National Cancer Institute

Merging Multiple Iterations of HINTS Data

A decade of hints 10
Quantifying the Health Information Revolution through Data Innovation and Collaboration

January 9-10, 2014 | Rockville, MD
NCI Shady Grove, 9609 Medical Center Drive

Sana N. Vieux, MPH

January 9, 2014

U.S. DEPARTMENT
OF HEALTH AND
HUMAN SERVICES
National Institutes
of Health

NATIONAL CANCER INSTITUTE®

Trending on an item: Factors to consider

- **Survey questions are comparable**
  - Questions wording
  - Response options
  - Universe of respondents (skip patterns)

- **For a complete list of items that can be used for trend analysis, visit hints.cancer.gov**

Applied example: Merging 4 HINTS iterations to test for trend

- **Question: Have you ever used email or the internet to communicate with a doctor or doctor’s office?**
    - Response options: Yes/No
- **Universe of respondents: Internet users**

Methods: Before merging

- **Need to ensure variable names and response options are coded *identically* across all datasets**

- **If using HINTS 3, need to first decide which weights to use before merging the data**
  - Test for mode effects
  - Refer to David Cantor’s presentation

Construction of statistical weights for a combined dataset

	Final sample weights	Replicate weights 1-50	Replicate weights 51-100
HINTS 3 Mail Sample	HINTS3 Mail Final Weight (mwgt0)	**HINTS3 Mail Replicate Weights (mwgt1-mwgt50)**	HINTS3 Mail Final Weight (mwgt0)
HINTS3 RDD Sample	HINTS3 RDD Final Weight (rwgt0)	HINTS3 RDD Final Weight (rwgt0)	**HINTS3 RDD Replicate Weights (rwgt1-rwgt50)**
Combined Data	Final Weight (twgt0)	Final Replicate Weights (twgt1-twgt50)	Final Replicate Weights (twgt51-twgt100)

Replicate weights for each respective iteration only contributes variance for that iteration

Testing for mode effects (SAS)

First: Create an array that combines weights from the RDD and Mail samples

```
data h07mergewts; **User-defined dataset names;
set c.hints2007;

array h07mwts[50] mwgt1-mwgt50; *Mail replicate weights;
array h07rwts[50] rwgt1-rwgt50; *RDD (Phone) replicate weights;
array h07twts[100] twgt1-twgt100; *Combined replicate weights;
**Note: Sampflag should be used to distinguish between mode;

if sampflag = 1 then do i = 1 to 50;*Address (Mail) sample;
 twgt0 = mwgt0;
 h07twts[i] = h07mwts[i];
 h07twts[i+50] = mwgt0;
 end;
else if sampflag = 2 then do i = 1 to 50;***RDD (Phone) sample;
 twgt0 = rwgt0;
 h07twts[i] = rwgt0;
 h07twts[i+50] = h07rwts[i];
 end;

run;
```

Testing for mode effects (SUDAAN)

Second: Run a t-test to test for differences in responses between RDD and Mail samples

```
***T Tests of differences in outcome by mode ***;
proc descript data=h07mergewts design=jackknife ddf = 98;
weight twgt0;
jackwgts twgt1-twgt100 / adjjack=.98;
class sampflag;
var talkdoctor; **Outcome of interest;
contrast sampflag = (1 -1);
run;
```

- **If the P-value is NS:**
  - **There are no significant differences in responses between the mail and RDD samples**
    - **Use HINTS3 combined weights (cwgt) to merge with the rest of the datasets.**

HINTS Statistical Weights

- **All HINTS iterations contain a final sample weight and 50 replicate weights**
- **Final sample weight is used to calculate population estimates**
- **Replicate weights are used to calculate accurate standard error of estimates using the jackknife replication method**

Construction of statistical weights for a combined data file (Table 2-1)

	Final sample weights	Replicate weights 1-50	Replicate weights 51-100	Replicate weights 101-150	Replicate weights 151-200
HINTS 1 (2003)	HINTS 1 Final Weight (fwgt)	**HINTS 1 Replicate Weights (fwgt1-fwgt50)**	HINTS 1 Final Weight (fwgt)	HINTS 1 Final Weight (fwgt)	HINTS 1 Final Weight (fwgt)
HINTS 2 (2005)	HINTS 2 Final Weight (fwgt)	HINTS 2 Final Weight (fwgt)	**HINTS 2 Replicate Weights (fwgt1-fwgt50)**	HINTS 2 Final Weight (fwgt)	HINTS 2 Final Weight (fwgt)
HINTS 3 (2008*)	HINTS 3 Final Weight	HINTS 3 Final Weight	HINTS 3 Final Weight	**HINTS 3 Replicate Weights**	HINTS 3 Final Weight
HINTS 4 (2011)	HINTS 4 Final Weight (person_finalwt0	HINTS 4 Final Weight (person_finalwt0)	HINTS 4 Final Weight (person_finalwt0)	HINTS 4 Final Weight (person_finalwt0)	**HINTS 4 Replicate Weights (person_finalwt1-person_finalwt50)**
Combined Data	Final Weight (nfwgt0)	Final Replicate Weights (nfwgt1-nfwgt50)	Final Replicate Weights (nfwgt51-nfwgt100)	Final Replicate Weights (nfwgt101-nfwgt150)	Final Replicate Weights (nfwgt151-nfwgt200)

- **Note: HINTS 3 allows for utilizing the RDD Weights (rwgt0), the mail weights (mwgt0), or the combined weights (cwgt0)

- Replicate weights for each respective iteration only contributes variance for that iteration
  - See Cochran, 1977 reference for formula to estimate the variance

Construction of statistical weights for a combined data file—5 Iterations

	Final sample weights	Replicate weights 1-50	Replicate weights 51-100	Replicate weights 101-150	Replicate weights 151-200	Replicate weights 201-250
HINTS 1 (2003)	HINTS 1 Final Weight (fwgt)	**HINTS 1 Replicate Weights (fwgt1-fwgt50)**	HINTS 1 Final Weight (fwgt)	HINTS 1 Final Weight (fwgt)	HINTS 1 Final Weight (fwgt)	HINTS 1 Final Weight (fwgt)
HINTS 2 (2005)	HINTS 2 Final Weight (fwgt)	HINTS 2 Final Weight (fwgt)	**HINTS 2 Replicate Weights (fwgt1-fwgt50**)	HINTS 2 Final Weight (fwgt)	HINTS 2 Final Weight (fwgt)	HINTS 2 Final Weight (fwgt)
HINTS 3 (2008*)	HINTS 3 Final Weight*	HINTS 3 Final Weight*	HINTS 3 Final Weight*	**HINTS 3 Replicate Weights***	HINTS 3 Final Weight*	HINTS 3 Final Weight*
HINTS 4-Cycle 1 (2011)	HINTS 4-Cycle 1 Final Weight (person_finalwt0)	HINTS 4-Cycle 1 Final Weight (person_finalwt0)	HINTS 4-Cycle 1 Final Weight (person_finalwt0)	HINTS 4-Cycle 1 Final Weight (person_finalwt0)	**HINTS 4 –Cycle 1 Replicate Weights (person_finalwt1-person_finalwt50)**	HINTS 4-Cycle 1 Final Weight (person_finalwt0)
HINTS 4-Cycle 2 (2012)	HINTS 4-Cycle 2 Final Weight (person_finalwt0)	HINTS 4-Cycle 2 Final Weight (person_finalwt0)	HINTS 4-Cycle 2 Final Weight (person_finalwt0)	HINTS 4-Cycle 2 Final Weight (person_finalwt0)	HINTS 4-Cycle 2 Final Weight (person_finalwt0)	**HINTS 4 –Cycle 2 Replicate Weights (person_finalwt1-person_finalwt50)**
Combined Data	Final Weight (nfwgt0)	Final Replicate Weights (nfwgt1-nfwgt50)	Final Replicate Weights (nfwgt51-nfwgt100)	Final Replicate Weights (nfwgt101-nfwgt150)	Final Replicate Weights (nfwgt151-nfwgt200)	Final Replicate Weights (nfwgt201-nfwgt250)

- *Note: HINTS 3 allows for utilizing the RDD Weights (rwgt0), the mail weights (mwgt0), or the combined weights (cwgt0)

- Replicate weights for each respective iteration only contributes variance for that iteration
  - See Cochran, 1977 reference for formula to estimate the variance

Jackknife Estimate of Variance

Full sample estimate	$\hat{\theta}$
Replicate estimate (i=1,…k)	$\hat{\theta}_i$
Jackknife estimate of variance	$Var(\hat{\theta}) = \frac{k-1}{k}\sum_{i=1}^{k}\left(\hat{\theta}_i - \hat{\theta}\right)^2$

Note: K= Number of replicate weights

Creating a combined dataset

- **Refer to Table 2-1 in the workbook**
- **Final combined dataset will have:**
  - 1 final sample weight (NFWGT0)
  - 200 replicate weights (NFWGT1—NFWGT200)
- **A note about the denominator degrees of freedom (DDF)**
  - 49*k, where k is the number of iterations of HINTS data used in analysis

Statistical Analysis

- Crosstabulation table of population estimates of the outcome for each HINTS iteration
- Decide which weights to use for HINTS3
  - No significant differences in the outcome between the modes
  - Therefore, we used cwgt0 and cwgt1-50
- SUDAAN code to test for mode effects in the appendix

Statistical Analysis

- **Multivariable logistic regression regressing the outcome on age, gender, and education**
  - Tested for three orthogonal trends
    - Cubic, Quadratic, and Linear
  - Computed predicted marginals
  - Gender*SurveyYear interaction

Measures

- **Outcome: “Have you ever used e-mail or the internet to communicate with a doctor or doctor’s office?”**
  - Yes/No
- **Sociodemographic variables**
  - Gender (Male/Female)
  - Age (18-34, 35-39, 40-44, 45+)
  - Education (Less than HS, HS Graduate, Some college, College graduate)
- **Survey Year**
  - Variable to indicate each HINTS iteration

Results (Table 2-2)

	HINTS1	HINTS2	HINTS3	HINTS4
In the last 12 months, have you used email or the internet to communicate with a doctor or doctor’s office?				
Yes	7.00%	9.62%	13.59%	19.11%
No	93.00%	90.38%	86.41%	80.89%

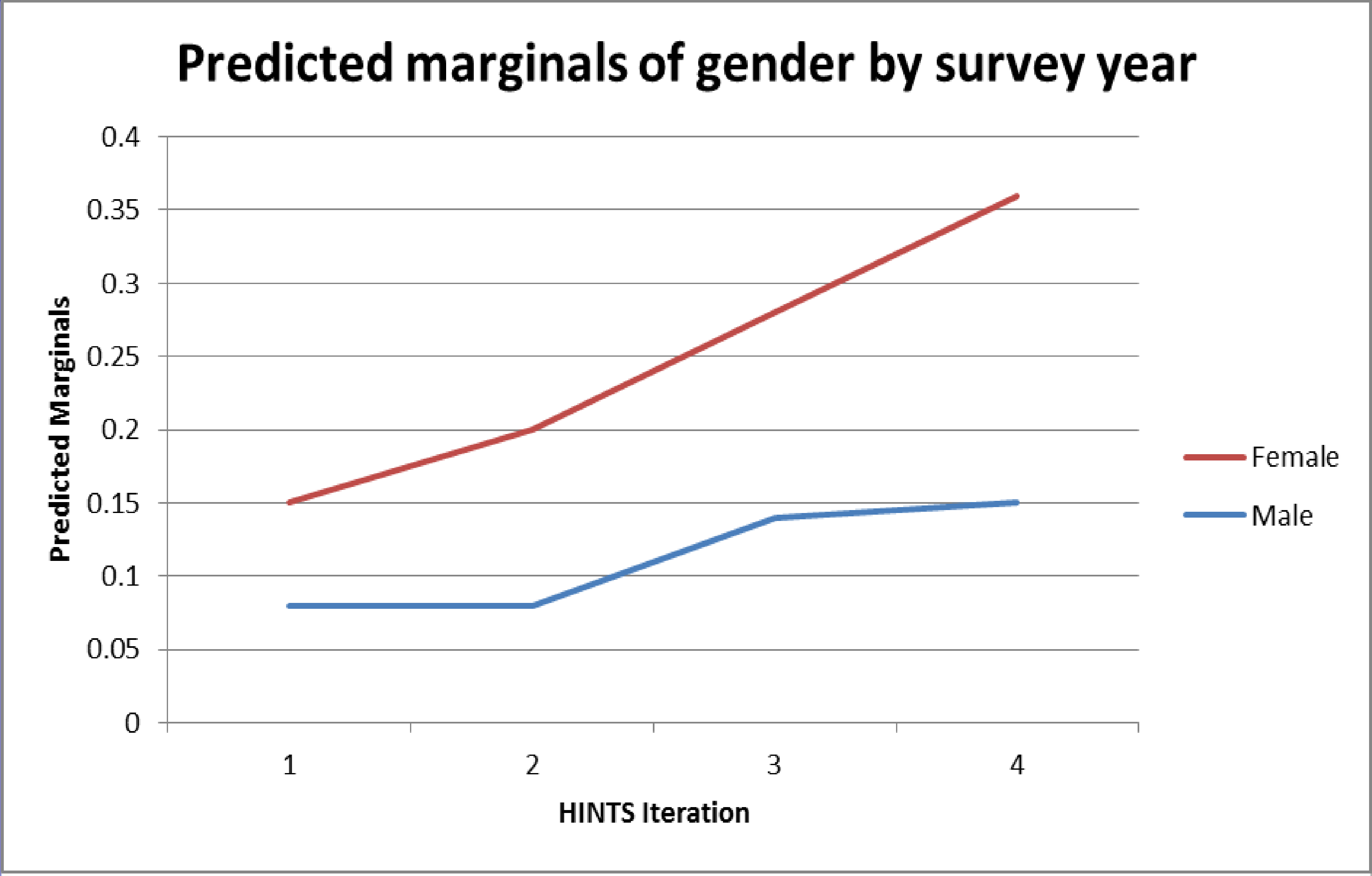
Proc rlogist (SUDAAN)

```
proc rlogist data = hintsmerge design = jackknife ddf=196;
weight nfwgt0;
jackwgts nfwgt1-nfwgt200 / adjjack = 0.98;
class survyear agegrpa educa gender;
model talkdoctor = survyear agegrpa educa gender
                   survyear*gender;
reflev survyear = 1 educa = 1 gender = 1;
predmarg survyear survyear*gender;
effects survyear = (-1 3 -3 1)/name = "Cubic trend";
effects survyear = (1 -1 -1 1)/name = "Quadratic trend";
effects survyear = (-3 -1 1 3)/name = "Linear trend";
run;
```

Results (Table 2-4)

Variable	OR	95% CI	P-Value
Survey Year			—
2003	1.00	---	
2005	1.02	0.70 - 1.48	
2008	1.91	1.42 - 2.57	
2011	2.14	1.48 - 3.09	
Education			0.0000
Less than HS	**1.00**	—	
HS Graduate	**1.02**	**0.61 - 1.71**	
Some College	**1.64**	**0.99 - 2.71**	
College Graduate	**2.57**	**1.58 - 4.16**	
Gender			—
Male	**1.00**	—	
Female	**0.82**	**0.61 - 1.09**	
SurveyYear*Gender			**0.0122**
2003, Male	**1.00**	**1.00 - 1.00**	
2003, Female	**1.00**	**1.00 - 1.00**	
2005, Male	**1.00**	**1.00 - 1.00**	
2005, Female	**1.82**	**1.17 - 2.85**	
2008, Male	**1.00**	**1.00 - 1.00**	
2008, Female	**1.20**	**0.83 - 1.75**	
2011, Male	**1.00**	**1.00 - 1.00**	
2011, Female	**1.82**	**1.17 - 2.83**	

Results



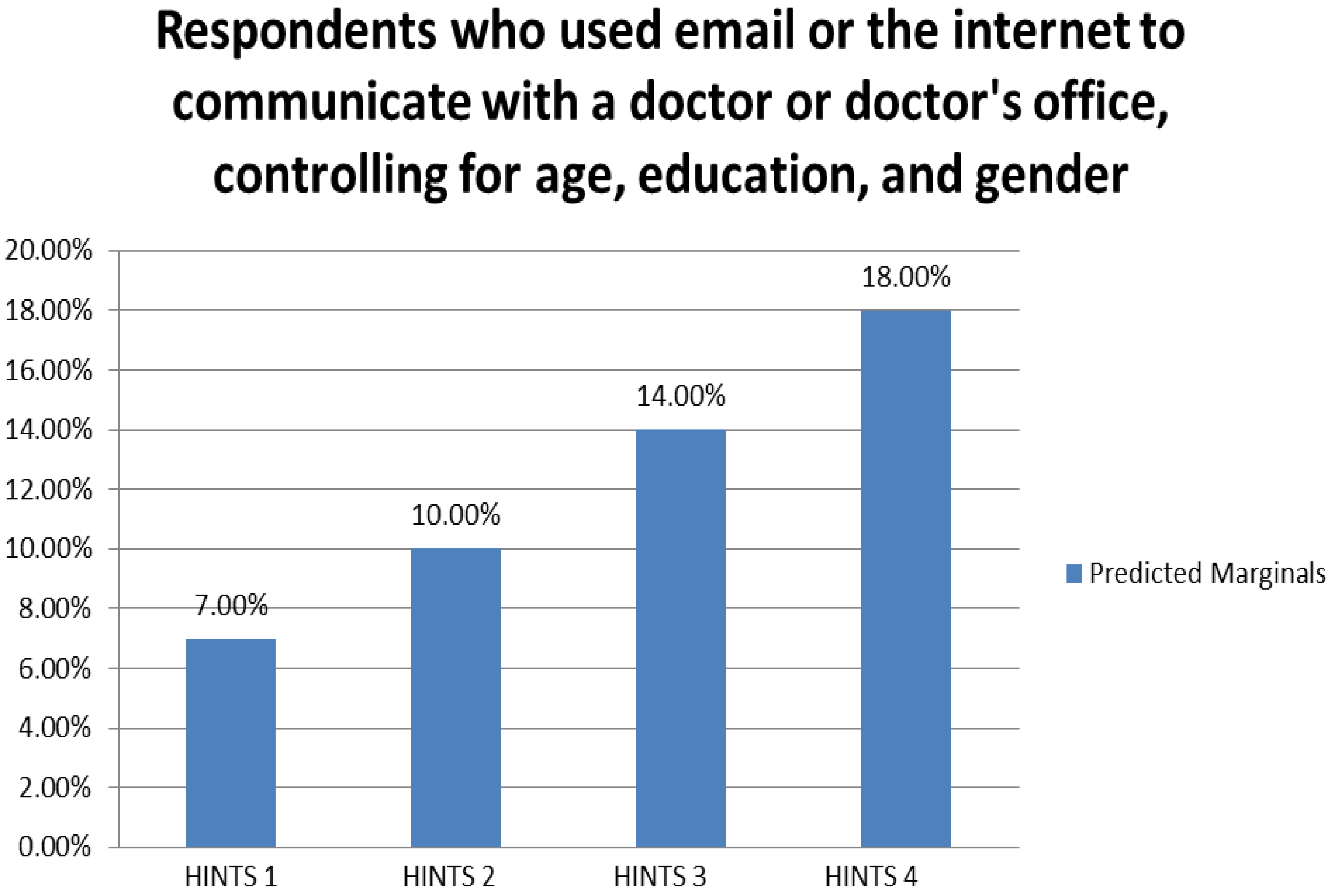
Wald F = 309.95, P-value < 0.0001

Results (Table 2-3)

- **Test of Trend**

Trend	F	P-value
Cubic Trend	0.14	0.7104
Quadratic Trend	0.00	0.9558
Linear Trend	99.36	0.000

Results


Respondents who used email or the internet to communicate with a doctor or doctor's office, controlling for age, education, and gender
20.00%
18.00%
16.00%
14.00%
12.00%
10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
7.00%
10.00%
14.00%
18.00%
HINTS 1
HINTS 2
HINTS 3
HINTS 4
Predicted Marginals

Merging HINTS 3 with HINTS-Puerto Rico

Overview of HINTS-Puerto Rico

- Spanish translation of the HINTS 3 (2008) survey
- N = 639
- 95% Hispanic
- RDD sample and weights
- See HINTS Brief #18 for more information about the HINTS-PR Survey implementation

HINTS3 and HINTS-PR Applied Example

- **Goal: Demonstrate how to merge HINTS 3 and HINTS-PR**

HINTS3 and HINTS-PR Applied Example

- **Question: Have you ever looked for information about cancer from any source?**
  - Response options: Yes/No
- **Universe of respondents: All Respondents**

Methodology

- **Data collection:**
  - RDD and CATI by experienced bilingual Puerto Rican interviewers

- **To keep mode consistent, only the RDD sample of HINTS3 will be used in this analysis**

Weights and Merging

- **The number and type of replicate weights differs between HINTS3 and HINTS-PR**

	HINTS 3 RDD Sample	HINTS PR
Replicate Weights	50	48
Replication Method	JK1	JKn
Sampling Strata	2	8
Jackknife Multiplier	0.98	0.83

Construction of statistical weights for a combined dataset (Table 3-1)

	Final sample weights	Replicate weights 1-50	Replicate weights 51-98
HINTS 3	HINTS 3 Final Weight (rwgt0)	**HINTS 3 RDD Replicate Weights (rwgt1-rwgt50)**	HINTS 3 Final Weight (rwgt0)
HINTS PR	PR Final Weight (r12wgt0)	PR Final Weight (r12wgt0)	**PR Replicate Weights (r12wgt1-r12wgt48)**
Combined Data	Final Weight (twgt0)	Final Replicate Weights (twgt1-twgt50)	Final Replicate Weights (twgt51-twgt98)

Replicate weights for each respective iteration only contributes variance for that iteration

Creating a combined dataset

- **Refer to Table 3-1 in the workbook**
- **Final combined dataset will have:**
  - 1 final sample weight (TWGT0)
  - 98 replicate weights (TWGT1—98)
- **DDF = 89**
- **Need additional code to properly apply the correct multipliers to each replicate weight in the combined dataset**

Statistical Analysis

- **Crosstabulation table of population estimates of the outcome for each HINTS iteration**
- **Chi-square tests were conducted for multiple comparisons between HINTS 3 and HINTS PR**
  - Mainland US vs. Puerto Rico
  - Non-Hispanics in Mainland US vs. Hispanics in Mainland US vs. Hispanics in Puerto Rico
  - Hispanics in Mainland US vs. Hispanics in Puerto Rico

Statistical Analysis

- **Two multivariable logistic regression models**
  - First: Regressing the outcome on HINTS iteration, controlling for age, gender, and education
  - Second: Regressing the outcome on ethnicity, controlling for age, gender, and education

Measures

- **Outcome: “Have you ever looked for information about cancer from any source?”**
  - Yes/No
- **Sociodemographic variables**
  - Gender (Male/Female)
  - Age (18-34, 35-39, 40-44, 45+)
  - Education (Less than HS, HS Graduate, Some college, College graduate)
  - Ethnicity (US Mainland Hispanics, US Mainland Non-Hispanics, and Puerto Rico Hispanics)
- **HINTS Iteration**
  - Variable to indicate each HINTS iteration

Results

Table 3-2: Comparing U.S. Mainland vs. Puerto Rico in seeking cancer information from any source

Seek Info about cancer	Mainland		Puerto Rico		Chi-Square	P-value
	N	%	N	%	36.83	0.0000
Yes	1911	39.40%	181	28.11%		
No	2162	60.60%	458	71.89%		
Total	4073	100.00%	639	100.00%		

Results

Table 3-3: Comparing percent of Hispanics on the Mainland U.S. vs. Non-Hispanics on the Mainland vs. Hispanics in Puerto Rico who sought information about cancer from any source

Seek Info about cancer	Non-Hispanics in Mainland US		Hispanics in Mainland US		Hispanics in PR		Chi-Square	P-value
	N	%	N	%	N	%	30.15	0.0000
Yes	1683	42.78%	90	21.19%	167	27.55%		
No	1718	57.22%	207	78.81%	428	72.45%		
Total	3401	100.00%	297	100.00%	595	100.00%		

Results

Table 3-4: Comparing percent of Hispanics on the Mainland vs. Hispanics in Puerto Rico who sought information about cancer from any source

Seek Info about cancer	Hispanics in Mainland US		Hispanics in PR		Chi-Square	P-value
	N	%	N	%	3.32	0.0717
Yes	90	21.19%	167	27.55%		
No	207	78.81%	428	72.45%		
Total	297	100.00%	595	100.00%		

Multivariable Logistic Regression

```
proc rlogist data = hintsmerge design = jackknife ddf= 89;
weight twgt0;
jackwgts twgt1-twgt98;
jackmult 50*0.98 48*0.83; **Applying different multipliers
to each respective dataset;
class survyear agegrpa educa gendern/nofreq;
model HC08SeekCancerInfo = survyear agegrpa educa gendern;
reflev survyear = 1 gendern=1 agegrpa=1 educa=1;
run;
```

Results

	Odds of seeking cancer information		
Variable	**OR**	**95% CI**	**P-Value**
Survey Year			0.0005
US Mainland	1.00	---	
Puerto Rico	0.64	0.50 - 0.82	
Age			0.0000
18 – 34	1.00	---	
35 – 39	1.78	1.06 - 2.98	
40 – 44	1.60	1.05 - 2.44	
45+	2.02	1.52 - 2.69	
Gender			0.0001
Male	1.00	---	
Female	1.56	1.27 - 1.92	
Education			0.0000
Less than HS	1.00	---	
HS Graduate	2.12	1.39 - 3.24	
Some College	3.71	2.43 - 5.67	
College Graduate	5.82	3.82 - 8.86	

Multivariable Logistic Regression

```
proc rlogist data = hintsmerge design = jackknife ddf = 89;
weight twgt0;
jackwgts twgt1-twgt98;
jackmult 50*0.98 48*0.83; **Applying different multipliers to each respective dataset;
class ethnicity agegrpa educa gendern/nofreq;
model HC08SeekCancerInfo = ethnicity agegrpa educa gendern ;
reflev ethnicity = 1 gendern=1 agegrpa=1 educa=1;
effects ethnicity = (1 0 -1); **Comparing U.S. Hispanics vs. Puerto Rico Hispanics;
effects ethnicity = (1 -1 0); **Comparing Mainland U.S. Hispanics vs. Mainland US non-Hispanics;
run;
```

Results (Table 3-6)

	Odds of seeking cancer information		
Variable	**OR**	**95% CI**	**P-Value**
Ethnicity			0.0004
Hispanics in the US	1.00	---	
Non-Hispanics in the US	1.64	1.11 - 2.42	
Hispanics in Puerto Rico	0.99	0.66 - 1.47	
Age			0.0003
18 – 34	1.00	---	
35 – 39	1.80	1.06 - 3.03	
40 – 44	1.62	1.06 - 2.48	
45+	1.94	1.44 - 2.60	
Gender			0.0001
Male	1.00	---	
Female	1.55	1.26 - 1.91	
Education			0.0000
Less than HS	1.00	---	
HS Graduate	1.91	1.22 - 3.00	
Some College	3.35	2.19 - 5.13	
College Graduate	5.21	3.33 - 8.16	

Results (Table 3-6a)

Comparing the odds of different ethnic groups in seeking information about cancer, controlling for age, education, and gender

	Wald F	P-value
Hispanics in Mainland US vs. Hispanics in Puerto Rico	<0.01	0.9490
Mainland US Hispanics vs. Mainland US Non-Hispanics	6.36	0.0133

Questions?

Thank you!

Rick Moser

moserr@mail.nih.gov

Sana Vieux

vieuxs@mail.nih.gov