Duke Clinical Research Institute

Linking Large Datasets to Advance Population Health: Challenges and Opportunities

Emily O'Brien, PhD

March 28, 2017

Outline

- What is data linkage and why do we use it?
- Existing and evolving data linkage methods
- Challenges of linkage
- Real world examples of linking datasets

What do we mean by “data linkage”?

- ***Definition*: a process of pairing records from two files and trying to select the pairs that belong to the same entity**
- Linkage may occur:
  1. ***Within*** the same dataset to combine multiple *records* per patient
  2. ***Between*** two or more different datasets to combine *different types of data* about a single patient

Within a Dataset

Pt #	Diagnosis
1	STEMI
2	STEMI
2	T2DM
3	UA

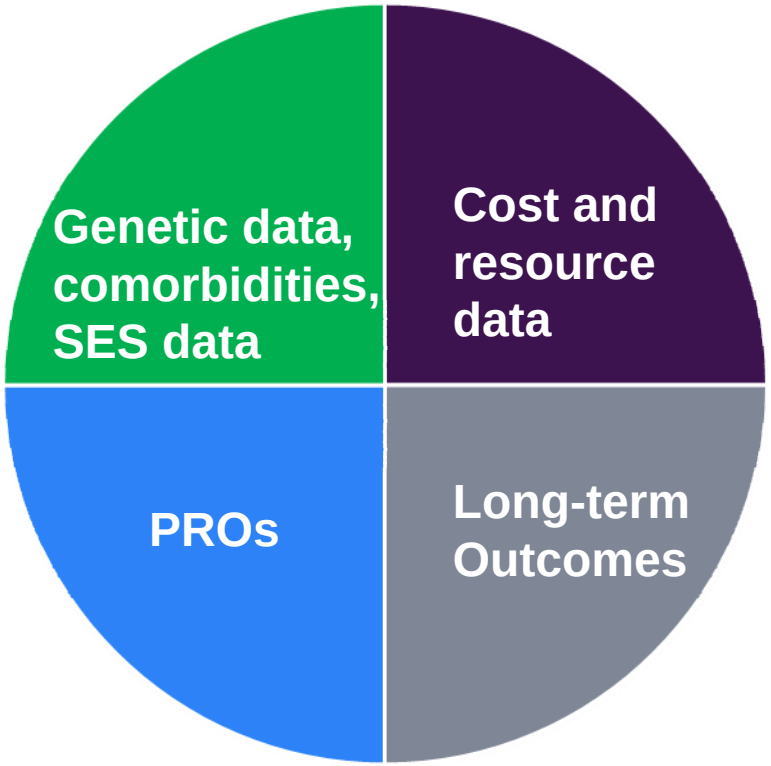
Between Datasets

Pt #	Diagnosis
1	STEMI
2	STEMI
2	T2DM
3	UA

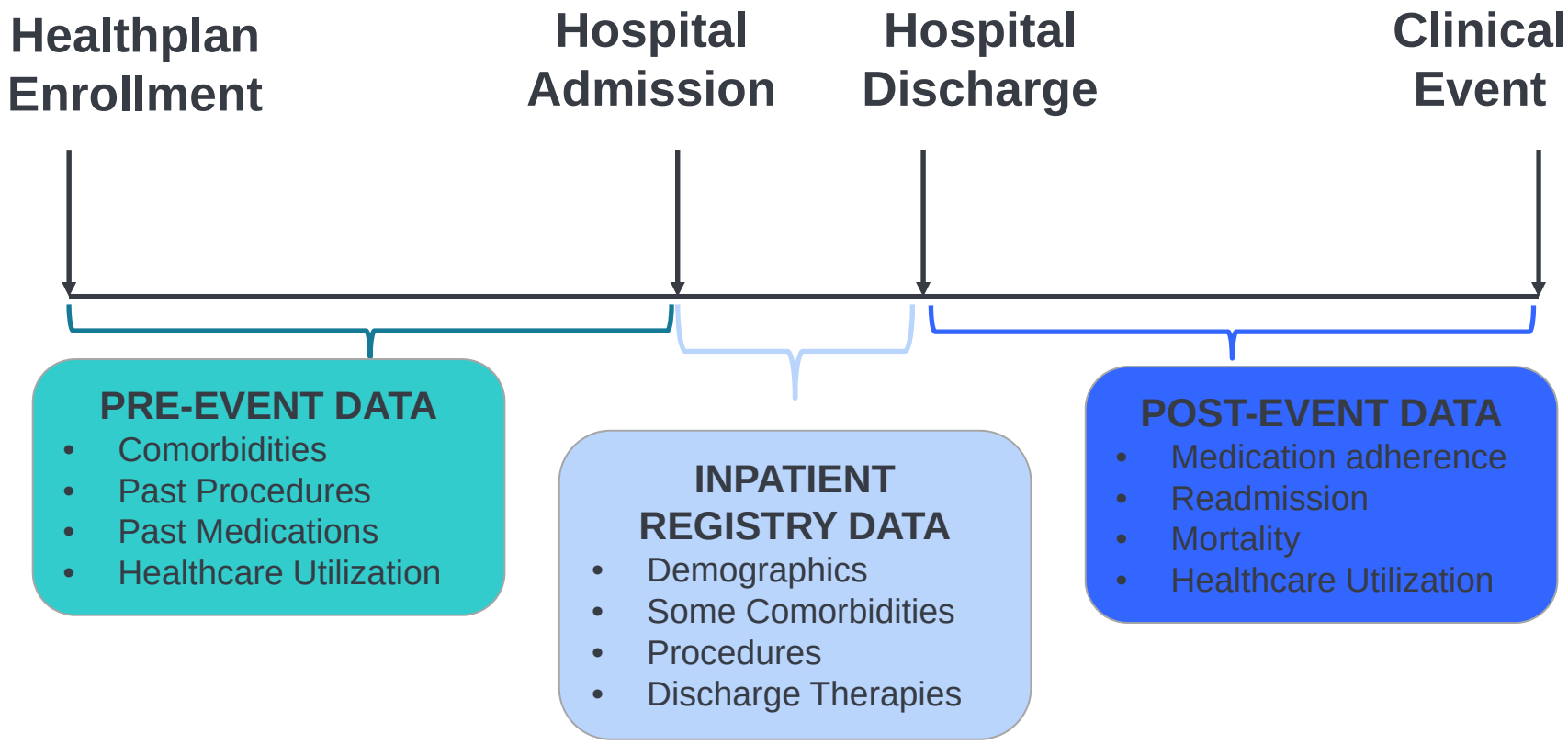
Pt #	LDL-C
1	108
2	134
3	122
4	95

Why link data sources for research?

- ***All data are fragmented***
- Linkages can support:
  - Better risk adjustment
  - Ascertainment of long-term clinical outcomes
  - Examination of non-clinical (e.g. patient-reported) outcomes
  - Ability to examine healthcare utilization/cost
- More information at potentially lower cost, inconvenience, & risk to the patient



Linkage Benefits: An Illustration



Why do we link?

Dataset 1

+

Dataset 2

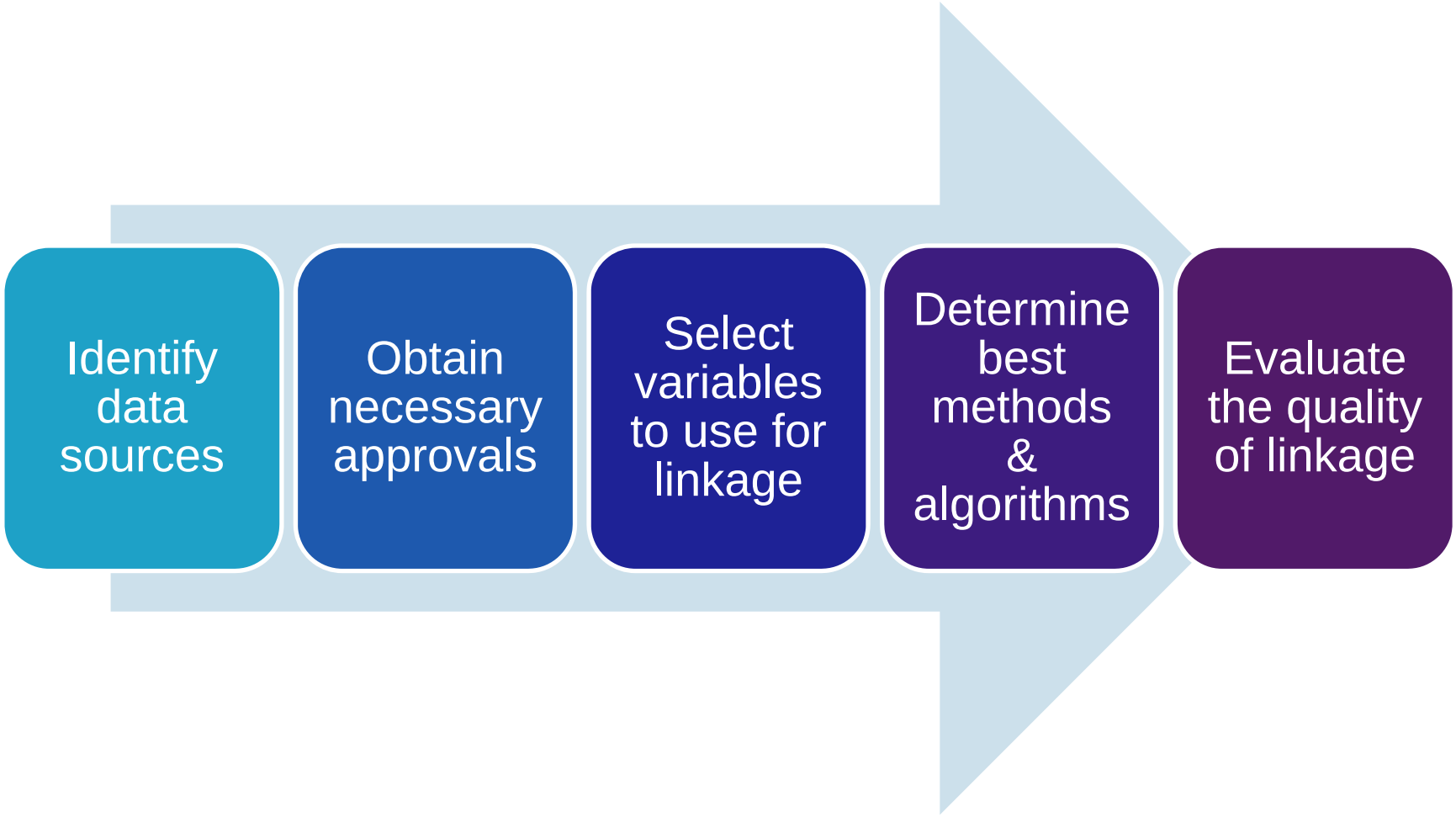
=

Craig's Crazy Carrot Cake Cheesecake™

Outline

- What is data linkage and why do we use it?
- Existing and evolving data linkage methods
- Challenges of linkage
- Real world examples of linking datasets

Data Linkage: A Basic Framework



Five Considerations for Data Linkage

- **Why?**
  - Q: *What is the added value of the information in the linked data source?*
- **Who?**
  - Q: *What is the overlap in datasets? Which patients are not included?*
- **Where?**
  - Q: W*here will the linkage be performed & who will have access to the datasets?*
- **How?**
  - Q*: What variables/algorithms are available to support high-fidelity linkage?*
- **What?**
  - Q: *Are there known/published match rates available for assessment of linkage "success"?*

Two strategies

- **Deterministic Matching**
  - 1 to 1
  - Unique identifier used to link data
  - Not always reliable; in some cases no single field can provide a completely reliable match between records
- **Probabilistic Matching**
  - “Fuzzy” matching
  - Multiple field values compared between records
  - Each field assigned a weight that indicates how closely the records match
  - Sum of individual weights = likelihood of a match

A new framework for collective record linkage in clustered data

- Records are organized into clusters (e.g. patients within hospitals)
- Aim: to make optimal match decisions using all of the data from each cluster pair
- Derive the “optimal” decision rule for clustered data in a clinical trial-CMS linked dataset
- Derive conditions to allow relaxing “conditional independence” assumption (usually violated by conventional linkage methods)
- Reduces to Fellegi-Sunter rule if 1 record/cluster

	Sensitivity (higher is better)	PPV (higher is better)
Fellegi-Sunter	92%	7%
Generalized Fellegi-Sunter	92%	99%

Outline

- What is data linkage and why do we use it?
- Existing and evolving data linkage methods
- Challenges of linkage
- Real world examples of linking datasets

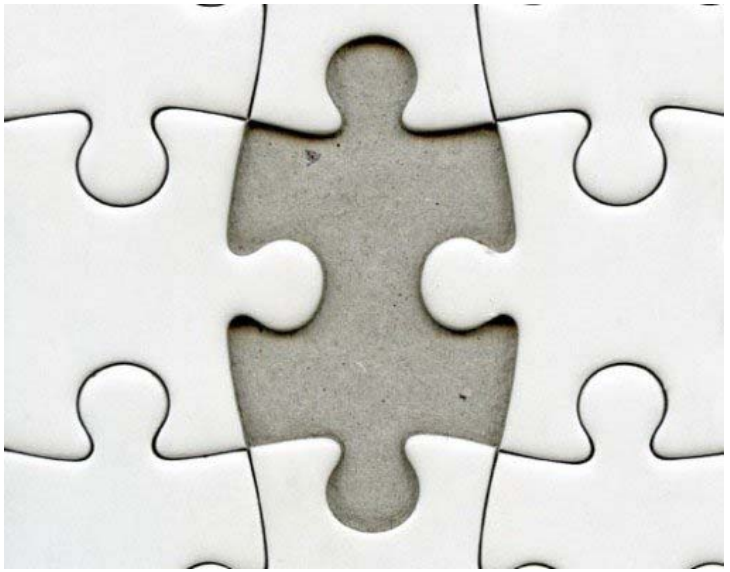
1. Not all identifiers are created equal

- ***Unique vs. non-unique***
  - Unique identifiers (SSN, HIC, MRN) may promote better linkages, but governance may also be different
- ***Linkage can generally be improved by also using non-unique identifiers***
  - Recommended: Sex, DOB
  - Not recommended: Race & ethnicity (inconsistently recorded or misreported)
- ***Identifiers not always standardized***

Dataset	SSN	Name	DOB
A	23685889	Smith, Bob	04/23/1956
B	23685889	Smith, Robert	1956/04/23

2. Informative Missingness

- ***Few data sources are 100% complete and consistent***
  - Incompleteness may vary by important patient characteristics (e.g, age, geographic region)
  - Missingness may be informative (e.g. female name change, lack of national health identifier in the United States)
- ***Comparing characteristics of linked/non-linked patients***
  - Key for evaluating potential bias and generalizability of results



3. Governance

- ***Ethical Considerations***
  - Different organizations (ie., healthcare systems, health insurance plans, state/federal databases) differ with respect to ethical requirements
  - Lack of clarity on who is the responsible party if there is a data breach
- ***Developing partnerships***
  - Healthcare data may represent a competitive advantage
  - All linkage carries some risk of re-identification *(sin of omission?)*
  - Requires incentivizing participation to assume additional risk of linkage

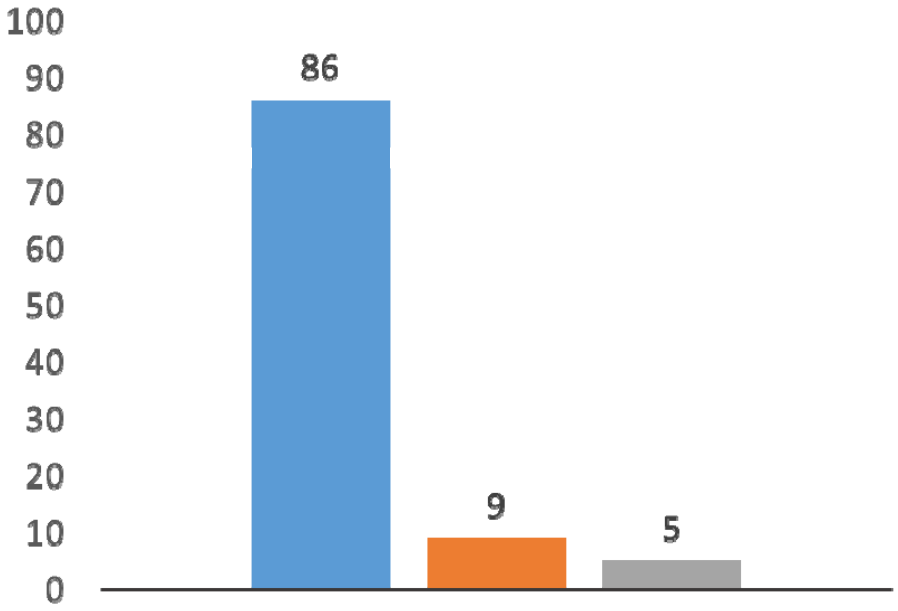
What is the patient perspective on data sharing for record linkage?

- **Survey Methods**
  - Recruitment from PatientsLikeMe, an online patient community representing over 2,500 health conditions
  - Survey developed using subject matter expertise and patient feedback from a concept elicitation phase (N=57 patients)
- **Population**
  - 3516 respondents
  - 73.8% women
  - 86.4% Caucasian
  - 14.5% 65 or older
  - 44.9% college or post-graduate education

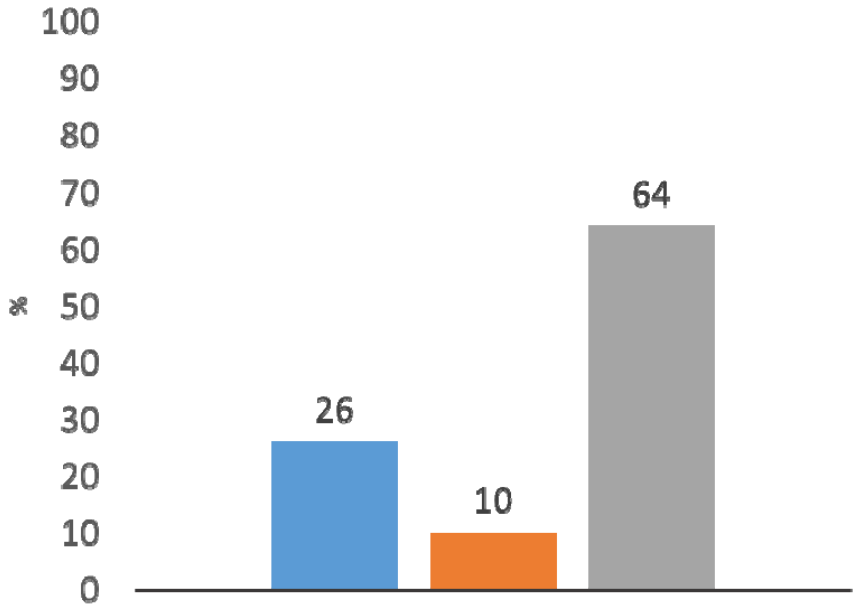
pcori®

patientslikeme™

Overall Data Sharing Comfort (n=3516)

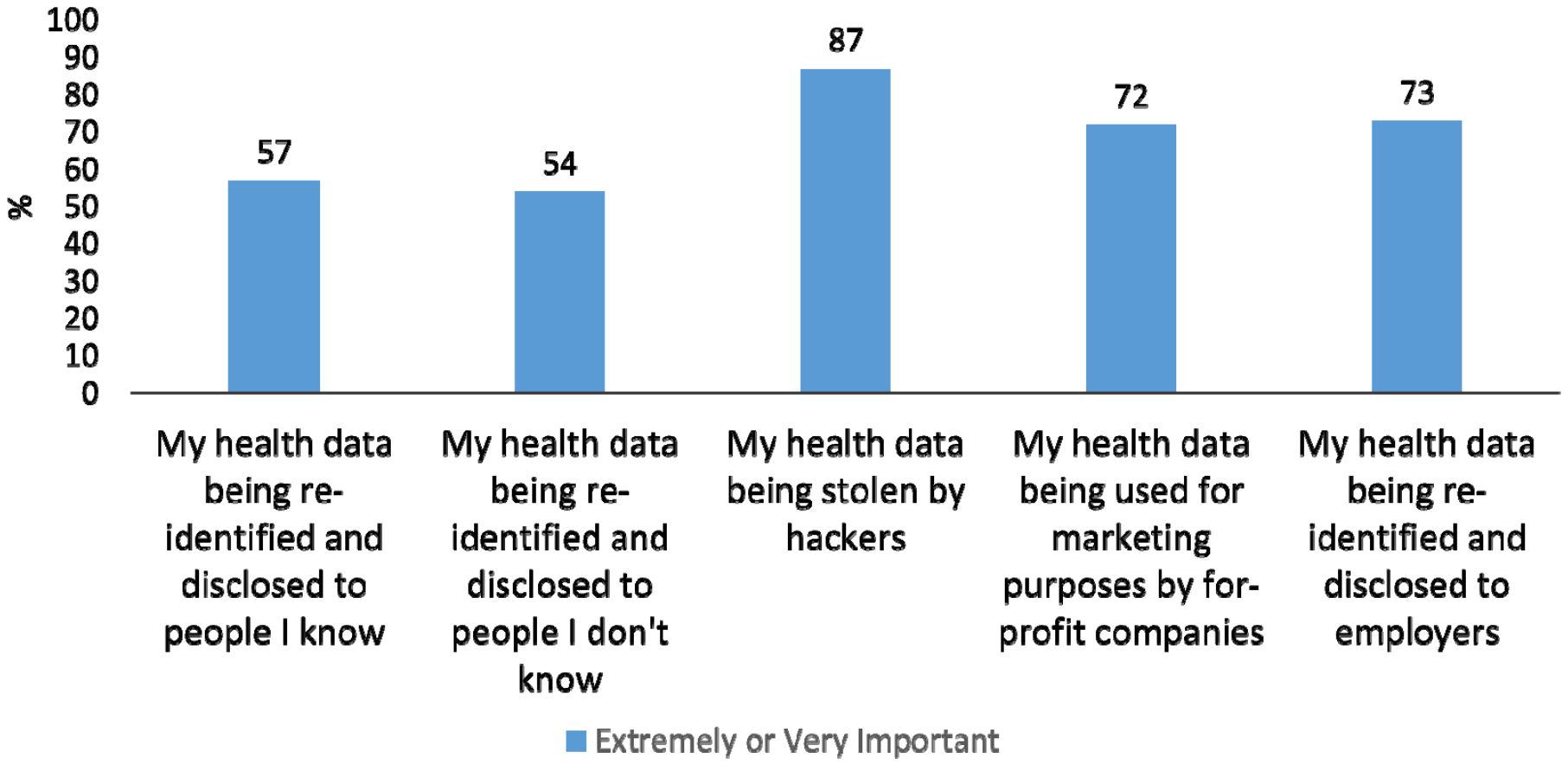


Comfortable with my health data being confidentially shared with researchers, as long as personal information like name and social security number is not available to researchers

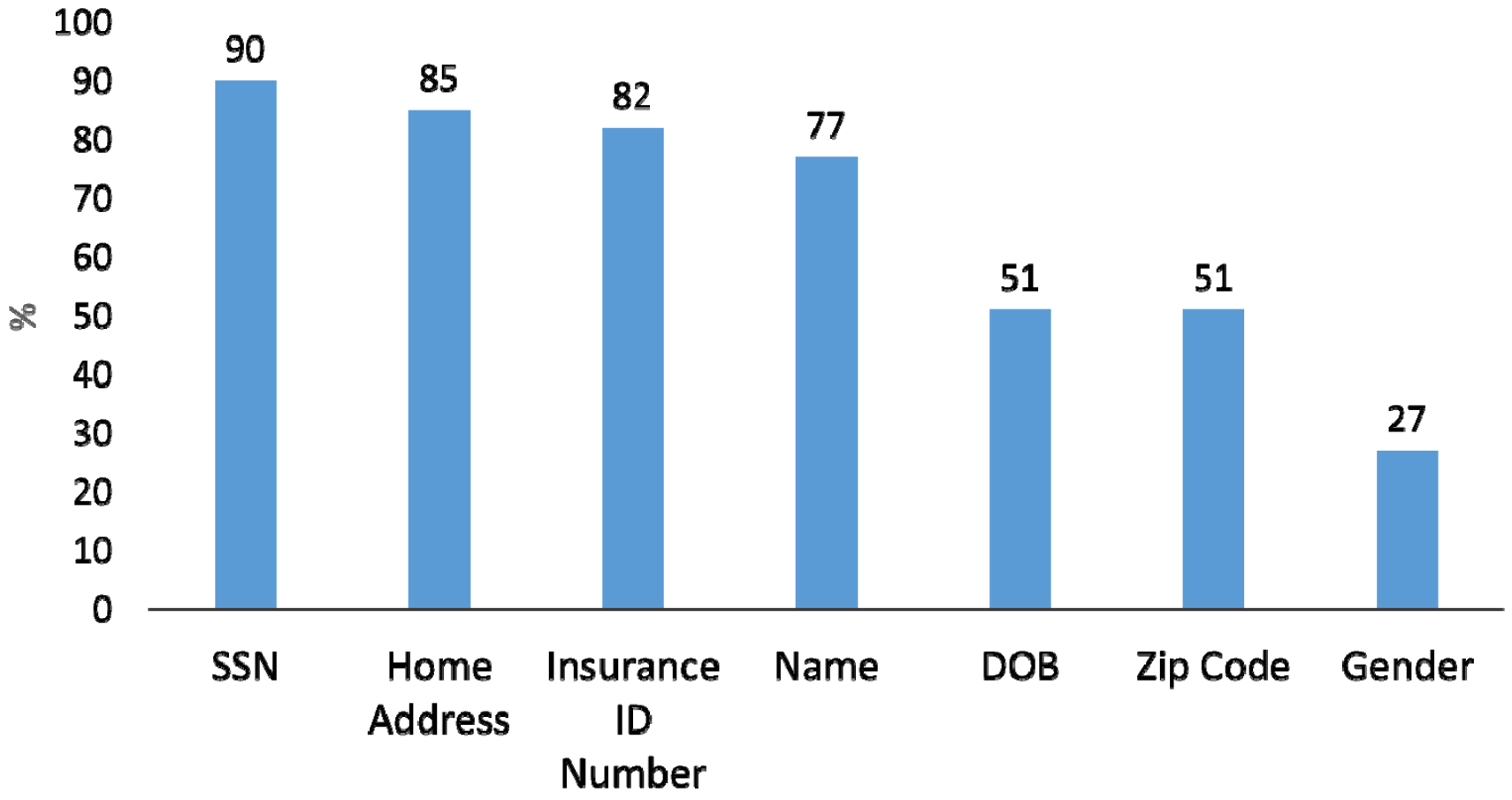


Comfortable with my electronic health data being confidentially shared with health care researchers, EVEN IF personal information like my name and social security number is available

Importance of Data Linkage Risks (n=3516)



% reporting they would be “extremely” or “much more comfortable” with removal of the following identifiers (n=3516):

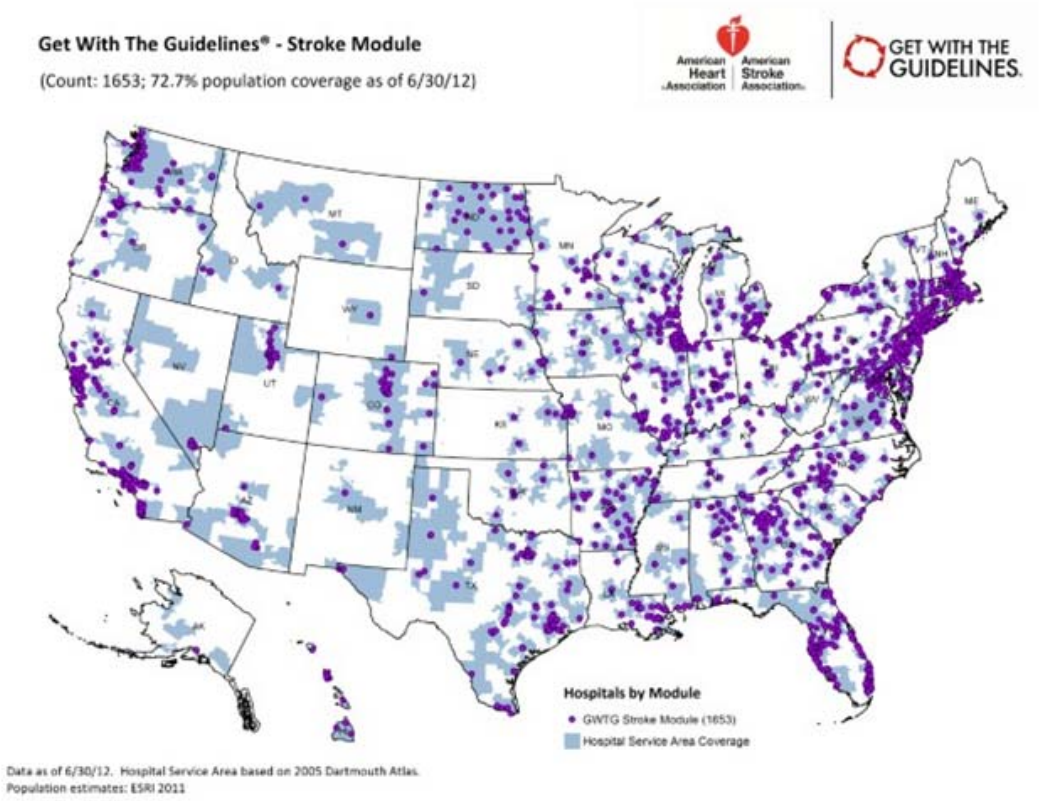


Outline

- What is data linkage and why do we use it?
- Existing and evolving data linkage methods
- Challenges of linkage
- Real world examples of linking datasets

PROSPER CER Analyses: Methods

- **Study population**
  - Get With the Guidelines-Stroke: National Quality Improvement Initiative
  - Acute ischemic stroke patients survived to discharge between 2009 and 2011
  - Linked to CMS for longitudinal outcomes up to 2 years after discharge



Leveraging Unique Data Sources

- Inpatient hospital data
- Detailed covariate information
- Medication use at discharge

- Post-discharge vital status
- Readmission
- SNF/Rehab information

Statin Use After Ischemic Stroke

	Statin (n=54,991)	No Statin (n=22,477)	Unadjusted HR (95% CI)	Adjusted HR (95% CI)
MACEs, %	**48.9%**	**57.9%**	**0.78 (0.76, 0.80)**	**0.91 (0.87, 0.94)**
P <.001				

	Statin (n=54,991)	No Statin (n=22,477)	Unadjusted Difference (99% CI)	Weighted Difference (99% CI)
Home time, days mean (SD)	**544 (255)**	**475 (285)**	**71 (65, 77)**	**28 (21, 34)**
*Weighted by proportion of follow-up; †Differences in days				P <.001

O'Brien EC et al *Circulation* 2015

Warfarin Use After Ischemic Stroke & Atrial Fibrillation

	Warfarin (N=11,039)	No Anticoagulant (N=1,513)	Unadjusted HR (95% CI)	Adjusted HR (95% CI)
MACEs	54.7 %	66.8%	0.73 (0.67-0.80)	**0.87 (0.78-0.98)**
				P =0.003

	Warfarin (N=11,039)	No Anticoagulant (N=1,513)	Unadjusted Difference (99% CI)	Weighted Difference (99% CI)
Home time, days mean (SD)	468 (255)	389 (267)	86 (68-104)†	**48 (27-68)†**
	*Weighted by proportion of follow-up; †Differences in days			P <.001

Xian Y et al BMJ 2015

Conclusions

- Data linkage offers significant promise but important challenges remain
- Requires a multidimensional approach
  - Development of new probabilistic methods
  - Patient advocacy
  - Effectively communicating value to stakeholders

Thank you!

Emily.Obrien@duke.edu

PCORI Methods Project: Accuracy of linkage algorithms using only indirect identifiers

- **Data Source**: Society of Thoracic Surgeons Adult Cardiac Database (N = 467,755) linked to CMS claims
- **Objective**: Estimate accuracy of various linkage rules treating linkage based on SSN + indirect identifiers as gold standard

Model-Based Accuracy Estimates (Preliminary Results)		
Linkage Rule	Sensitivity	1-PPV
SSN + admission date	93.5%	<0.00001%
DOB + sex + admission date	92.6%	0.004%
DOB + sex + discharge date	96.3%	0.004%
2 of 3 DOB components + sex + discharge date	97.8%	0.25%