Data Integration Innovations to Enhance Analytic Capacity and Inform Healthcare Policy and Practice

Steven B. Cohen

Presentation

- Data Integration efforts: A model for enhanced analytic capacity and data quality

- Highlight potential gains in efficiency, accuracy, capacity and quality

- Attributes that impact on survey designs

- Applications to Project Data Sphere

- Model for consideration

The Project Data Sphere data integration effort is being funded by a grant from the Robert Wood Johnson Foundation

Data Integration Model

- Data integration is a process in which related and supplemental data from multiple sources are connected into a unified structure.

- The resultant integrated data resource serves as a platform to enhance analytic efforts.

- The data integration model facilitates greater analytic utility for each of the component data sets as a consequence of their "connectivity".

- Data integration is often implemented in a data warehouse or data enclave setting to ensure the extraction, linkage and structure of the combine data resources are presented in a unified manner.

Parameters

- Accuracy

- Relevance

- Timeliness

- Accessibility

- Clarity

- Cost-efficient

Integrated Survey Design Features

- Direct linkage between sample members in core survey with larger host survey; administrative records; or follow-up surveys

- Use of secondary data (e.g. aggregate data at the county/state level) as core component of survey

- Prior survey record of call data informs data collection strategies

- Informs sample design, nonresponse and poststratification adjustments, imputation and data supplement for item nonresponse

- Facilitates reductions in measurement error

- Need for greater attention to ensuring confidentiality: limitations in public use data

Administrative Records and National Surveys Serving as Sampling Frames

- The Medicare Current Beneficiary Survey (MCBS) conducted by the Centers for Medicare & Medicaid Services (CMS) is a continuous, multipurpose survey of a nationally representative sample of the Medicare population: *the sample is selected from Medicare enrollment files*

- The National Health Interview Survey (NHIS) has a central role in the ongoing integration of household surveys in DHHS.
  - The National Survey of Family Growth has used the NHIS as a sampling frame and
  - the Medical Expenditure Panel Survey currently uses the NHIS as a sampling frame.
  - Other linkage includes linking NHIS data to Medicare and Medicaid administrative data from CMS and death certificates in the National Death Index (NDI).

Medical Expenditure Panel Survey (MEPS)

Annual Survey of 14,000 households:

provides national and state estimates (most populous) of health care use, expenditures, insurance coverage, sources of payment, access to care and health care quality

Permits studies of:

- Distribution of expenditures and sources of payment
- Role of demographics, family structure, insurance
- Expenditures for specific conditions
- Trends over time

Sponsored by the Agency for Healthcare Research and Quality

MEPS Medical Provider Component

Targeted Sample

- All associated hospitals and sample of associated physicians

- Sample of associated office-based physicians

- All associated home health agencies

- All associated pharmacies

Data Collected

- Dates of visit

- Diagnosis and procedure codes

- Charges (except prescriptions) and payments

- Problem: Recent changes in the provision and delivery of health care have led to concerns about how these changes interact with provider organizational characteristics to affect access, use, and expenditures for care.

- Analytic Need: Linked information is needed on both the characteristics of providers and the individuals receiving care.

- *Prior to the MEPS MOS, there was no nationally representative data source that provides this essential information.*

Grant Provided by Robert Wood Johnson Foundation to Support MEPS-MOS

MOS Objectives

- Previously, the MEPS did not acquire essential data on provider, practice and organizational characteristics, policies and treatment protocols, penetration of ACOs, medical homes, and health information technology (HIT).

- The MOS fills a critical gap in content to permit analyses assess the impact of the following factors on individuals' access, use and affordability of health care services:

  o Organizational characteristics (e.g., size, specialties covered, practice rules and procedures, patient mix and scope of care provided, membership in an ACO, certification as a primary care medical home)

  o Use of health information technology

  o Policies and practices with potential Affordable Care Act impact

  o Financial arrangements (e.g., reimbursement methods, number and types of insurance contracts, compensation arrangements within the practice)

Analytic Capacity

It supports analyses designed to answer questions such as:

- How does the organization of physician practices affect access to care for individuals with different health characteristics and types of insurance?

- How does the provider organization affect use of different types of services?

- What is the relationship between physician practice organization and overall medical expenditures for care?

- What is the relationship between physician practice organization and individuals' out-of-pocket costs for care?

- How is provider organization related to the health status of the individuals receiving care?

MEPS Integrated Design

Household Component (HC)

Medical Provider Component (MPC)

Medical Organization Survey (MOS)-support provided by *the Robert Wood Johnson Foundation*

Insurance Component (IC)

- Longitudinal design

- Linkage to CMS claims data

- Linkage to National Health Interview Survey

- Linkage to National Death Index

- Data Supplementation at the state and county levels.

Data Integration Innovations

Challenge: Enhancements to the analytic capacity and utility of cancer clinical trial data hosted by *Project Data Sphere, LLC ("PDS") :* Limited demographic information currently available on PDS patients to ensure confidentiality

Approach/Innovation: Data integration efforts employed to join PDS patient-level data with nationally representative health and healthcare related data.

Collaborations/Partnerships: *Project Data Sphere, LLC ("PDS")-*

RTI International-The Robert Wood Johnson Foundation

Impact/Next steps: Permit examination of the level of variation in patient outcomes attributable to differentials in access to care, health care service utilization, socioeconomic characteristics, and to health behaviors and preferences

- Launched **April 8, 2014**
- **Phase III cancer trials** (Industry and National Cancer Institute)
  - **Comparator arm data** at launch
  - **Experimental arm data** now available
- **De-identified patient data, data dictionary, protocol, & CRFs**
- Free, powerful **analytic tools**
- **Easy-to-use,** with favorable IP
- An independent initiative of the CEO Roundtable on Cancer's *Life Sciences Consortium*
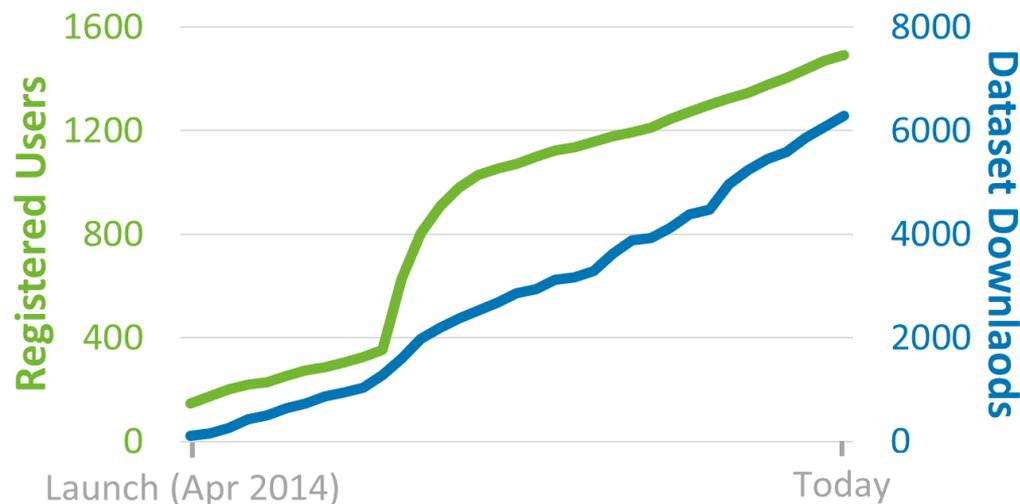

PDS-inspired "Sounding Board" published in *New England Journal of Medicine,* March 23
Join the conversation: bit.ly/NEJMDataSharing

- **~63,000 pt. lives** of data from 28 providers
- **>1,600 researchers** accessed data ~7,000 times
- **Triple the usage** of other major, clinical trial data-sharing initiatives combined[1]

Registered Users / Dataset Downlaods

Launch (Apr 2014) — Today

- **17 peer-accepted publications** to date cite PDS data, including:
  - Prognostic model to predict Prostate Cancer overall survival (OS)
  - Tumor growth model with statistically valid intermediate endpoint for OS
  - Efficacy differentiation for multiple FDA-approved Prostate Cancer therapies
  - Prediction of Pancreatic Cancer OS in patients treated with gemcitabine

THE NEW ENGLAND JOURNAL *of* MEDICINE

THE LANCET Oncology

SCIENTIFIC DATA
a natureresearch journal

EUROPEAN UROLOGY

The Oncologist

[1]Navar A, et al. Use of Open Access Platforms for Clinical Trial Data. JAMA. 2016;315(12):1283-1284.

Project Data Sphere Applications

- For most patient-level records on the PDS platform, demographic measures available for statistical linkage are generally limited to age, race, and sex to reduce the possibility of re-identification.

- A data integration effort limited to these three demographic measures would produce a multitude of many-to- many exact linkages.

- To ameliorate this problem, our approach to data integration uses an additional measure that further distinguishes patients by their health-related quality of life assessments.

- This measure is the EQ-5D™ index score, derived from the EuroQOL five dimensions questionnaire, one of the most commonly used measures of health-related quality of life.

Project Data Sphere Applications

- The EQ-5D™ descriptive system consists of the following five health-related components: Mobility, Self-care, Usual activities, Pain/discomfort, and Anxiety/depression.

- Each dimension has three levels, reflecting no health problems, moderate health problems, and extreme health problems.

- Consequently, there are 3^5=243 health states defined by the instrument, with the associated 5-digit response profiles ranging from 11111 for perfect health to 33333 for the worst possible state.

- To calculate the EQ-5D™ index score based on the U.S. population-based preference weights, a scoring algorithm has been created and operationalized.

Project Data Sphere Applications

- The EQ-5D has also been administered in the past in the MEPS, which also includes administration of the 12-Item Short Form Health Survey (SF-12).

- The SF-12 is a general health status instrument with 12 questions producing two summary scores, the Physical Component Summary (PCS-12) and the Mental Component Summary (MCS-12).

- EQ-5D index scores can be derived from MEPS using an algorithm developed by Sullivan and Ghushchyan (2006) that only requires the availability of the MCS-12 and PCS-12 scores.

This method uses the following prediction equation: EQ-5D = 0.057867 + 0.010367·(PCS-12) + 0.00822·(MCS-12) - 0.000034·(PCS- 12·MCS-12) - 0.01067.

Both the direct values of the index scores (when available) and the predicted values of the EQ-5D index scores are used as an additional discriminatory variable in the statistical linkage.

Project Data Sphere Data Enhancements

PDS Data Used for Linkage

- PDS data file *LungNo_MerckKG_2007_145* includes 507 lung cancer patients, representing the intent to treat population

MEPS Data Used for Linkage

- MEPS lung cancer survivors were identified among all MEPS cases from the 2000-2013 Household Component (HC) Survey Full Year
- MEPS cases with ICD9CODX = 162 were identified as lung cancer survivors.
- There are 653 MEPS lung cancer survivors

Criteria Used for Linkage

Summary of Criteria Used for Linkage	EQ-5D Value Used for Linkage	
	MEPS	PDS
2000 - 2003 MEPS		
1st Step: Single year age, sex, race, direct EQ-5D	**Direct**	**Direct**
2nd Step: Categorized age, sex, race, direct EQ-5D	**Direct**	**Direct**
3rd Step: Collapsed categorized age (75+), sex, race, decile categories of EQ-5D	**Decile Cat (Dolan)**	**Decile Cat (Dolan)**
2004 - 2013 MEPS		
1st Step: Single year age, sex, race, decile categories of EQ-5D	**Decile Cat (SG)**	**Decile Cat (Dolan)**
2nd Step: Collapsed categorized age (75+), sex, race, decile categories of EQ-5D	**Decile Cat (SG)**	**Decile Cat (Dolan)**

Age, sex, race, and measures of the EQ-5D were used to link to PDS cases.

Age categories:18-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75-84, 85+.

Many-to-many linkages were allowed, with the variable LINKMETHOD indicating the method under which each linkage was attained.

Linkage Results

- 401 PDS cases (507 in PDS) achieved a linkage to MEPS lung cancer survivors.

- 283 MEPS lung cancer survivors (653 in MEPS) achieved a linkage to PDS lung cancer cases.

MEPS Survey Weights and Sample Design Variables

It is advised to produce average nationally representative estimates of lung cancer survivors using the MEPS lung cancer survivors included in the linked dataset.

To do so, the file should either be de-duplicated by MEPSID to account for many-to-many linkages or the weights averaged among the multiple MEPS linkages

Assessment of Factors that Distinguished the Characteristics of Lung Cancer Cases in the PDS clinical trial

Socio-demographic factors:

- Age, race/ethnicity, sex, marital status, employment status, education level, income level, year in MEPS

Access related factors:

- Insurance coverage, ability to obtain necessary medical care

Health related:

- EQ5D, health status, work limitations, smoker status

Healthcare related:

- Office based physician visits, in-patient stays, ER visits, Rx purchases, total expenditures

Factors that Distinguished the Characteristics of Lung Cancer Cases in the PDS clinical trial

Contrast	DF	Wald F	P-value
Overall Model	9	6.95	<0.0001
Model minus intercept	8	7.10	<0.0001
Marital Status	1	6.16	0.0134
Sex	1	11.04	0.0010
MEPS Year	1	6.47	0.0113
EQ5DDEC	1	14.81	0.0001
Race/ethnicity	2	25.94	<0.0001
Difficulty in access to necessary medical care	1	3.17	0.0758
Smoker Status	1	4.46	0.0352

Exploratory Assessments of Factors that Suggest Association with Survivorship in the Comparator Arm

PDS Measures:

- Age, race/ethnicity, sex
- ECOG Performance: scale used to assess how a patient's disease is progressing, assess how the disease affects the daily living abilities of the patient, and determine appropriate treatment and prognosis
- Microscopic verification, Surgery / non-surgery, Smoking history
- Response to chemo-radiotherapy
- Type of chemo-radiotherapy
- N Stage

Supplemented with MEPS Measures

Exploratory Assessments of Factors that Suggest Association with Lung Cancer Survivorship in the Comparator Arm

Contrast		DF	Wald F	P-value	Pseudo R^2
Overall Model		11	7.11	0.0000	0.1104
Model minus intercept		10	3.49	0.0002	
Response to chemo-radiotherapy		1	2.82	0.0940	
Type of chemo-radiotherapy		1	2.49	0.1155	
N Stage		2	4.24	0.0151	
MEPS Measures					
Income		1	2.33	0.1273	
Medicaid coverage		1	7.43	0.0067	
Private HMO coverage		1	3.34	0.0684	
Current smoker		1	3.81	0.0515	
Believes health insurance is not needed		1	3.64	0.0573	
Had labtests		1	6.05	0.0143	

Project Data Sphere Applications

The PDS-MEPS Data Integration will permit studies that examine:

Are the demographic characteristics of those cancer patients enrolled in specific phase III clinical trials comparable to cancer patients with the same disease in the general population?

How are variations in cancer patients' access to health care and income impacting patient outcomes in specific phase III clinical trials?

What variations in patient outcomes are associated with specific demographic, socioeconomic, and health-related factors?

Summary

- Data Integration efforts: A model for enhanced analytic capacity and data quality

- Highlight potential gains in efficiency, accuracy, capacity and quality

- Attributes that impact on survey designs

- Applications to Project Data Sphere