Cancer Care Outcomes Research and Surveillance Consortium (CanCORS)

Using Multiple Imputation to both Uncover and Hide

Dave Harrington (dph@jimmy.harvard.edu)

Dana-Farber Cancer Institute, Harvard Statistics and Biostatistics

10 June 2014

Collaborators

Alan Zaslavsky, HMS

Yulei He, CDC

Bronwyn Loong, Australian National University

Paul Catalano, DFCI and HSPH

MaryBeth Landrum, HMS

Many members of the CanCORS Consortium

Outline . . .

The Thread

CanCORS Sites

Group Health Cooperative
(Seattle metropolitan area)

Seattle, WA

Kaiser Permanente Northwest
(Portland metropolitan area)

Portland, OR

Northern California
(8 counties in San Jose, San Francisco/Oakland, and Sacramento areas)

Los Angeles County

Kaiser Permanente Hawaii
(4 major islands)

State of Iowa

Minneapolis, MN

Chicago, IL
(Lakeside and Hines)

Indianapolis, IN

Henry Ford Health System
(Detroit metropolitan area)

Harvard Pilgrim Health Care
(Boston metropolitan area)

New York, NY

Baltimore, MD

Durham, NC

North Carolina
22 central/eastern counties

Nashville, TN

Atlanta, GA

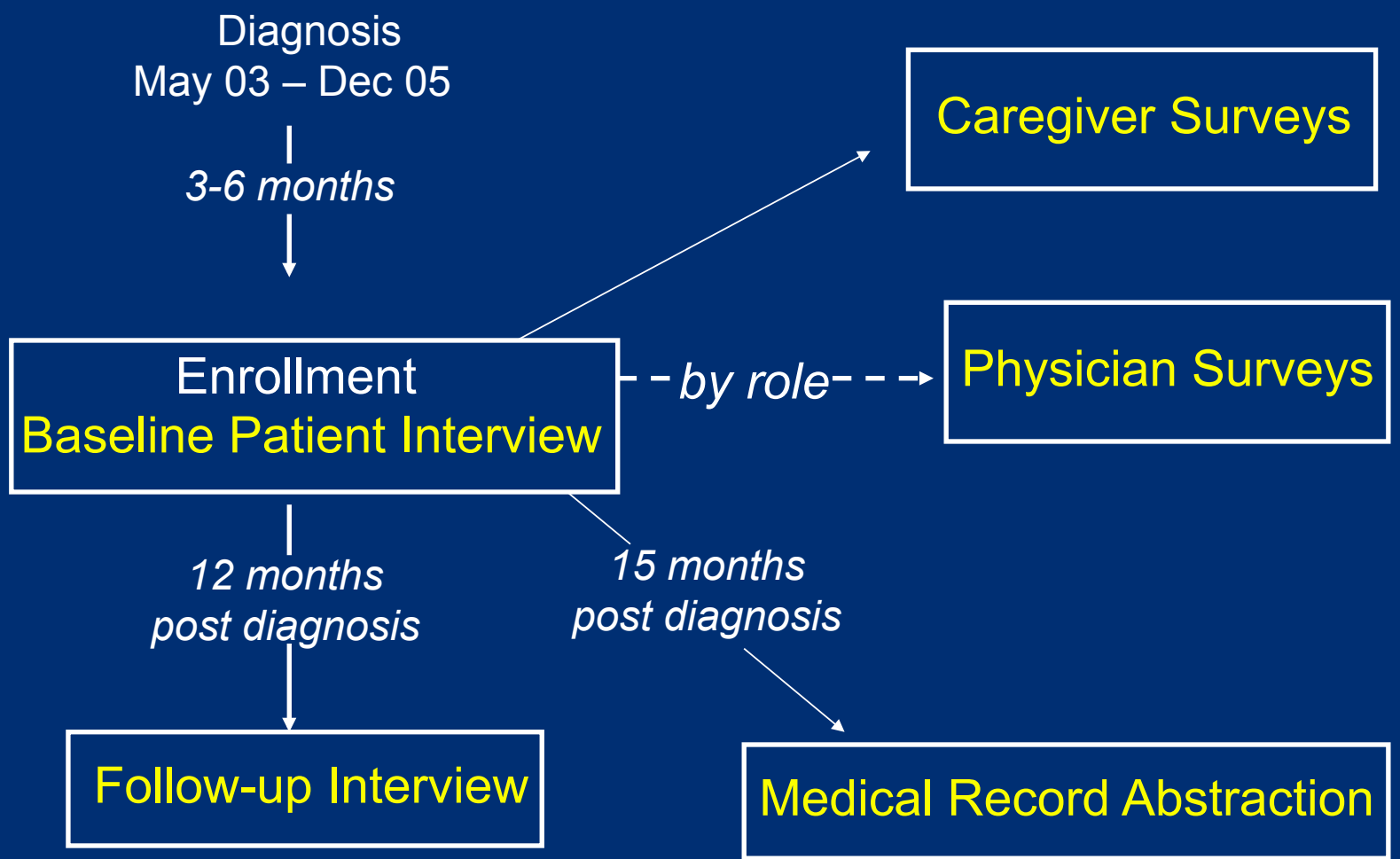
Biloxi, MS

Houston, TX

State of Alabama

- Patients from population-based cohorts in geographic areas
- Patients from integrated health-care delivery systems
- Patients at Veterans Health Administration hospitals

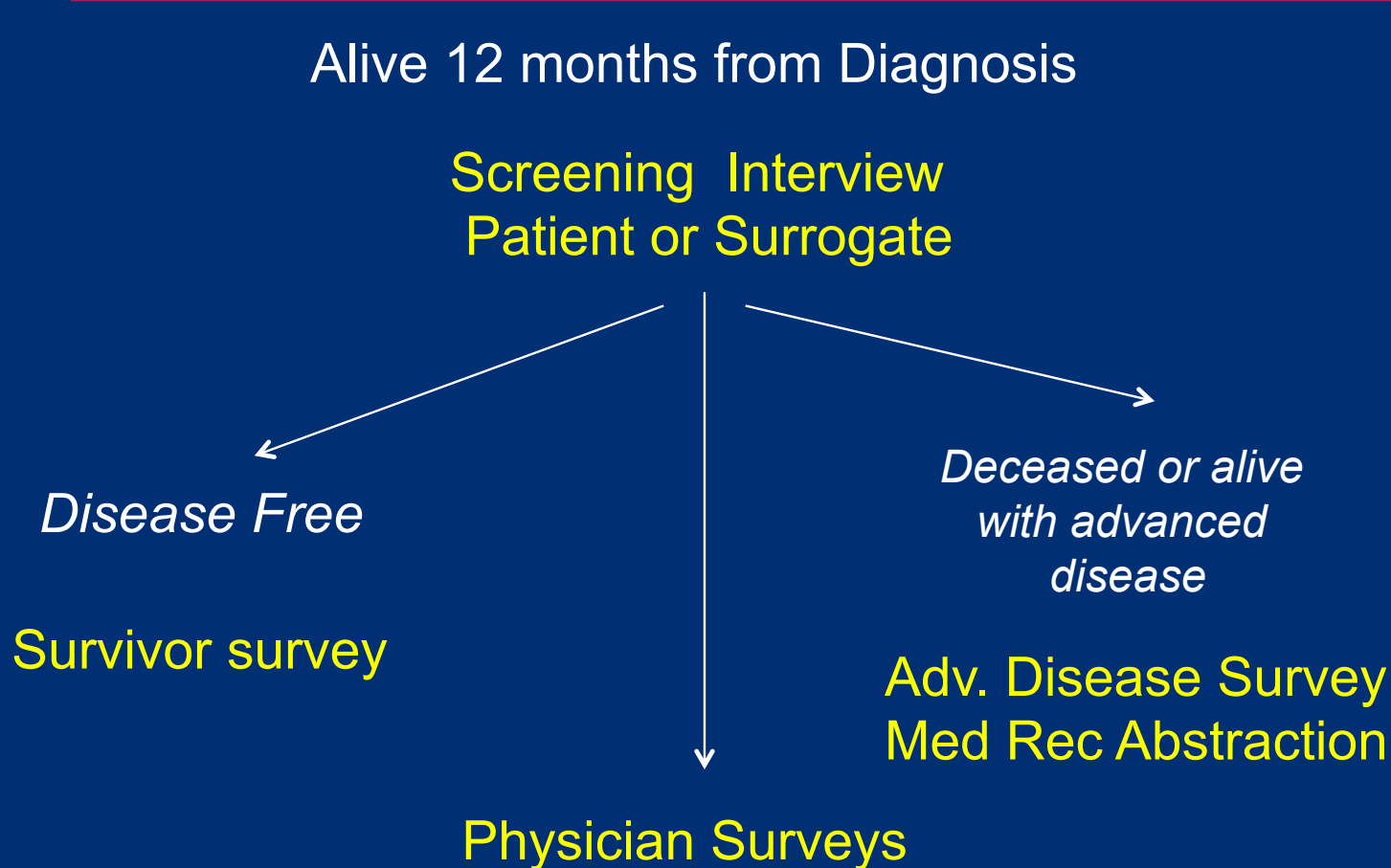
Enrollment Phase DataCollection



Interview Instrument Types

- Full baseline
- Brief baseline
- Surrogate living and decedent

Second wave Phase Data Collection



Enrollment

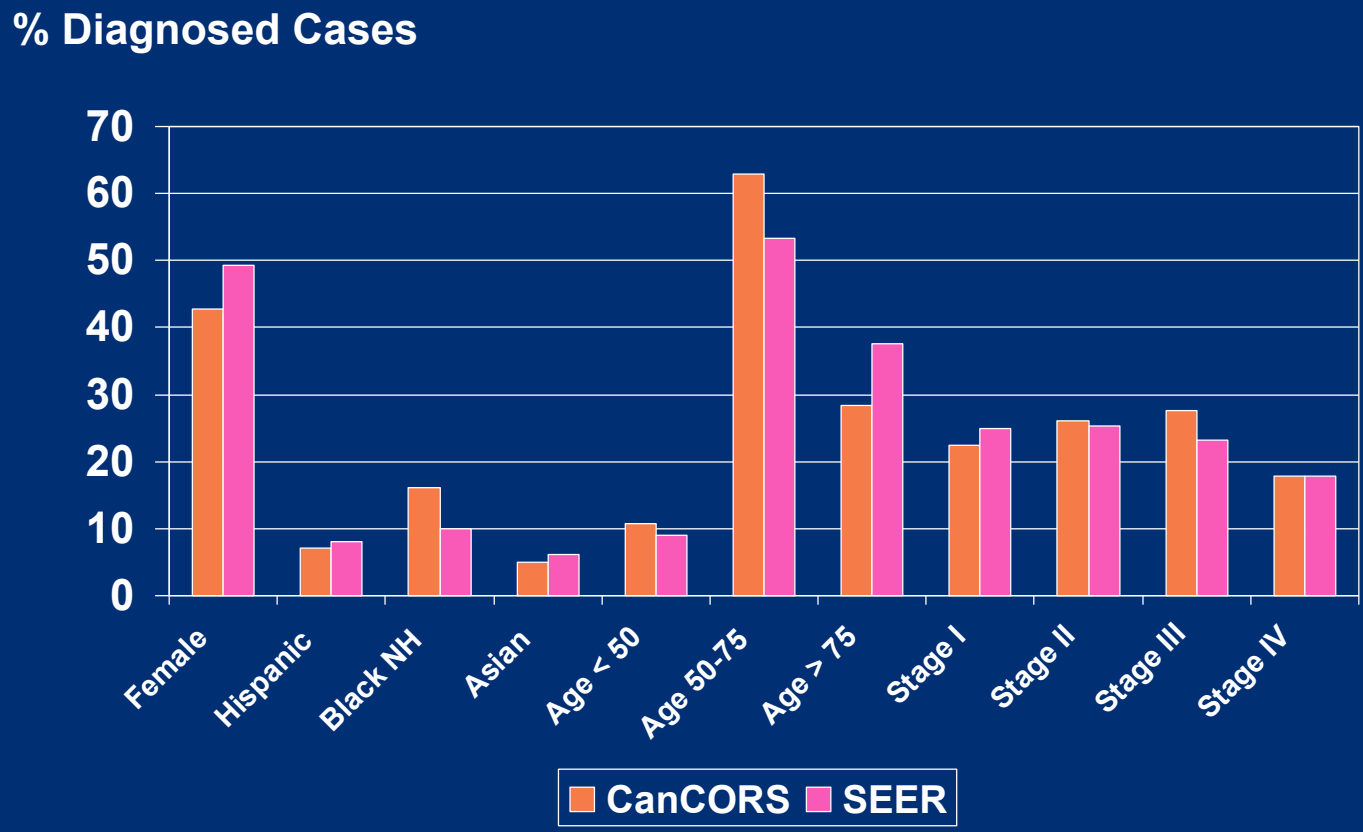
10,061 patients with lung or colorectal cancer enrolled, with data from

- Patient interviews
- Physician surveys
- Medical record abstraction

2,995 patients recontacted in second wave

- Revisions of earlier instruments
- Assessing recurrence surveillance using electronic records in integrated health plans and Medicare Claims

CanCORS vs SEER Colorectal Cancer



(Catalano, *Med Care*, 2012)

Challenges in the Design and Analysis

- Uniform data collection/standards of analysis
- Unit non-response
- Item non-response
- Confidentiality

The Thread

Multiple Imputation

Proposed by Rubin in 1977; Rubin's 1996 JASA review is accessible.

- Schafer 1999 SMMR paper is a primer
- Multiple draws (usually 5 - 10) used to create versions of completed dataset.
    - Draws taken from posterior predictive distribution of missing data, given observed data.
- Each dataset analyzed using traditional complete case analysis.
- Variance estimates account for within and between imputation variability.

The Thread

Missingness in the surveys

Survey	No. Vars	Sample Size	Skips(%)		DO/DK/REF(%)	
			Mean	Range	Mean	Range
Full Baseline	535	5763	50	0-99	1.23	0-37
Brief Baseline	156	1397	48	0-99	2.49	0-22
Surrogate Live	413	1083	50	0-99	1.61	0-28
Surrogate Deceased	473	1827	62	0-99	1.46	0-29
Follow-up	384	6087	65	0-99	0.55	0-33

DO/DK/REF = Drop Out/Don't Know/Refused

Some complete case analyses drop approximately 30% of respondents.

Challenges to MI

- Several different data types
- 4 instruments used in two cancer types
- Block missingness because of missing domains in shorter surveys.
- Extensive skip patterns in interview structure
- Data set periodically refreshed
- Diagnostics difficult

Imputation Strategy

- Use Sequential Regression Multiple Imputation (SRMI), implemented in SAS module IVEWare. (Raghunathan, 2001, *Surv. Meth*)
- Imputation used to create 5 datasets for each of the two cancer types.
- Within a cancer, imputation based on dataset with instrument types combined.
- Skipped questions and missing blocks in shorter surveys imputed
- Skips and block missingness restored after imputation
- To incease *congeniality*, imputation did not use data available to Coordinating Center but not available to data analysts.

Diagnostics

- Examined changes in marginal means, pairwise correlation coefficients.
- Comparisons with complete case and missingness indicator analyses.
- Comparisons with weighted analyses: treat brief/surrogate surveys as item non-response, non-response weights calculated for participation in full survey.

Comparative Analyses of Hospice

Full analysis in Huskamp, et al., Arch Int Med (2009)

Did you participate in a discussion of hospice with your physician?

Table 5 Hospice care analysis results

Predictor	Ref.	CC			Missingness indicator			SRMI		
		EST	SE	*p*-value	EST	SE	*p*-value	EST	SE	*p*-value
Intercept		−1.97	0.51	0.00	−2.08	0.43	0.00	−2.17	0.42	0.00
Race	White									
Black		0.03	0.21	0.90	0.03	0.17	0.86	−0.02	0.17	0.89
Hispanic		−0.51	0.31	0.10	−0.68	0.24	0.00	−0.66	0.24	0.01
Asian		0.24	0.29	0.42	0.20	0.25	0.41	0.21	0.25	0.40
Other		0.46	0.27	0.08	0.46	0.23	0.04	0.42	0.22	0.06

Table is small piece of full logistic regression.

Complete account in He, et al. 2010, *SMMR*

Weighted analyses: QoL scores

Table 7 Estimates of quality-of-life scales

Sample	Scale	CC		Weighting		SRMI	
		Mean	SE	Mean	SE	Mean	SE
All	EORTC_LC	79	0.31	77	0.58	75	0.26
	EORTC_QLQ	77	0.39	74	0.60	71	0.28
	DYSPNEA	63	0.59	59	0.96	59	0.50
Stage I/II	EORTC_LC	82	0.48	81	0.69	78	0.42
	EORTC_QLQ	80	0.60	79	0.83	75	0.51
	DYSPNEA	64	0.95	62	1.34	61	0.73
Stage IV	EORTC_LC	78	0.62	76	1.07	73	0.40
	EORTC_QLQ	75	0.81	73	1.18	68	0.43
	DYSPNEA	63	1.23	59	2.00	58	0.82

The Thread

Challenges in preserving confidentiality with publicly available healthcare microdata

- Minimize disclosure risk.
  - Reduce risk of participant re-identification by malicious user should be acceptably low.
- Maintain data utility.
  - Modifications to the observed data set should maintain variable relationships that are both clinically plausible and reproduce (approximately) relationships in the data.
- Low maintenance for Coordinating Center

Background on partially synthetic data

Idea: Replace the observed values for sensitive variables or key identifiers with synthetic data. Do not release original values of these variables.

- First proposed by Rubin (1993) based on the concepts of multiple imputation.
- Synthetic values are created using samples drawn from the posterior predictive distribution of target population responses given the observed data set.
- Analysts can draw approximately valid inferences about the target population of interest using standard methods for multiply imputed data.

Synthetic data in practice

- Survey of Consumer Finances
- American Communities Survey
- Survey of Income and Program Participation
- US Longitudinal Business Database
- Longitudinal Employer - Household Dynamics
- German IAB Establishment Panel

Identification of variables to synthesize

- **Direct identifiers:** name, postal address, SSN: not available to internal and external investigators, so not synthesized.
- **Quasi-identifiers:** Age, education, sex, marital status, race.
- **Clinical:** Not synthesized; complex structure - difficult to synthesize, accessible by insiders only (not wider population); subject to judgemental variation and incomplete medical record acquisition
- **Sensitive:** All information gathered can be considered sensitive; full synthesis not practical and not useful to data analysts

Details reported in Loong, et al. (2013) *Stat. Med.*

Imputation models

$$f\left(Y_{\text{age}}^{(\text{j})}, Y_{\text{educ}}^{(\text{j})}, Y_{\text{marstat}}^{(\text{j})}, Y_{\text{race}}^{(\text{j})}, Y_{\text{sex}}^{(\text{j})} \middle| Y_0, \mathbf{S}_0\right) , \tag{1}$$

Not practical to draw directly from conditional joint distribution of identifiers, given observed data.

We used sequential regression multiple imputation here as well

Parametric approach to imputation - logisitic regression model to impute sex, multinomial logit models for other variables.

$m = 5$ partially synthetic data sets created

Imputations done in R package *MI* for this project.

Imputation models

- Select predictor variables using stepwise regression within each of 12 sections of survey.
- On average ≈ 50 predictors for each variable to be imputed.
- No interactions, main effects only - reduce risk of overfitting (which increases disclosure risk), reduce computational burden and avoid multicollinearity and separation issues.
- Variables included in all imputation models: survey disposition code, stage, histology, vital status, study site.

Disclosure risk

- **Identification disclosure risk**: potential identification of sampled units in the released data
- **Inferential disclosure risk**: inference of new information about a known participant in the survey, e.g., all participants with the same identifiers have the same income.

Inferential disclosure risk a difficult problem, that is still not well-understood. We focus on identification disclosure risk assessment.

Identification disclosure risk assessment

Duncan and Lambert and risk framework (1989)

- Mimic the behavior of an ill-intentioned public user (an intruder) who has the true values of unique or quasi-identifiers for select target units and wants to identify the records in the synthetic data that have matching identifier values.
- We investigated 3 sets of quasi-identifying values representing varying assumed levels of intruder information
  (i) Set 1: Age, sex, marital status, and race
  (ii) Set 2: Set 1 + education + income level
  (iii) Set 3: Set 2 + disease stage + study site

Identification disclosure risk assessment

A potential identification risk for target record i occurs when its quasi-identifying values match the corresponding values for a record k in synthetic data set j.

Some assumptions/definitions about risk of identification:

- The intruder knows the target is in the survey and the quasi-identifiers of all units in the population.
- If target record has no matching quasi-identifiers among the m imputed data sets, risk of identification is zero.
- Expected match risk: probability of correct match if intruder randomly guesses the match from the candidates with same quasi-identifiers
- Maximum match risk: intruder correctly always correctly identifies record from among potential candidates
- Total match risk: probability that a record is correctly *and* uniquely identified in synthetic data.

Disclosure risk . . .

Compute probabilities of identification under three sets of quasi-identifiers.

(i) *Set 1*: Age, sex, marital status, race

(ii) *Set 2*: Set 1 + income level + education

(iii) *Set 3*: Set 2 + disease stage + PDCRID

Disclosure risk estimates, lung cancer cohort

Using Set 2 of identifiers:

Synthetic data:

- Expected match risk: 217/5000 = 4.3%
- Maximum risk: 925/5000 = 18.5%
- Total match risk: 88/5000 = 1.7%

Observed data:

- Expected match risk: 1770/5000 = 35.4%
- Maximum risk: 5000/5000 = 100%
- Total match risk: 989/5000 = 19.8%

Calculating disclosure risk

Expected match rate, identifier set 2, lung cancer:

- *Total Match Risk*: 88 cases where original record correctly and uniquely matched to synthetic data record.
- *Maximum risk*: 1158 instances where a true record was among identical sets of identifiers in 5 synthetic data sets. Corresponds to 958 unique records
- *Expected matches:* Expected number of matches if intruder guessed randomly among potential set of matches. This is a weighted average first within then across synthetic data sets.

Disclosure risk by set of indentifiers

Table: *Disclosure risk - CanCORS lung cancer* ***partially synthetic*** *data*

Quasi-identifier set	**EMR**		**TMR**	
	Obs.	Syn.	Obs.	Syn.
1	70	38	0	0
2	1770	217	989	88
3	4495	890	4000	717

Data utility assessment

Characterize the quality of what can be learned from the synthetic data, relative to what can be learned from the observed data set.

- Confidence interval overlap (Karr et al. (2006))
- Coverage error due to bias (Cochran (1977))

Data utility

Observed versus synthetic data inferential results

Analytical comparisons based on published studies which analyzed the observed data set.

Reference papers:

- Huskamp et al. (2009). *Discussions with Physicians about Hospice among Patients with Metastatic Lung Cancer.* Arch. Intern. Med., 169 (10), 954-962
- Keating et al. (2010). *Cancer patients' roles in treatment decisions: do characteristics of the decision influence roles?*. Journal of Clinical Oncology, 28:4634 -4370.

Data utility - analytic results comparison

Table: *Descriptive characteristics and estimated probabilities of hospice discussion by race, unadjusted for other covariates. (Standard errors in parentheses)*

Characteristic	**Patients %**		**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.	Obs.	Syn.
Race/ethnicity					< 0.001	0.62
White	73.7	72.0	55.2 (1.3)	54.4 (1.5)		
Black	10.7	11.4	42.6 (1.3)	50.0 (4.1)		
Hispanic	5.9	5.7	40.4 (1.3)	45.8 (5.3)		
Asian	5.1	5.3	49.4 (1.3)	51.6 (6.0)		
Other	4.7	4.8	64.5 (1.2)	54.0 (6.9)		

Data utility - Confidence interval overlap and coverage error

Table: Data utility for coefficient estimates in a logistic model for hospice discussion, unadjusted for other covariates.

Characteristic	$\frac{Bias(q_{syn})}{SE(q_{syn})}$	**CI. overlap**	**CI Cov. Error**
Race/ethnicity			
Black	1.101	0.767	0.196
Hispanic	0.101	0.831	0.051
Asian	0.024	0.803	0.050
Other	1.479	0.707	0.316

Data utility - uncongeniality

What explanation can be given for the change in conclusion of significance?
$\rightarrow$ **Uncongeniality** Meng (1994)

"Analysts and imputer have access to different types of information and data, and assess and use the information and data in different ways"

That is, there are systematic differences between the imputation model and analysis procedure inputs.

For our study, the imputation model conditioned on the entire data set (5000 records) but the analysis procedure analyzed the subset of Stage IV lung cancer patients (1517 records) .

$\rightarrow$ If the imputation model does not capture all the important subgroup relationships, results from the synthetic data may be biased.

Data utility - uncongeniality

Re-ran imputation models conditional on Stage IV lung cancer patients only - large bias for 'black' race factor level is removed.

Table: *Descriptive characteristics and estimated probabilities of hospice discussion by race, unadjusted for other covariates. (Standard errors in parentheses)*

Characteristic	**Patients %**		**Discussed Hospice %**		**p-value**	
	Obs.	Syn.	Obs.	Syn.	Obs.	Syn.
Race/ethnicity					< 0.001	0.003
White	73.7	74.6	55.2 (1.3)	55.6 (1.5)		
Black	10.7	9.9	42.6 (1.3)	42.0 (4.2)		
Hispanic	5.9	6.1	40.4 (1.3)	40.4 (5.3)		
Asian	5.1	5.1	49.4 (1.3)	50.2 (5.8)		
Other	4.7	4.3	64.5 (1.2)	58.5 (6.5)		

The Thread

Currently, no public use data sets released.

Grant officially closes July 31, 2014.

Public web site (www.cancors.org/public) will be available with

- Full study and data documentation
- Bibliography of published papers
- All study instruments and protocols

Outside investigators will be urged to collaborate with members of CanCORS team