Population Health Assessment in Cancer Center Catchment Areas – All Grantee Meeting

Data Integration and Preliminary Analysis Plan

Presented by:
Tonja Kyle, MS, ICF
Ronaldo Iachan, PhD, ICF
Lew Berman, PhD, ICF

2/24/2017

ICF

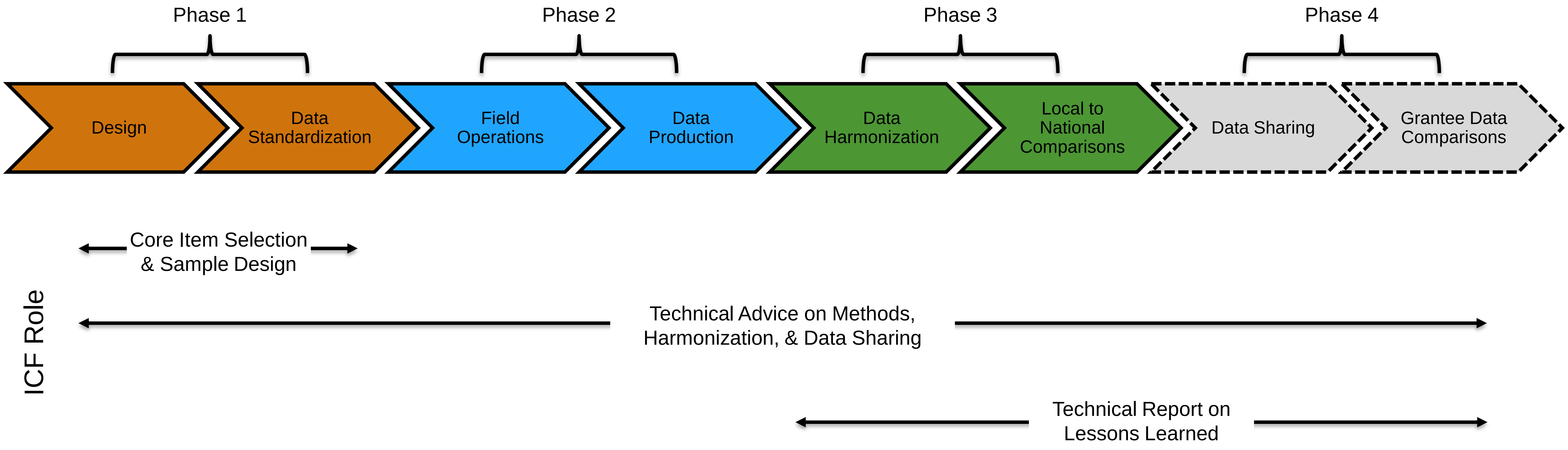
Session Agenda

- Project phases and ICF role – Lew Berman
- Data production practices – Tonja Kyle
- Data harmonization – Ronaldo Iachan
- Open data and data sharing – Lew Berman
- Selected relevant NIH funding opportunities - Rick Moser

ICF Team:

Audie Atienza, PhD	Ronaldo Iachan, PhD	Matt Thomas, PhD
Lew Berman, PhD	Tonja Kyle, MS	Bob Tortora, PhD
John Boyle, PhD	Kelly Martin, MPH	
Bridget Beavers, MS	Deirdre Middleton, MPH	

Project Phases & ICF Role



Data Production – Why it's important…

- **How often have you….**
  - Tried to use a dataset for analysis and couldn't find information you needed?
  - Had to recode/reorganize the dataset significantly to prepare it for analysis?
  - Tried to replicate something such as a computed variable, but weren't completely sure you had all the information required?
  - Found poorly described variables in your dataset?
- **These types of experiences are the result of missteps in data production**

Data Production Defined

- **Data production is the process of curating, documenting and preparing data for use by other researchers and/or audiences**

- **Common elements associated with data production can include:**
  - –Selecting data file formats
  - –Determining and applying variable types and naming conventions
  - –Deciding on variable values/formats
  - –Developing, implementing and documenting data cleaning/editing specifications
  - –Documenting data limitations and analytical considerations
  - –Developing codebooks and documenting study methods

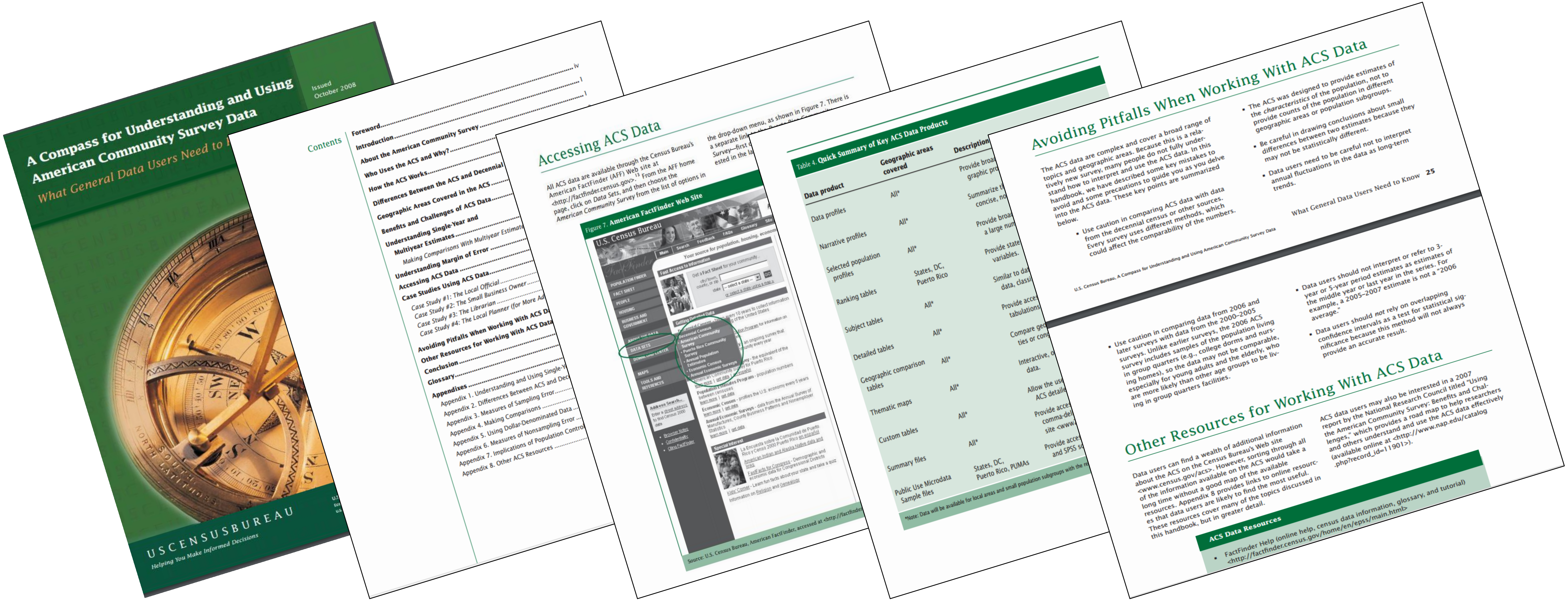
- **Careful and comprehensive data production makes the data more useful to you and to others**

The Key to Producing User Friendly Data

- **Quality Data Documentation**
  - Study Plan and Operation Guide (including information on study mode, sample design, response rates, eligibility rates, cooperation rates, refusal rates)
  - Codebook
  - Analytic Guidelines

- **How can ICF help?**
  - Review data management plans
  - Review data structures and survey logic
  - Provide recommendations for data user documentation

Examples of Data Documentation – American Community Survey Data Handbook



https://www.census.gov/content/dam/Census/library/publications/2008/acs/ACSGeneralHandbook.pdf

Examples of Data Documentation – Calculated Variables

Section 15: Breast and Cervical Cancer Screening

_MAM502Y *Calculated variable for women respondents aged 50+ who have had a mammogram in the past two years.* _MAM502y is derived from SEX, AGE, HADMAM, and HOWLONG.

1	Yes	Female respondents aged 50 and older who have received a mammogram within the past two years. (SEX=2 and AGE >= 50 and HADMAM=1 and HOWLONG=1,2)
2	No	Female respondents aged 50 and older who have not received a mammogram within the past two years. (SEX=2 and AGE >= 50 and HADMAM=2 or HADMAM=1 and HOWLONG=3,4,5)
9	Don't know/ Not Sure/ Refused	Female respondents aged 50 and older with don't know, not sure, or refused responses for HADMAM or HOWLONG or female respondents with don't know, not sure, refused or missing responses for AGE, HADMAM or HOWLONG. (SEX=2 and HADMAM=7,9, missing or HOWLONG=7,9, missing or AGE=7,9,missing)
.	Missing or Age less than 50 or Male	Female respondents less than 50 years old, or male respondents. (SEX=1 or SEX=2 and AGE < 50)

SAS Code:

```
IF SEX=2 AND AGE GE 50 THEN DO;
IF HADMAM=1 THEN DO;
IF HOWLONG IN (1,2) THEN  MAM502Y=1;
ELSE IF HOWLONG IN (3,4,5) THEN  MAM502Y=2;
ELSE IF HOWLONG IN (7,9) THEN _MAM502Y=9;
END;
ELSE IF HADMAM=2 THEN _MAM502Y=2;
ELSE IF HADMAM IN (7,9,.) THEN _MAM502Y=9;
END;
ELSE IF SEX=2 AND AGE IN (.,7,9) THEN _MAM502Y=9;
ELSE  MAM502Y=.;
```

Considerations for Catchment Area to National Comparisons…

- Source calculation
- Exclusion criteria
- Other questions that are used in calculation of the analytical variable(s)
- Cleaning methods that impact the variable(s) of interest
- Imputation and the impact on the variable(s) of interest
- Linking accuracy and validation
- Contextual data from the outside data source needed to produce meaningful results

Harmonization: Steps Towards Comparing and Combining Grantees' Data

- Set the stage for harmonization: goals and limitations
- Harmonizing constructs
- NCI Grantees: clusters of similar designs which can be more easily compared
- Community HINTS provides example comparisons with local and national data
- How and when can ICF help?

Standardization and Harmonization

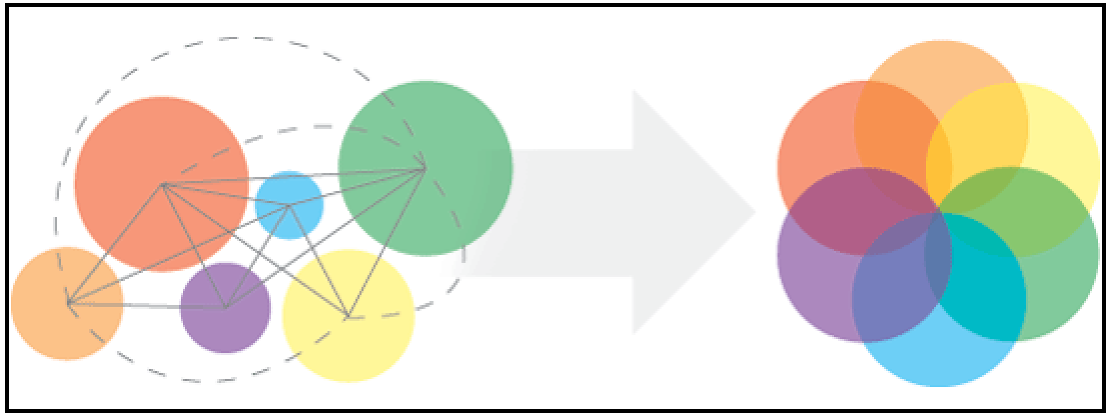
- Harmonization is a process by which variables are made comparable across locations, modes of data collection or survey years.
- Harmonization framework—ex-ante and ex-post:
  - Try to harmonize at the ex-ante (pre) phase as much as possible: STANDARDIZATION
- Improve comparability of different surveys and measures collected
- Applied to sampling, data collection, instruments and measures, etc.
  - If all else is harmonized, comparisons between the populations are more accurate

Standardized Constructs

Demographics	Behavioral constructs
Age	Health information seeking
Sex	Health information access
Place of birth	Breast cancer screening—ever had
Education	Colorectal cancer screening—ever had
Race/ethnicity	Cancer screening knowledge
Income	Tobacco use
Financial security	Cancer beliefs
Homeownership	Risk awareness
Health insurance	Health care access
Employment	Health care barriers

Comparisons for Each Grantee

- National HINTS
- State BRFSS
- Selected Metropolitan Area Risk Trends (SMART) BRFSS: County and MSA levels
- Other grantee(s)



Comparing and Combining Grantee Data

- Restrict comparisons and integration to clusters of sites with similar methods and populations
- Can also compare subgroups targeted by different grantees
  - For example, rural subpopulation, African Americans, homeless
- Summary table highlights challenges in combining data from diverse sites
- Primary clusters based on probability sampling—representing the general population of the catchment area-- versus non-probability sampling
- Clusters represent our initial attempt in our work with grantees to define such clusters

Classification by Methods and Populations, Cluster #1: Probability samples

Grantee	*Target Population*	*Other subpopulations of interest*	*Survey methodology (sampling and data collection)*
Dana-Farber	State (MA)	5 sub-populations	Probability panel sample plus community partners
University of Pennsylvania	General population in the catchment area	---	Dual frame RDD: subsample from the SEPA 2015 survey
Dartmouth	New Hampshire and Vermont	Urban, large rural, small rural, isolated rural	Stratified random sample (RDD)
MD Anderson	State (TX)	Rural and urban areas	ABS and panel samples stratified by urban/rural statewide
University of Pittsburgh	29 counties in Western PA (mostly rural)		Random digit dialing (RDD)
University of Kentucky	54 Kentucky counties designated as Appalachian	Metro, Slightly rural, Rural, Completely rural	Probabilistic, ABS/mail survey Convenience, community-based settings/in-person

Classification by Methods and Population, Cluster #2: Probability samples of patients

Grantee	*Target Population*	*Other subpopulations of interest*	*Survey methodology (sampling and data collection)*	*Grantee*
Indiana University	IU patient population – those seen at IU Health in past year	Rural/urban and Black/White comparisons	Mail survey using list probability sample (stratified)	Indiana University

Classification by Methods and Populations, Cluster #3: Non-probability samples

Grantee	*Target Population*	*Other subpopulations of interest*	*Survey methodology (sampling and data collection)*
Dartmouth	New Hampshire and Vermont	Urban, large rural, small rural, isolated rural	Convenience sample (MTurk)
UC San Francisco	N California (48 counties)	5 subgroups	NPS: community org based
Memorial Sloan Kettering	Two Bronx (NYC) districts	Blacks and Hispanics	NPS: Community partner recruitment
Hawaii	Hawaii and Guam ethnic subgroups	3 ethnic subgroups	Respondent driven sampling (RDS)
Ohio State University	State (OH)	6 subpopulations	Quota samples for six subpopulations; two done in person
Albert Einstein	General population in the Bronx (NYC)		Venue based sampling
Roswell Park	General population in catchment area	5 subgroups	Web panel, and org-based for subgroups
Temple University	Catchment area (counties in PA-NJ)	Underserved subpopulations	Two NPS samples: a) Block subsample, and b) Org-based

Community HINTS (CHINTS): Examples of Comparisons with Local, State and National Data

- Pilot internet panel surveys in selected communities
- Use HINTS questions plus BRFSS questions for comparisons
- Communities included Cleveland, New York City and Seattle (plus Los Angeles County--not presented here)
- The CHINTS studies highlight the importance of specifying a reference population, and in particular, the geographic scope (e.g. one or more counties, or the entire state)—for weighting and for comparisons
  - If all else is harmonized, comparisons between the populations are more accurate
- Also provide best practices in comparing and combining non-probability sample data

CHINTS: Geographic Scope and Data Sources

		Cleveland	Seattle	New York City
	Community HINTS (CHINTS)	MSA	County	Five boroughs
Post-Stratification Adjustments	American Community Survey (ACS) 2013	MSA	County	Five boroughs
Post-Stratification Adjustments	The Behavioral Risk Factor Surveillance System (BRFSS) 2012	MSA	County	MSA

CHINTS Post Stratification (Weighting)

Raking to 2013 ACS one-year estimates

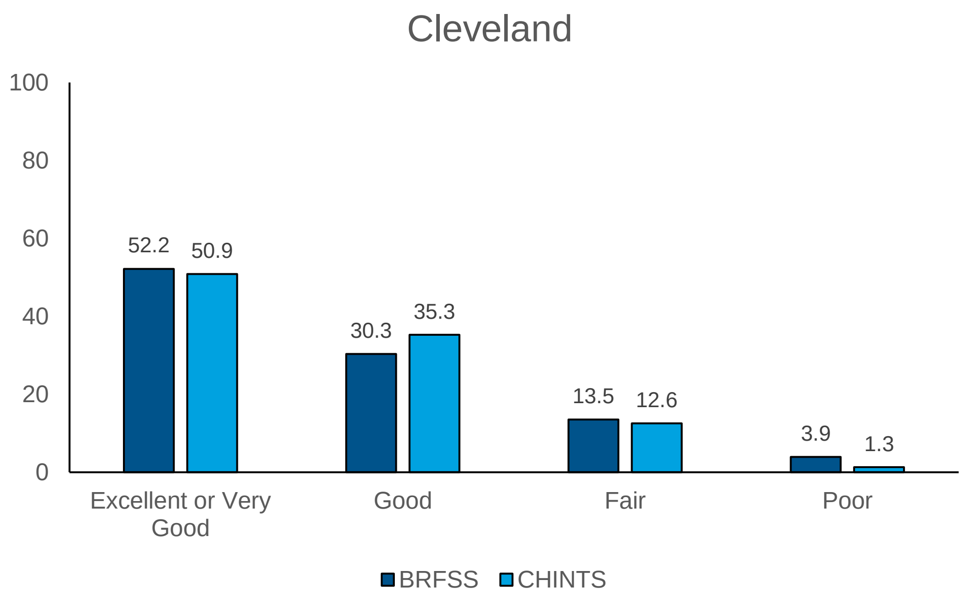
- Gender
  - Male and female
- Age
  - 18 – 34, 35 – 54, and 55+ years of age
- Race and Hispanic Ethnicity
  - Non-Hispanic (NH) white, NH black, Hispanic, and other
- Education
  - High school (HS) or less, more than HS
- Marital Status
  - Married and not married

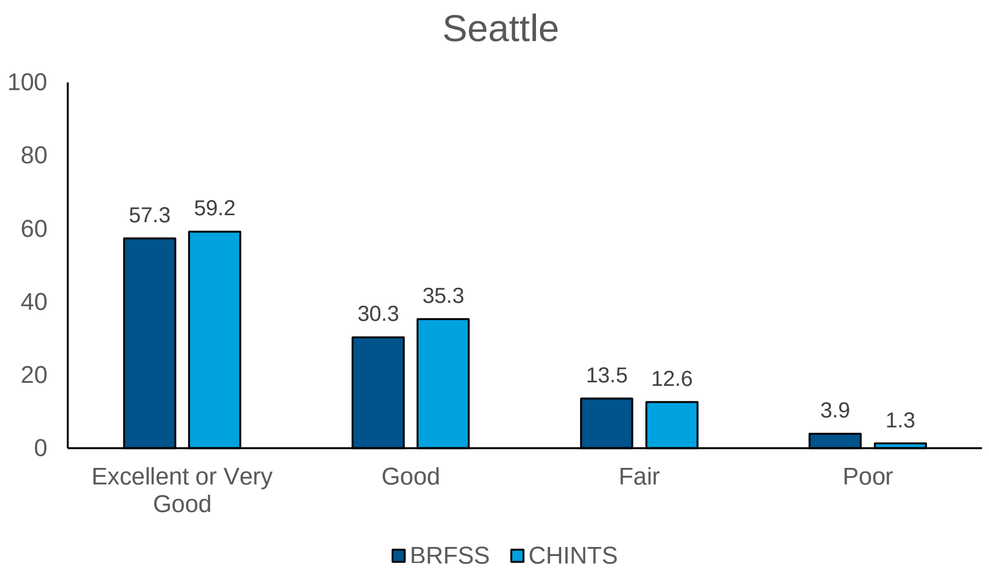
CHINTS Comparison Across Sites and with the SMART BRFSS

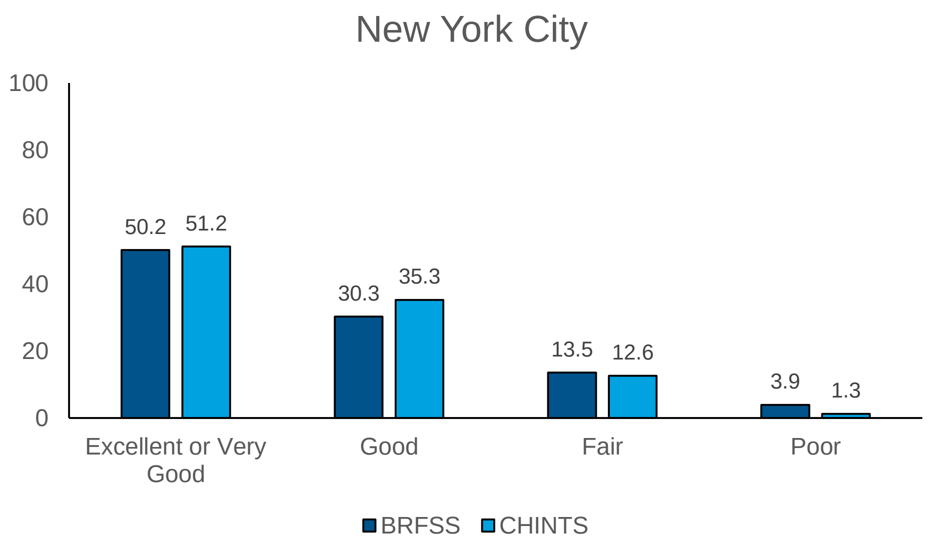
Height (inches)	Mean (SE)	
	BRFSS	CHINTS
Cleveland	67.1 (0.1)	67.2 (0.2)
Seattle	67.2 (0.1)	67.3 (0.3)
New York City	66.4 (0.1)	66.2 (0.3)

Weight (pounds)	Mean (SE)	
	BRFSS	CHINTS
Cleveland	178.7 (1.6)	191.8 (3.1)
Seattle	172.9 (1.1)	175.5 (3.1)
New York City	167.5 (0.8)	172.8 (2.6)

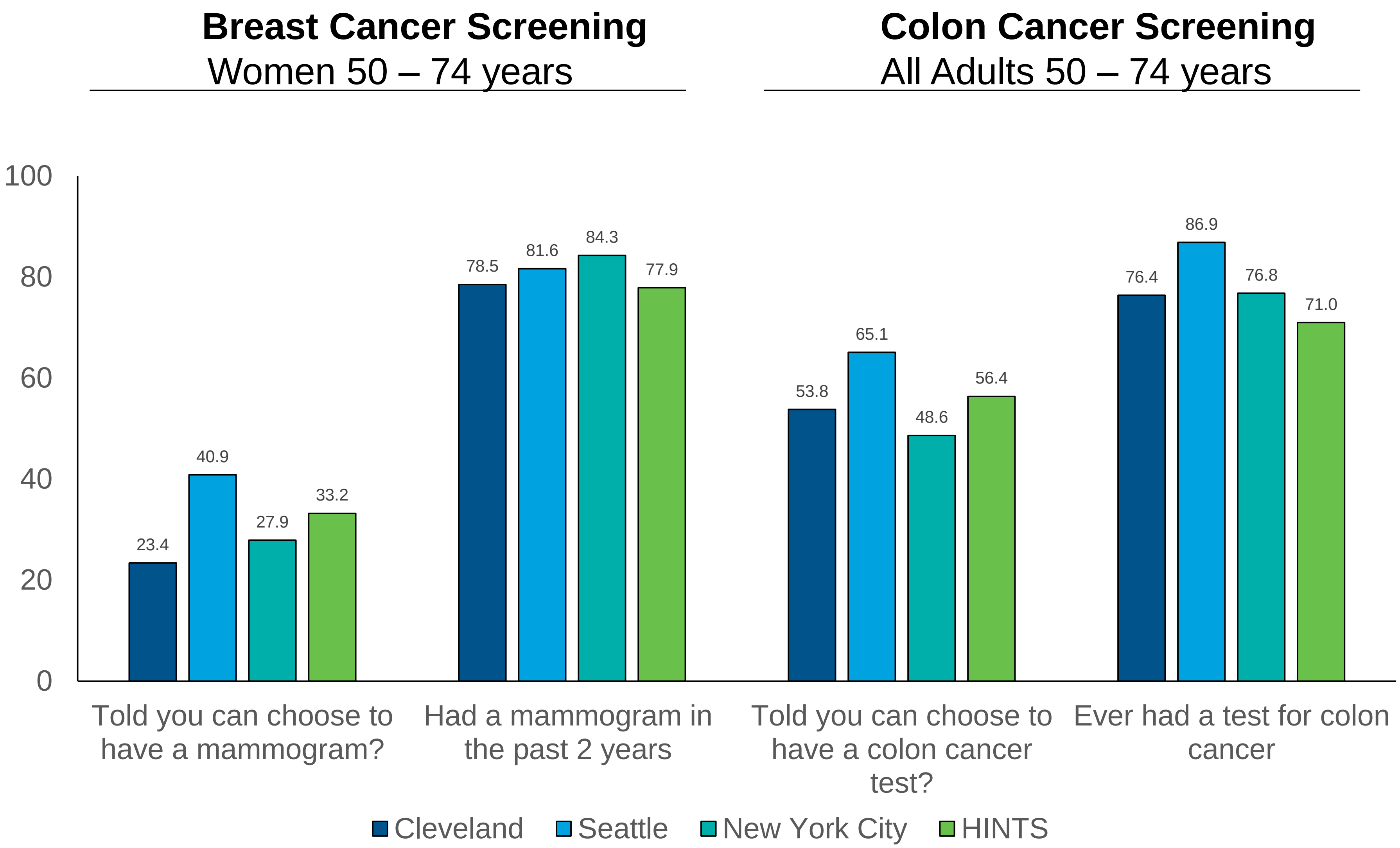
CHINTS Comparisons with SMART BRFSS (General Health)







CHINTS Comparison with HINTS (Cancer Screening)

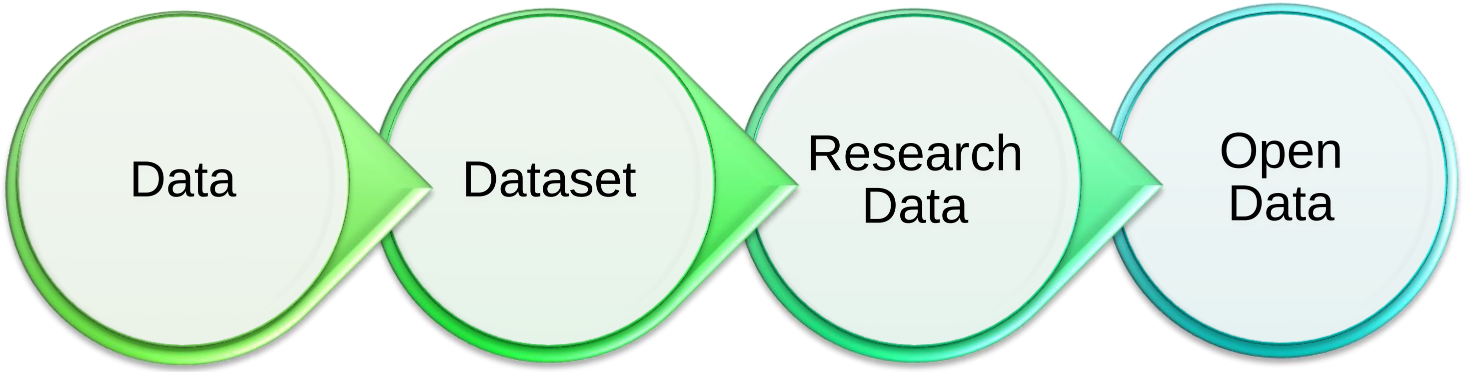


Summary

- It is recommended to do as much ex-ante standardization as possible: target population, measures, mode
- It is important to specify reference population for weighting and comparisons
- CHINTS studies provide examples of comparisons with local and national data using weighted data

What is Open Data[1-6]?

- **Definition: "Publicly available data structured in a way that enables the data to be fully discoverable and usable by end users"**
- **Purpose**
  - Increase data access to businesses, academia, and the general public
  - New insights, innovative solutions, and economic gains



All ***structured information***, unless otherwise noted

A collection of data presented in ***tabular or non-tabular f***orm

Recorded ***factual*** material commonly accepted in the scientific community as ***necessary to validate*** research findings

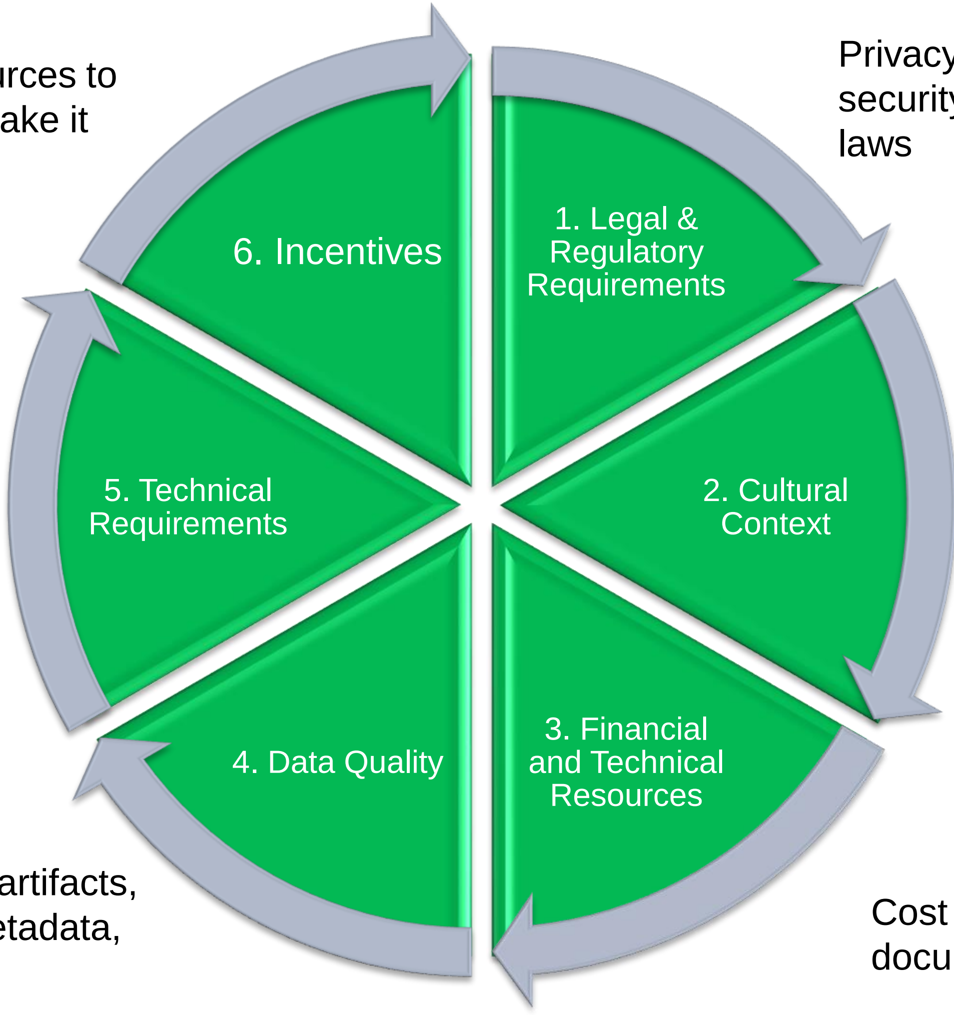
Publicly available data structured in a way that enables the data to be ***fully discoverable*** and usable by end users

Public data: Data asset is or could be made publicly available to all without restrictions
Restricted public data: Data asset is available under certain use restrictions
Non-public data: Data asset is not available to members of the public

Open Data / Data Sharing Considerations

Financial and other resources to encourage activity and make it sustainable

Privacy, confidentiality, and system security. Federal, State, and local laws



Organizational understanding and beliefs, barriers to making data available, leadership directives and support, governance

Preparation, transfer, release, management, and storage of data and related artifacts.

Completeness of data and artifacts, QA/QC, version control, metadata, standards use, timeliness

Cost to prepare, manage, store, document, and access data

Why Open Data Matters[7-8]?

Hysterectomy-Corrected Cervical Cancer Mortality Rates Reveal a Larger Racial Disparity in the United States

Anna L. Beavis, MD, MPH[1]; Patti E. Gravitt, PhD[2]; and Anne F. Rositch, PhD, MSPH[3]

BACKGROUND: The objectives of this study were to determine the age-standardized and age-specific annual US cervical cancer mortality rates after correction for the prevalence of hysterectomy and to evaluate disparities by age and race. **METHODS:** Estimates for deaths due to cervical cancer stratified by age, state, year, and race were derived from the National Center for Health Statistics county mortality data (2000-2012). Equivalently stratified data on the prevalence of hysterectomy for women 20 years old or older from the Behavioral Risk Factor Surveillance System survey were used to remove women who were not at risk from the denominator. Age-specific and age-standardized mortality rates were computed, and trends in mortality rates were analyzed with Joinpoint regression. **RESULTS:** Age-standardized rates were higher for both races after correction. For black women, the corrected mortality rate was 10.1 per 100,000 (95% confidence interval [CI], 9.6-10.6), whereas the uncorrected rate was 5.7 per 100,000 (95% CI, 5.5-6.0). The corrected rate for white women was 4.7 per 100,000 (95% CI, 4.6-4.8), whereas the uncorrected rate was 3.2 per 100,000 (95% CI, 3.1-3.2). Without the correction, the disparity in mortality between races was underestimated by 44%. Black women who were 85 years old or older had the highest corrected rate: 37.2 deaths per 100,000. A trend analysis of corrected rates demonstrated that white women's rates decreased at 0.8% per year, whereas the annual decrease for black women was 3.6% ($P < .05$). **CONCLUSIONS:** A correction for hysterectomy has revealed that cervical cancer mortality rates are underestimated, particularly in black women. The highest rates are seen in the oldest black women, and public health efforts should focus on appropriate screening and adequate treatment in this population. *Cancer* **2017;000:000-000.** *© 2017 American Cancer Society.*

KEYWORDS: Behavioral Risk Factor Surveillance System, cervical cancer, disparities, hysterectomy, mortality, Surveillance, Epidemiology, and End Results (SEER).

- **>12,000 women are diagnosed with cervical cancer / year**
- **>4,000 women die from cervical cancer / year**
- **Correction for hysterectomy prevalence to adjust population-at-risk *N***
- **Mortality ratio for black women versus white women increased from 1.8 to 2.2**

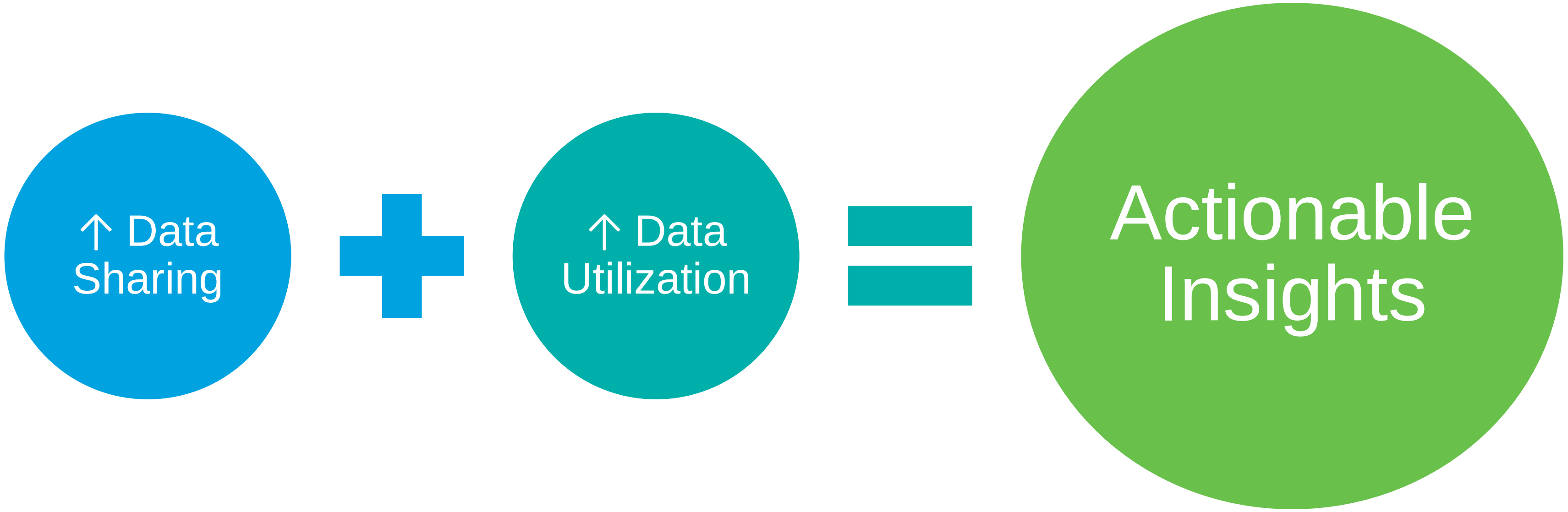
Age-standardized cervical cancer mortality rate, 2000-2012, per 100,000

	All races	White	Black
Uncorrected	3.4 (95% CI, 3.3-3.4)	3.2 (95% CI, 3.1-3.2)	5.7 (95% CI, 5.5-6.0)
Corrected	5.0 (95% CI, 4.9-5.1)	4.7 (95% CI, 4.6-4.8)	10.1 (95% CI, 9.6-10.6)

Data Sources

- *CDC BRFSS*: public
- *NCI SEER*: restricted public
- *CDC Mortality Data*: public, restricted public

Potential for Open Data



"A new understanding from data analysis that leads to something of practical value"

Selected Relevant NIH Funding Opportunities

- **Cancer-Related Behavioral Research through Integrating Existing Data**
  - R01: http://grants.nih.gov/grants/guide/pa-files/PAR-16-256.html
  - R21: https://grants.nih.gov/grants/guide/pa-files/PAR-16-255.html

- **Methodology and Measurement in the Behavioral and Social Sciences**
  - R01: http://grants.nih.gov/grants/guide/pa-files/PAR-16-260.html
  - R21: http://grants.nih.gov/grants/guide/pa-files/PAR-16-261.html

- **NCI Small Grants Program for Cancer Research (Omnibus)**
  - R03: https://grants.nih.gov/grants/guide/pa-files/PAR-16-416.html

Questions?

Contact Information:

Tonja Kyle

Tonja.Kyle@icf.com

Ronaldo Iachan

Ronaldo.Iachan@icf.com

Lew Berman

Lewis.Berman@icf.com

References

1. M-13-13 — Memorandum for the Heads of Executive Departments and Agencies. https://project-open-data.cio.gov/policy-memo/. Accessed on February 7, 2017.
2. NIH Data Sharing Policy. https://grants.nih.gov/grants/policy/data_sharing/data_sharing_chart.doc. Accessed on February 7, 2017.
3. Martin EG, Helbig N, Birkhead GS. Opening health data: what do researchers want? Early experiences with New York's open data portal. J Public Health Manag Pract. 2015 Sep–Oct; 21(15): E1–E7.
4. Martin EG, Helbig N, Shah NR. Liberating data to transform healthcare: New York's open data experience. JAMA. 2014;311(24):2481–2482.
5. Hardy, K, Maurushat, A. Open up government data for Big Data analysis and public benefit. Computer Law and Security Review. 2016.
6. Project Open Data Metadata Schema v1.1. https://project-open-data.cio.gov/v1.1/schema/. Accessed on February 8, 2017.
7. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016;66(1):7–30. doi: 10.3322/caac.21332
8. Beavis, A. L., Gravitt, P. E. and Rositch, A. F. (2017), Hysterectomy-corrected cervical cancer mortality rates reveal a larger racial disparity in the United States. Cancer. doi:10.1002/cncr.30507.
9. Billy Wisniewski and Holly Newman. Challenges Facing the Disclosure Review Board at Census, August 3, 2016. Available at: https://www.census.gov/content/dam/Census/newsroom/press-kits/2016/20160803_challenges_facing_the_disclosure_review_board.pdf. Accessed on 1/18/2017.