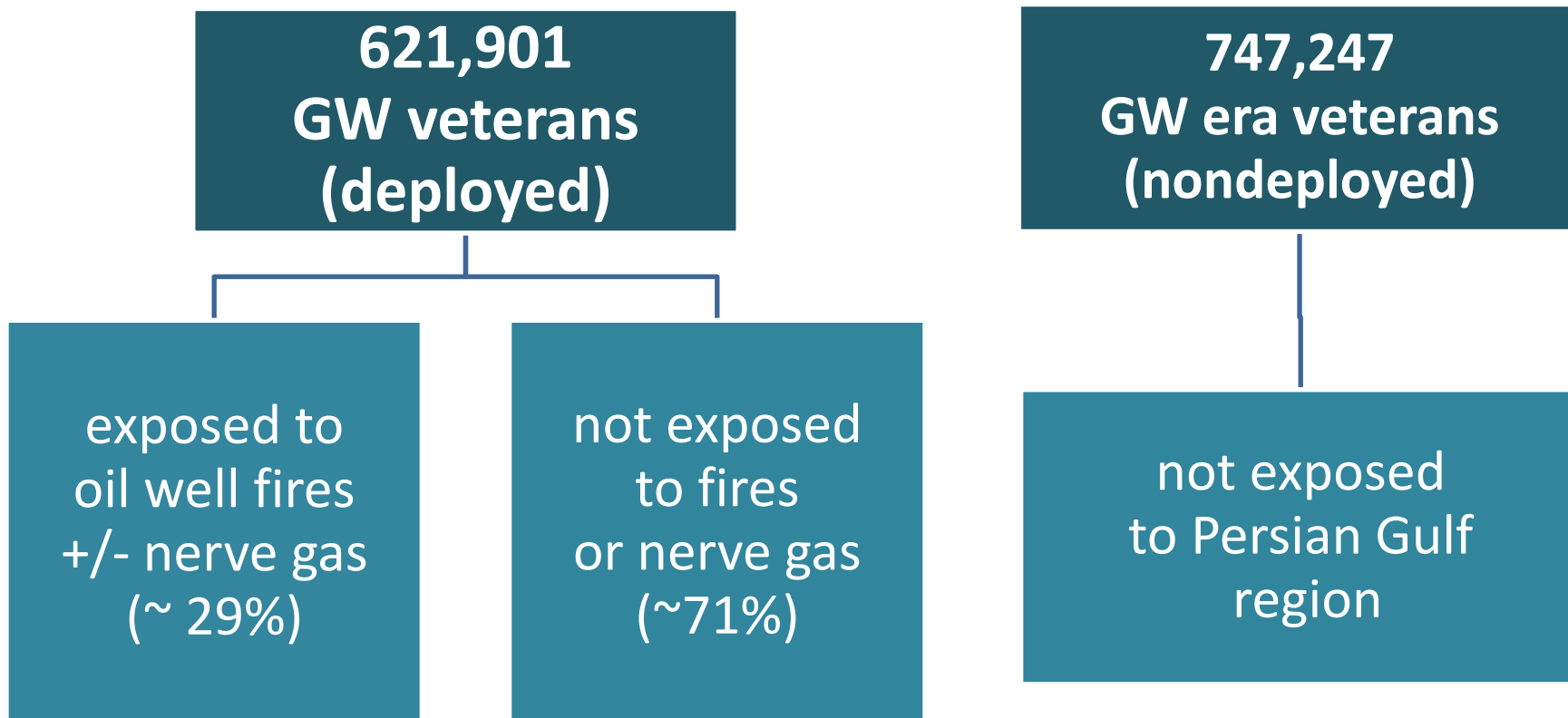
BOARD ON THE HEALTH OF SELECT POPULATIONS

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

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**
- 

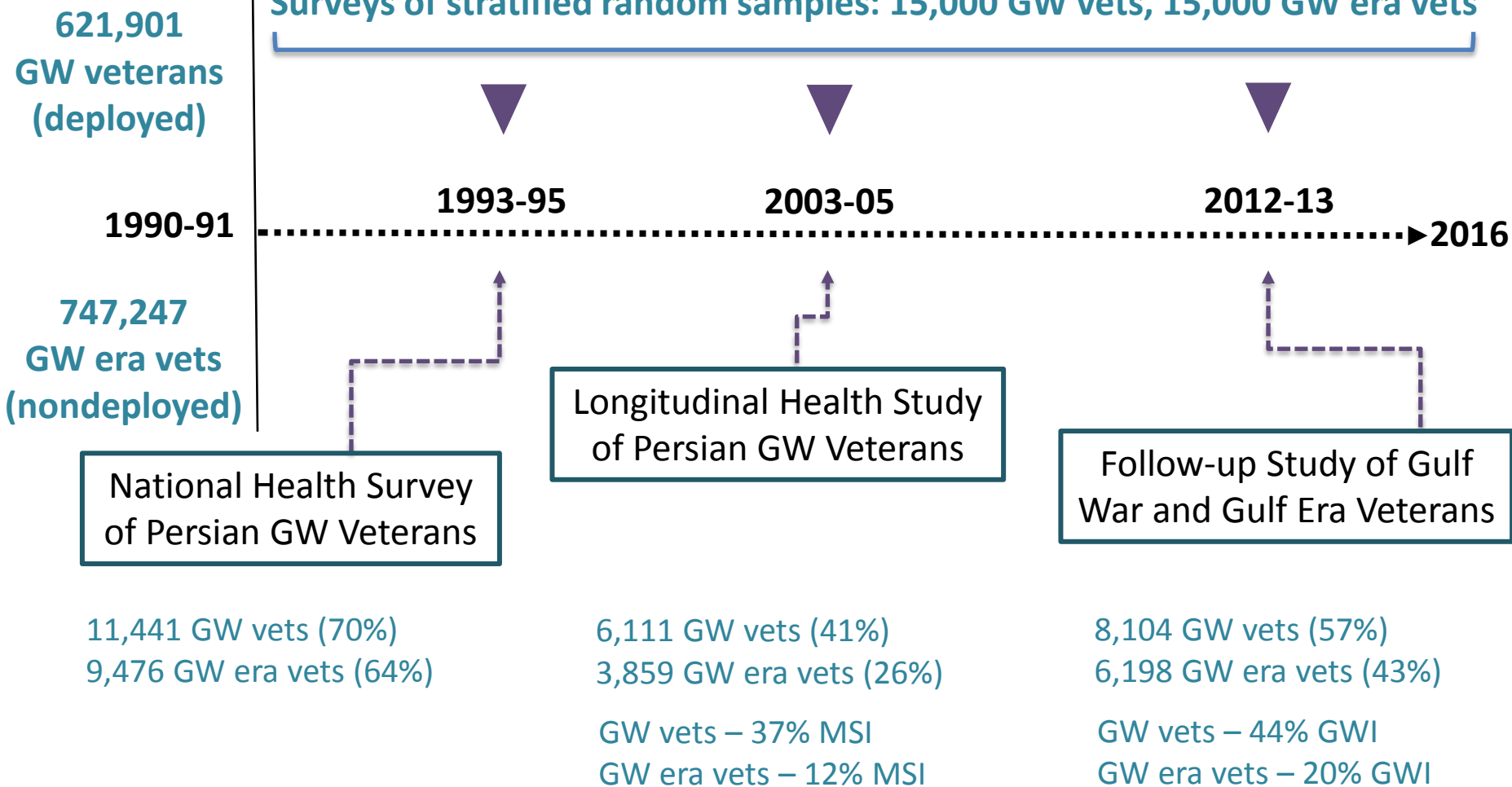
Gulf War Cohorts – 1990-1991



Types of Study Designs: Surveys

Surveys

Surveys of stratified random samples: 15,000 GW vets, 15,000 GW era vets

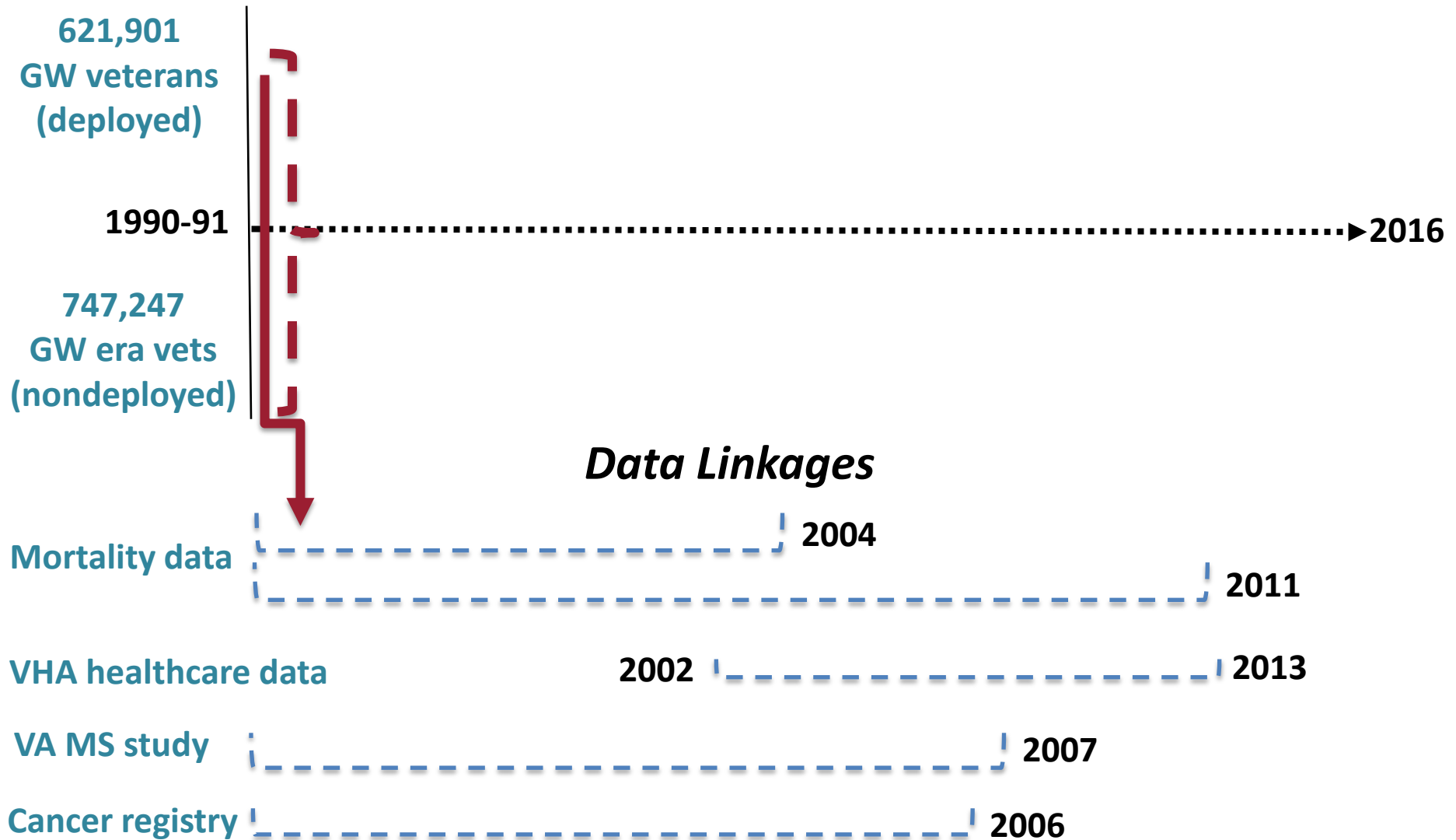


[Kang et al. JOEM, 2000]

[Kang, et al. JOEM, 2009]

[Dursa et al., JOEM 2016]

Types of Study Designs: Data Linkages



Possible Biases by Type of Study Design

Type of Study Bias	Surveys	Healthcare Utilization Data Linkages
Selection bias - subjects are represented in a study in such a way that they do not represent their original cohort.		
Volunteer bias	+++	
Differences in VHA eligibility or coverage, or selective attrition by cohort		+++
Confounding - 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

Type of Study Bias	Surveys	Mortality data	VHA Data Linkages	Disease Registries
Selection bias	Strong	Weak	Moderate	Weak
Confounding	Weak	Moderate	Moderate	Weak
Misclassification	Weak	Strong	Strong	Weak



Strength / Likelihood of Bias
Strong
Moderate
Weak

Misclassification of Health Conditions by Type of Study

Health Condition	Surveys	Mortality data	VHA Data Linkages	Disease Registries
Gulf war illness	+++	-	+/-	N/A
Migraine	+++	-	-	N/A
MS	<i>too rare in survey samples</i>	-	++	N/A
PD		-	++	N/A
ALS		++	++	+++
Cancer		-	++	+++
Brain tumor		++	++	+++

+++ 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

Death Certificates

- Data
 - ◆ primary cause of death
 - ◆ underlying cause of death
 - ◆ demographic variables
- ICD-9 coded

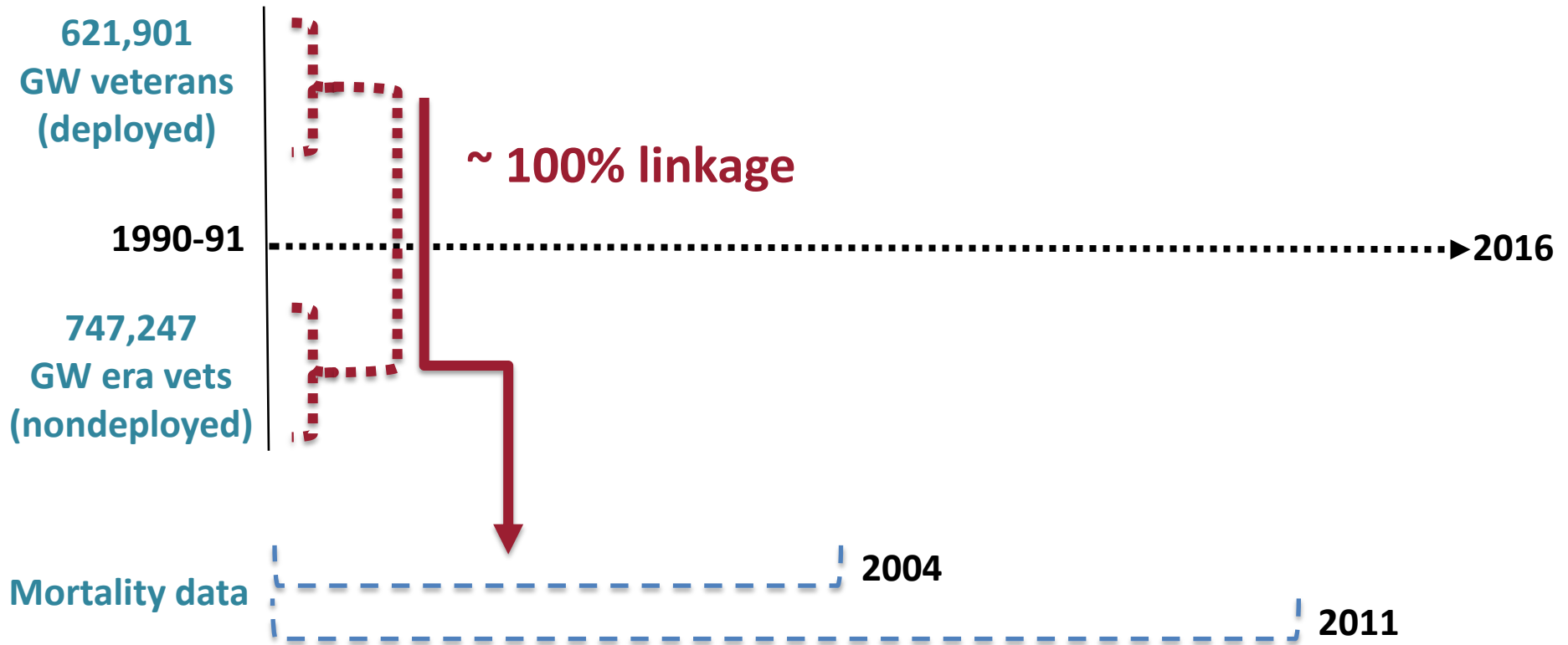
Strengths:

- 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



Strengths & Limitations

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

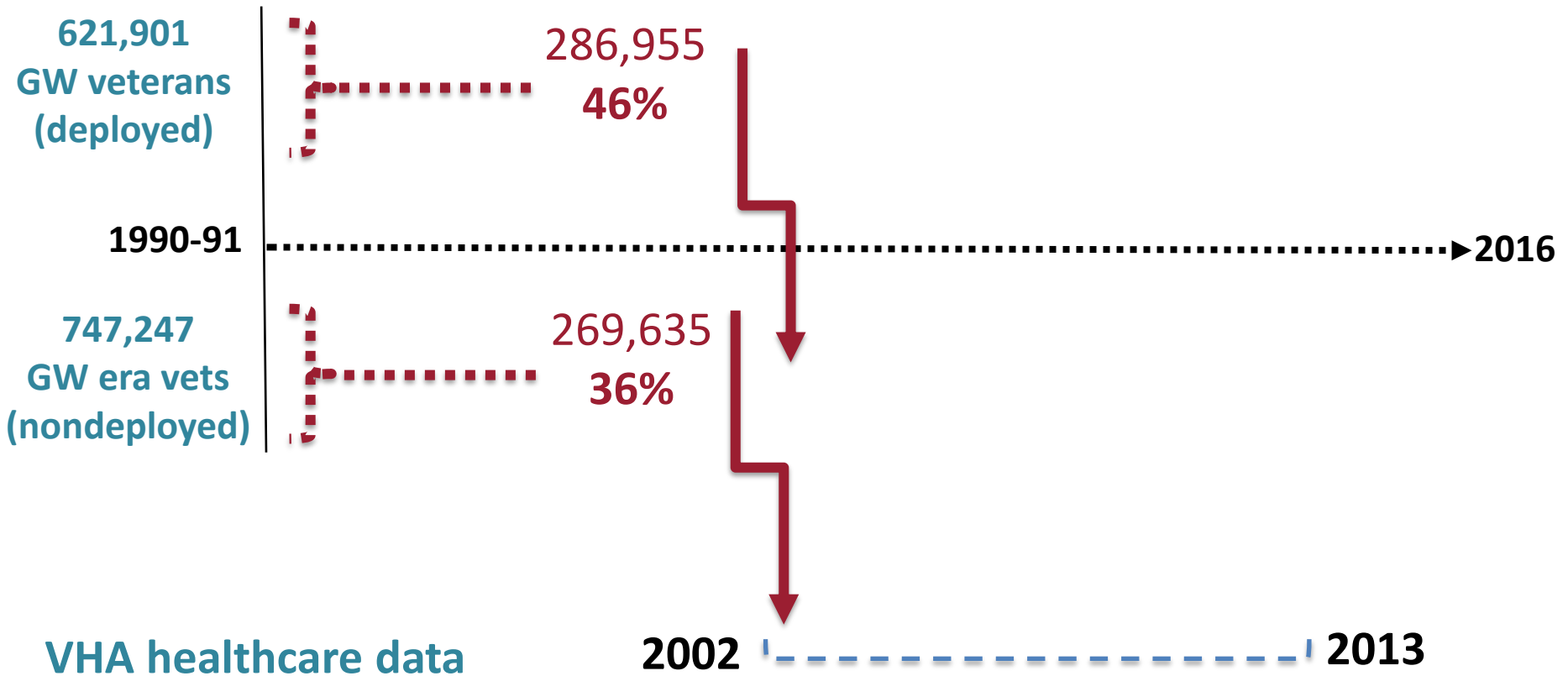
Limitations:

- ICD-9 diagnoses not always accurate
- Lab results usually not available
- No physical measures (BMI, BP,..)

Types of Administrative Health Care Data

Government	Population-based?	Hospital (inpatient)	Physician (outpatient)	Medications
Medicare - national	Yes, ≥ 65 years	X	X	X
Medicaid- national (by state)	No, low income & disabilities	X	X	X
VHA - national	No, veterans	X	X	X
Indian Health Service - national	No, Indian and Alaskan natives			
National Hospital Discharge Survey	Yes (probability sample)	X		
Private Organizations	Population-based?	Hospital (inpatient)	Physician (outpatient)	Medications
HMO organizations – select regions	No	X	X	X
Commercial insurance claims – select regions	No	X	X	X

Possible Biases in the VHA Care Subset?



Strengths & Limitations

Electronic Health Record

- Medical record generated at the point of care
 - ◆ text notes
 - ◆ laboratory test results
 - ◆ imaging results
- Standard format for data exchange (HL-7), rest of record unstructured

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

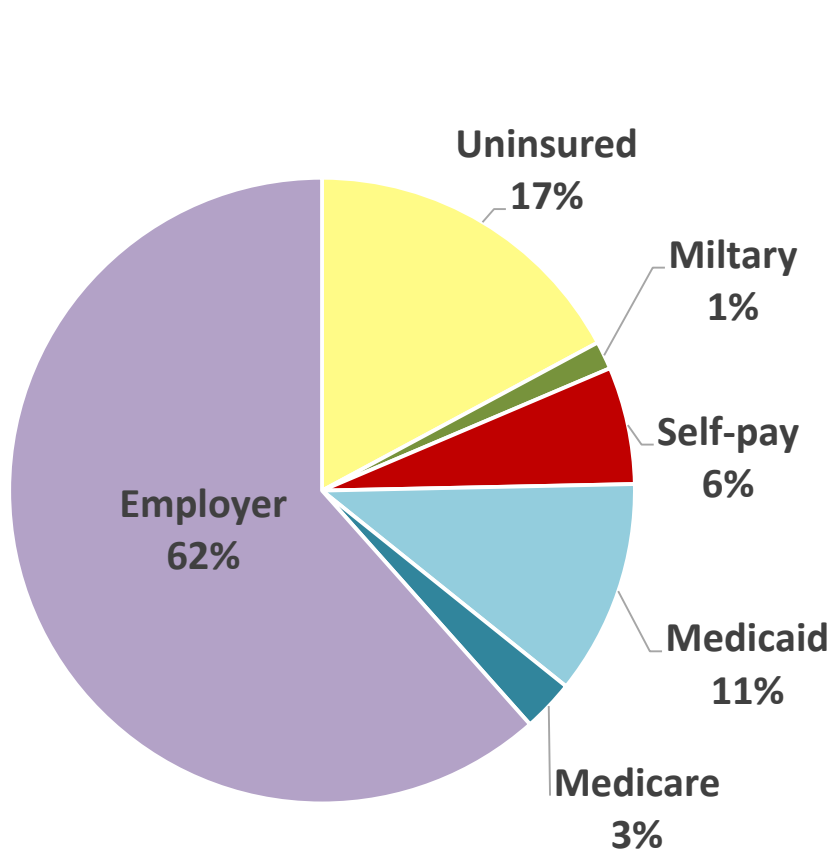
- **Sources of bias by study type**
 - Selection bias
 - Confounding
 - Misclassification of health outcomes
- **Secondary Data Sources**
 - Healthcare utilization data (claims data, HMO data)
 - Electronic health record (EHR) data
 - Death certificates
- **Examples**
 - Estimating national prevalence of MS
 - Following Gulf War Veteran cohort for health outcomes
- **Conclusions**

Estimating the National Prevalence of MS 2008-2010

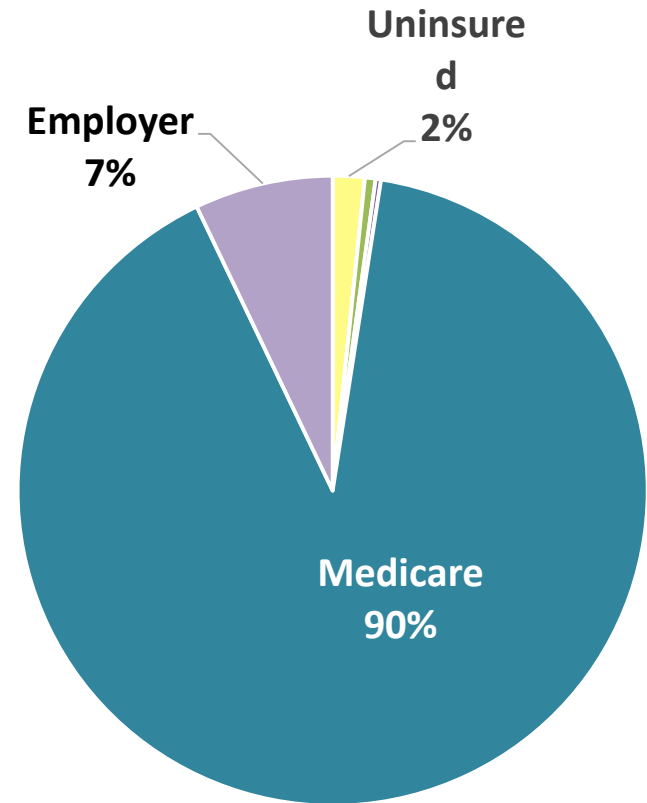
- Other than SEER cancer registry (and now ALS National Registry), no nationwide registries for chronic diseases.
- Much of what we know about the descriptive epidemiology of MS comes from intensive studies in small populations
- Very hard to get information on temporal trends, differences according to race/ethnicity
- How do we do this in a fragmented U.S. health care system?



U.S. Sources of Health Insurance Coverage, 2007



Age < 65

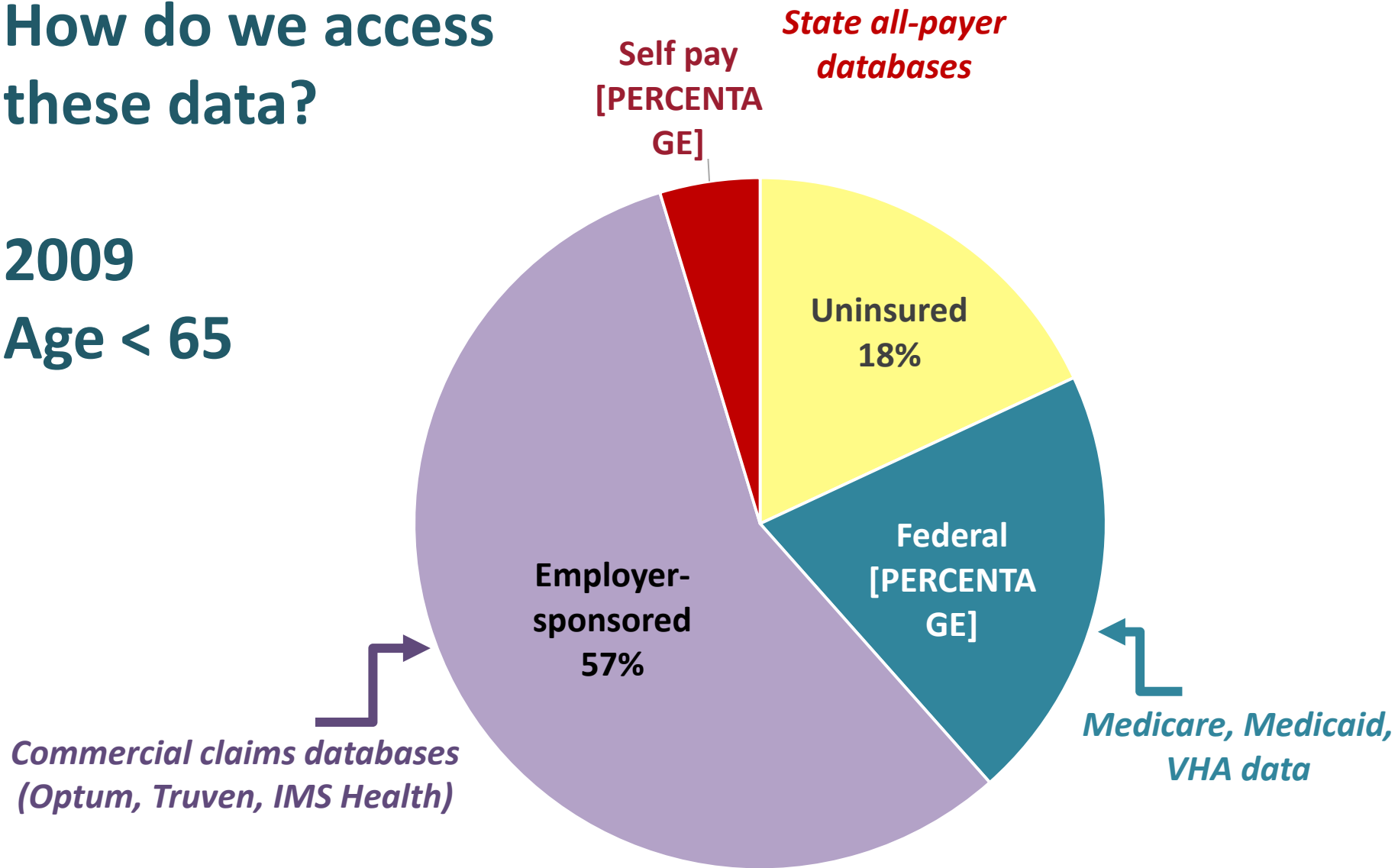


Age ≥ 65

Based on S. R. Collins, C. White, and J. L. Kriss, *Whither Employer-Based Health Insurance? The Current and Future Role of U.S. Companies in the Provision and Financing of Health Insurance* (New York: The Commonwealth Fund, Sept. 2007) and analysis of the Current Population Survey, March 2008, by Bisundev Mahato of Columbia University.

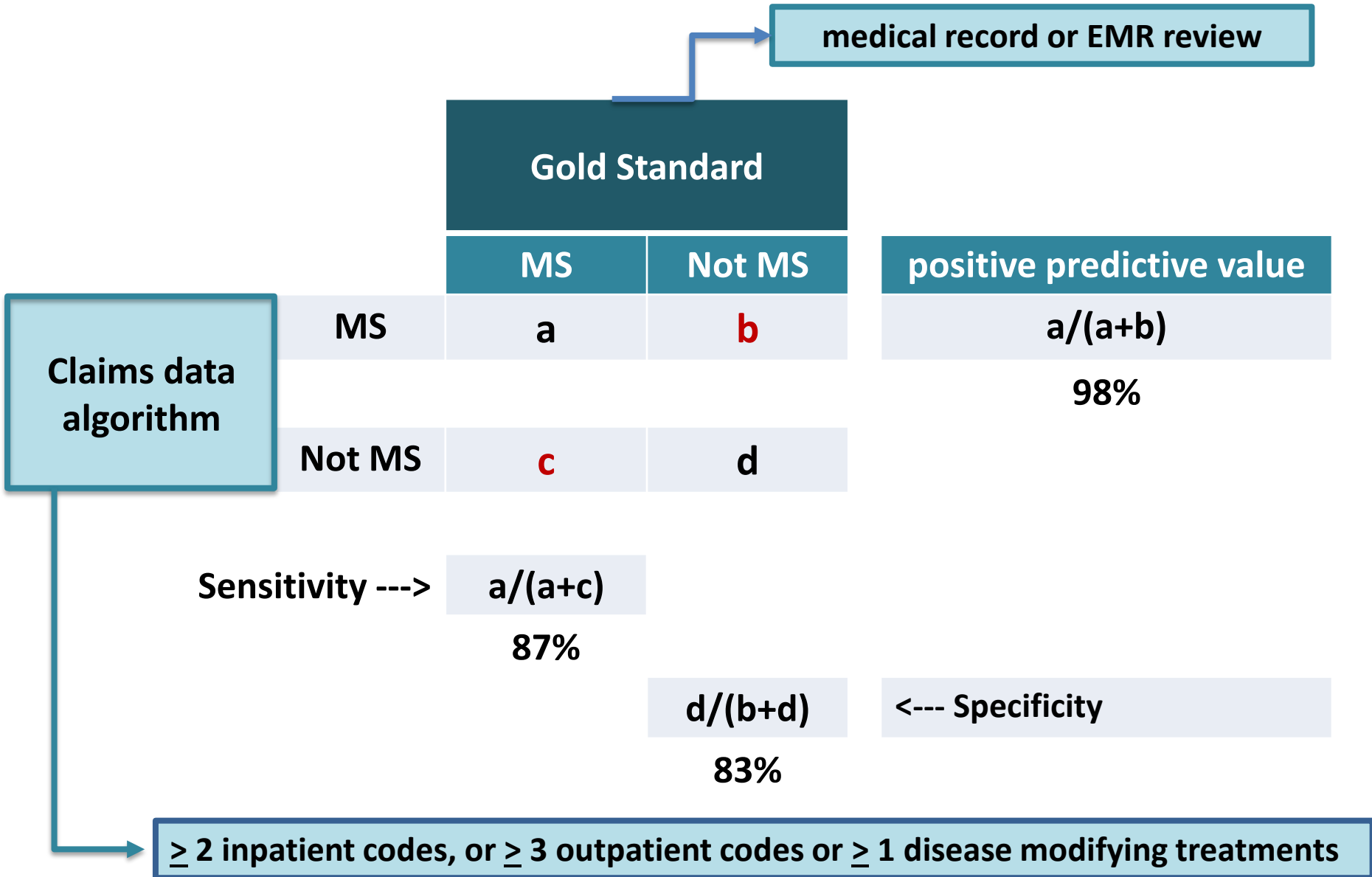
How do we access these data?

2009
Age < 65



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

MS misclassification when using claims data



PD misclassification when using claims data

medical record or EMR review

Gold Standard

	MS	Not MS
MS	a	b
Not MS	c	d

positive predictive value

$a/(a+b)$

MS 98%

PD ~ 80-85%

Claims data algorithm

Sensitivity ---> $a/(a+c)$

MS 87%

PD ~70-73%

$d/(b+d)$

83%

<--- Specificity

PD ~ 80-85%

≥ 2 inpatient codes, or ≥ 3 outpatient codes or ≥ 1 disease modifying treatments

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)

	Gulf War Deployed (286,995)	Gulf War Nondeployed (269,635)	Unadjusted Odds Ratio (95% CI)
Migraines	16,327	14,115	1.09 [1.07-1.12]
Multiple sclerosis	1,040	1,089	0.90 [0.82-0.98]
Parkinson's disease	403	487	0.78 [0.68-0.89]
Brain tumor	342	332	0.97 [0.83-1.13]

BOARD ON THE HEALTH OF SELECT POPULATIONS

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