Challenges to Connecting Data Across Agencies: Health Claims and other CMS data

Beth Virnig, PhD, MPH

Professor

PI, Research Data Assistance Center

SCHOOL OF
PUBLIC HEALTH
UNIVERSITY OF MINNESOTA

CMS: Centers for Medicare and Medicaid Services

- Source for Medicare and Medicaid data

- Directly manage Medicare program
  - Almost all elders 65+
  - Select people under age 65 with disabilities, ESRD, ALS

- Work with states to manage Medicaid program(s)
  - Select low income people

What type of data does CMS provide?

- Enrollment information
  - Who is enrolled in each program? Monthly status
  - Demographic information
  - Dates of birth and death
- Health claims data
  - Record of bills/claims/encounters for health services
  - Covered benefits
- Assessments
  - SNF
  - Home health services
  - Inpatient Rehab services

Linking with CMS data

- Policies:
  - Limited number of possible matching variables
  - Names are never released
  - Exact addresses are not released
  - Finder file sent to CMS contractor, validated matches returned.
  - Only returned if a single match is identified
    - e.g., if there are two potential matches, neither is returned

Linking…

Linking CMS data

Medicare

- Social Security Number
  - Personal or used to justify benefits
- Medicare Beneficiary ID (on the Medicare card)
  - used after 2018
- HIC (Health Insurance Claim Number)
  - used before 2018
- Name plus date of birth

Medicaid

- State Medicaid ID
  - Can have multiple within and across states
- Social Security number
  - Not all states collect
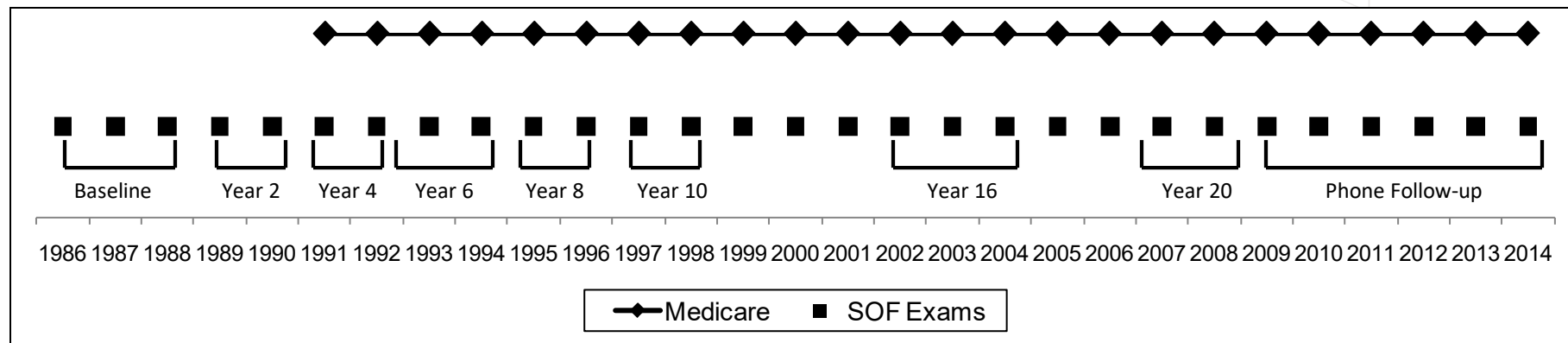- Name plus date of birth

MrOS & Medicare Claims Data Collection Timeline
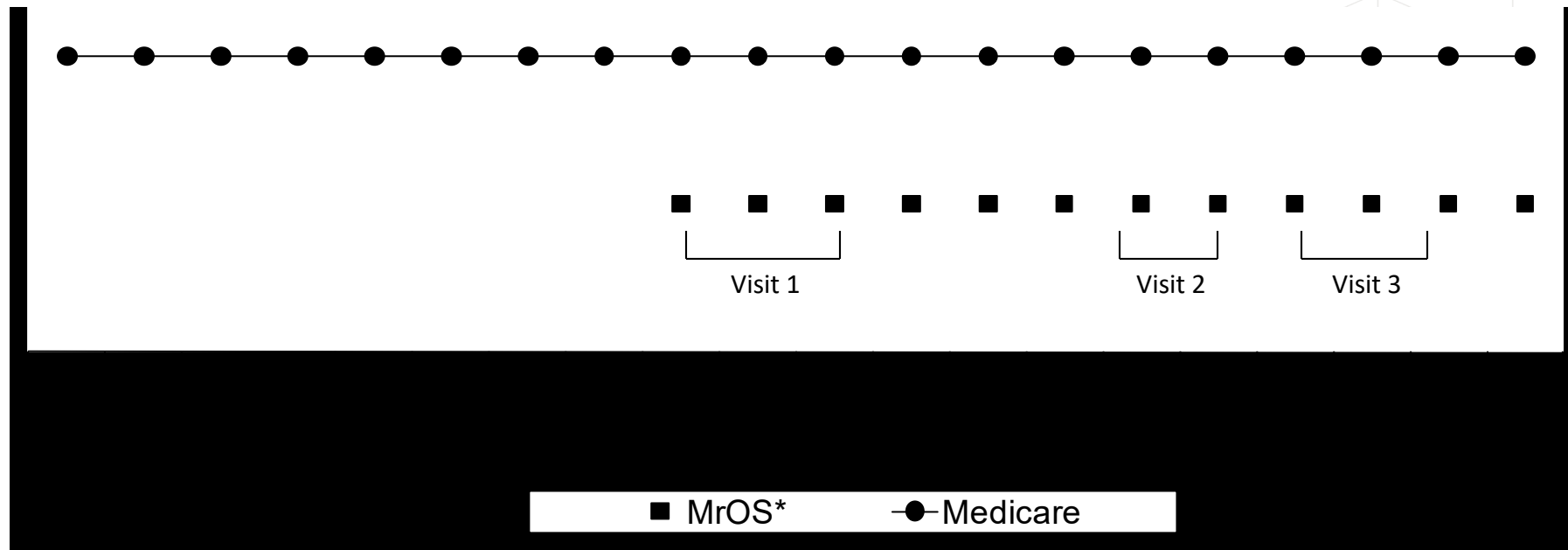
*5,995 men were recruited between 1999-2001

NOTE: MrOS participants are followed-up every 4 months via postcard questionnaire to track endpoints of falls, incident non-spine fractures, hip fractures, and deaths

SOF & Medicare Claims Data Collection Timeline

*9,704 Caucasian women were recruited between 1986-1988; an additional 662 African American women were recruited in 1997-1998 for Visit 6

NOTE: Until Fall 2009, SOF participants were followed-up every 4 months via postcard questionnaire to track endpoints of falls, incident non-spine fractures, hip fractures, and deaths. Starting Fall 2009, SOF participants will be followed-up every 6 months via questionnaire administered on the telephone and expand our assessment of factors related to successful aging, including whether or not a participant lives independently in the community, their ability to perform activities of daily living, and measures of optimism, resilience and social engagement.

Figure 1: Availability of survey and Medicare data
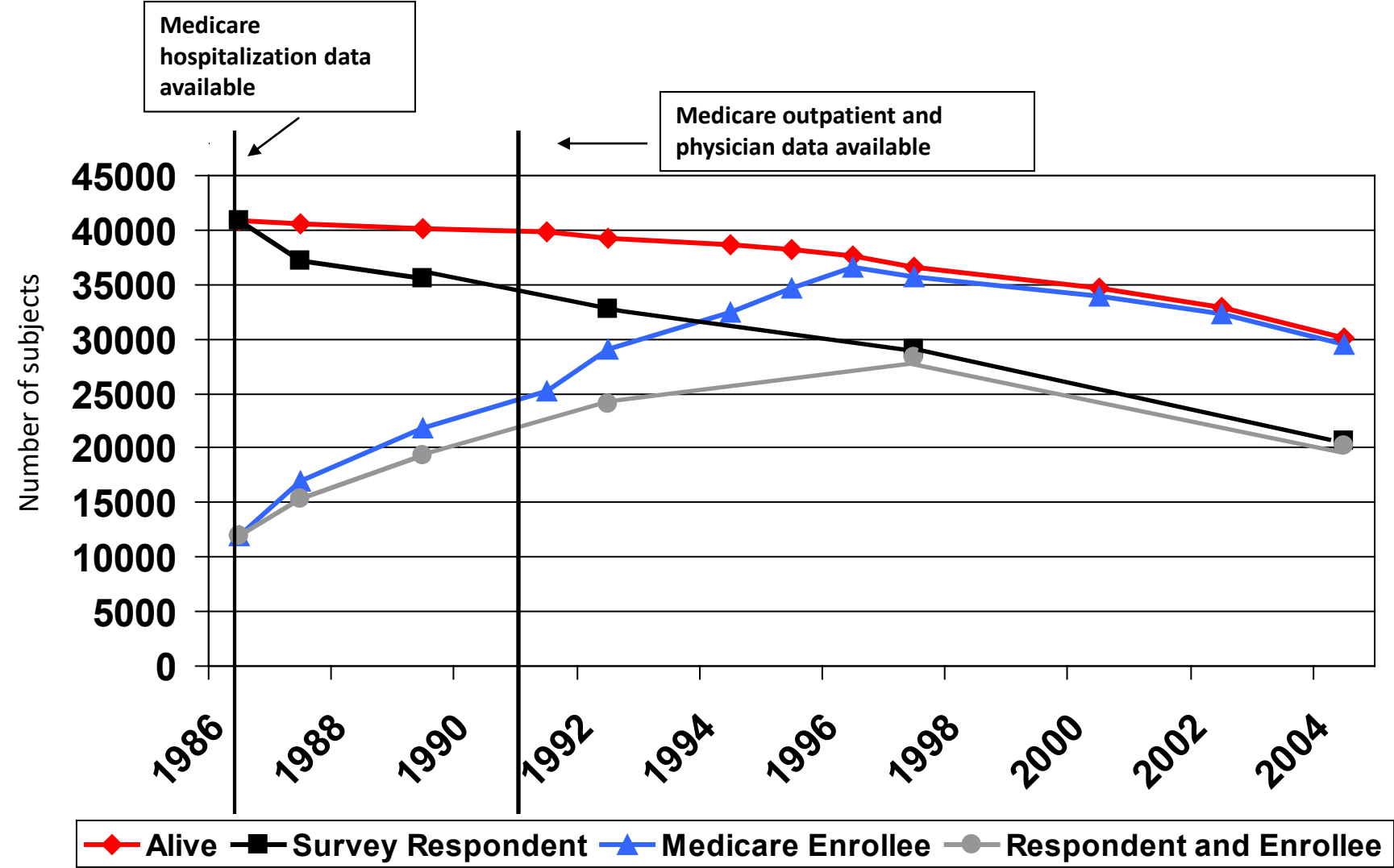
Challenges that need to be considered around linking

- Left truncation—gap before CMS observation but then no gap (both, but Medicare in particular)

- Interval censoring—mix of observations and gaps (Medicaid)

- Impact on sampling weights

Working with the claims linked data

Key variables

- Beneficiary demographics
- Coverage options—FFS, MC, pharmacy, etc.
- Diagnosis (ICD-9/ICD-10 codes)
- Date(s) of service
- Procedure (ICD-9/ICD-10 or CPT/HCPCS codes depending on location of care)
- Types of care: ER services, hospice, rehabilitation services
- Provider (CCN, UPIN)

Data are collected in real-time

- Advantage—no recall bias
- Disadvantage—no correction of past errors such as missed diagnosis
- Data are a record of what was done, not what should have been done, was intended to be done, what we now realize to be important, etc.
- Use caution when looking for claims-based measures of behaviors
  - Smoking, obesity, drug and alcohol use
- Conditions that are underdiagnosed in clinical settings will be under-represented in claims
  - Dementia!!!

Key difference vs. surveys: Variables need to be created

- Both an advantage and disadvantage
- Can use dates to align health information with survey dates
- Can sequence events—A happened before B, A and B happened together, etc.
- Can explore new hypotheses that were not considered at the time of survey launch (without worry about recall bias)
- But... variables typically need to be created. There is no "had a stroke in 2016" variable
- Flexibility provides tremendous opportunity but can be overwhelming
- Using the data requires consideration of the structure of the healthcare system, thoughts about insurance and payments, etc.

The Medicare structure is challenging to work with...

File	Structure	Frequency	% FFS with any record in a CY	Median # records per CY
Master Beneficiary Summary File	Base plus segments	One record per enrollee per year	100%	1
MedPAR	Flat file	One record per hospitalization	14.6%	1
Outpatient	Base plus segments	One base per bill	71.8%	4
Carrier	Base plus segments	One base per bill	100%	16
Durable Medical Equipment	Base plus segments	One base per bill	23.4%	3
Skilled Nursing Facility	Base plus segments	One base per bill	4.1%	2
Home Health	Base plus segments	One base per bill	8.1%	1
Hospice	Base plus segments	One base per bill	3.9%	2
Pharmacy	Very complex	One per fill		

For longitudinal studies

- Medicare coverage, payment and reporting policies change over time

- Codes for procedures and diagnoses change over time
  - ICD-9 to ICD-10 conversion in October 2015
  - CPT codes updated annually or more often
  - COVID codes created April 2020 and updated after that

- Medicaid policies vary by both state and time and do not have common claims submission and data processing rules.

Other CMS considerations

- Re-release policy—cell sizes <11 must be suppressed

- Non-US located researchers may not access data

- Structured request process—forms, approvals, etc

Resources

- Research Data Assistance Center (ResDAC)
- www.resdac.org
- Website with many resources
- Help Desk that will assist with data requesting and questions about data use

Questions?

Beth Virnig
virni001@umn.edu
resdac@umn.edu **attention Beth Virnig**