

# Examining Changes Across Years Using HINTS 2003 and 2005 Data



May 5, 2007

Richard P. Moser, Ph.D.<sup>1</sup>

William W. Davis, Ph.D.<sup>1</sup>

William Waldron, B.S.<sup>2</sup>

Timothy McNeel, B.A.<sup>2</sup>

<sup>1</sup>National Cancer Institute; <sup>2</sup> Information Management Services, Inc.

# Goals of the Training



- ❖ Demonstrate how separate HINTS 2003 and HINTS 2005 data can be used to:
  - Test for differences in outcomes between survey iterations
    - ❖ Across groups or by subgroups
- ❖ Demonstrate using a combined HINTS 2003 and HINTS 2005 data set to:
  - Test for differences in outcomes between survey iterations
    - ❖ Across groups or by subgroups
  - Test for differences in outcomes controlling for covariates
    - ❖ Across groups or by subgroups
  - Gain a larger sample size
    - ❖ Used to calculate means and variances
    - ❖ Most useful for variables not expected to change over time

# Overview of Analyses



- ❖ Outcome for all analyses: “Have you ever visited an Internet web site to learn specifically about cancer?”
  - HC-29 in HINTS 2003
  - CA-15 in HINTS 2005
  - These were modified for the analyses
- ❖ Covariates:
  - Agegroup (5 levels)
  - Education (4 levels)
  - Race/Ethnicity (4 levels)
  - Gender
  - Income (4 levels)
  - Hintsyear (2 levels)
- ❖ Syntax examples
  - Exclusive use of SAS and SUDAAN
  - Other programs can be used (e.g., STATA, WesVar)



# Overview (cont.)



- ❖ Techniques here are general
  - Can be used for other HINTS analyses
  - Can be used with other data sets with multiple years
- ❖ Assumptions
  - Two independent cross-sectional surveys
  - Same questions, formats, and interpretation
  - Replicate weights for both surveys
- ❖ References
  - Korn and Graubard (1999) Analysis of Health Surveys
  - Lee, Davis, *et al.* (2006) Examining trends and averages using combined cross-sectional survey data from multiple years.  
[www.chis.ucla.edu/pdf/chis\\_pooling\\_10302006.pdf](http://www.chis.ucla.edu/pdf/chis_pooling_10302006.pdf)

# Overview (cont.)



## ❖ We will demonstrate

- Analyses using separate (2003 and 2005) data
- Analyses using combined (2003 and 2005) data

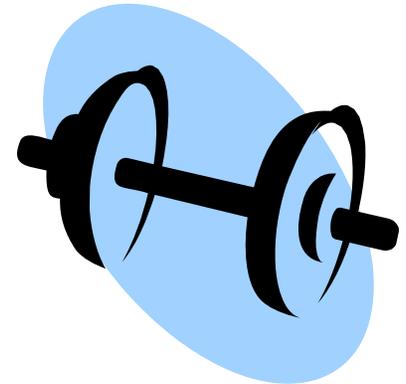
## ❖ Many analyses can be done using either

- However, once the combined file is constructed many analyses are obtained easily

# HINTS Statistical Weights



- ❖ Both HINTS 2003 and 2005 contain full sample and 50 replicate weights.
- ❖ Weights derived from
  - selection probabilities,
  - response rates,
  - post-stratification adjustment.
- ❖ HINTS 50 replicate weights obtained by deleting  $1/50^{\text{th}}$  of the respondents (and re-weighting)
  - Each replicate is similar to a HINTS yearly sample
  - The variability in replicate estimates can be used to estimate variance



# Jackknife Estimate of Variance for HINTS



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



# Analyses Using Separate Data Sets



# Testing for Change Using Separate Datasets



- ❖ Do not need combined data
- ❖ Do need the following information:
  - Estimates and variances from each survey year\*

Year	True value	Estimated value	Variance of estimate
2003	$\theta_{2003}$	$\hat{\theta}_{2003}$	$v(\hat{\theta}_{2003})$
2005	$\theta_{2005}$	$\hat{\theta}_{2005}$	$v(\hat{\theta}_{2005})$
<i>Change</i>	$\Delta = \theta_{2005} - \theta_{2003}$	$\hat{\Delta} = \hat{\theta}_{2005} - \hat{\theta}_{2003}$	$v(\hat{\Delta}) = v(\hat{\theta}_{2005}) + v(\hat{\theta}_{2003})$

\*From SUDAAN proc descript or proc crosstab

# Assessing For Change Using Separate Data Sets



- ❖ A: Has there been a change in the percentage of respondents who looked for cancer information online between 2003 and 2005?
- ❖ B: Has there been a change in the percentage of respondents who looked for cancer information online by demographic subgroup?

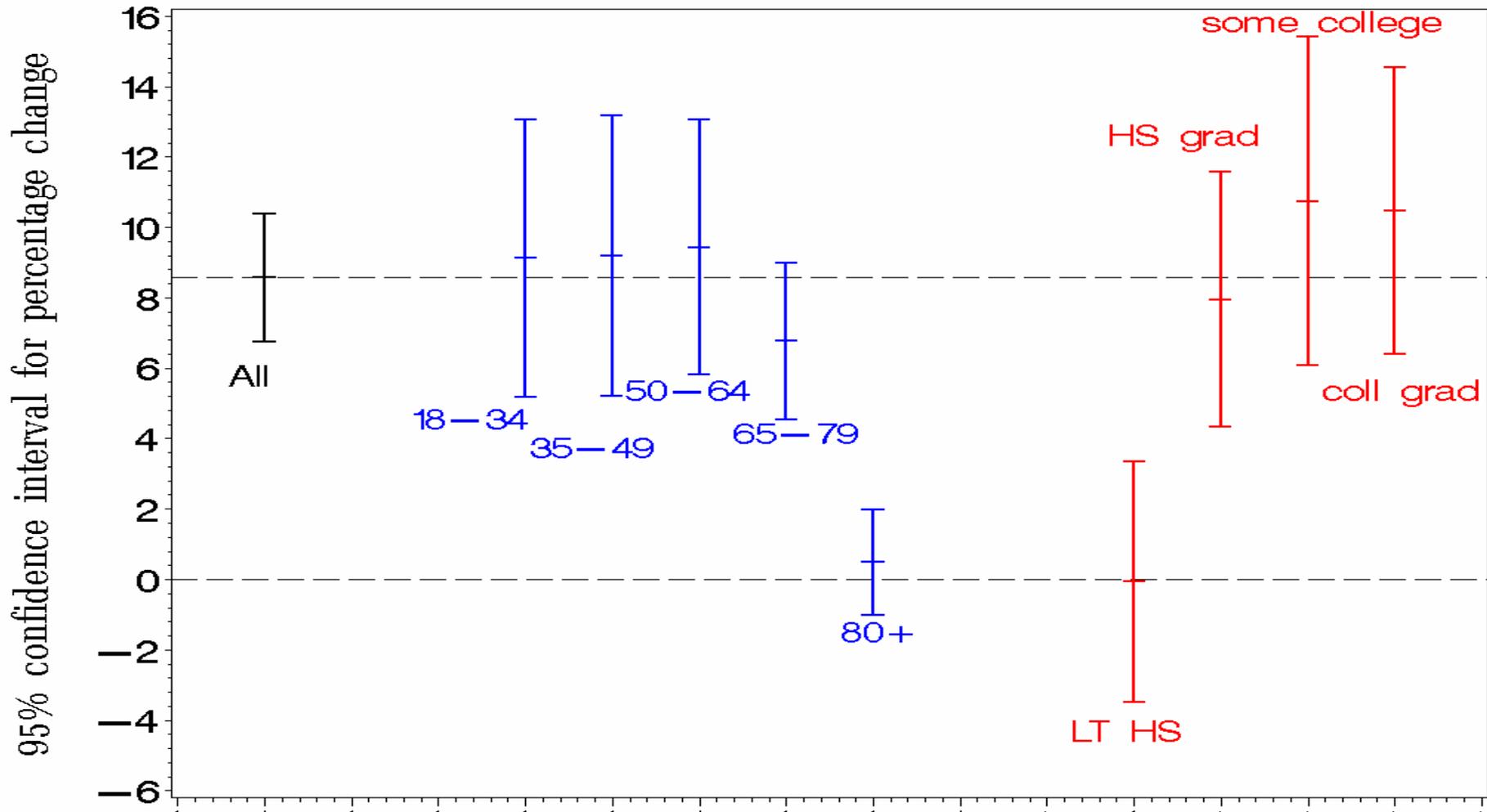


# Results



	2003		2005				
Variables	Weighted %	SE	Weighted %	SE	Diff	SE Diff	z
<b>All</b>	19.67	0.59	28.25	0.7	<b>8.58</b>	<b>0.92</b>	<b>9.37</b>
<b>Age</b>							
18-34	23.45	1.27	32.58	1.53	<b>9.13</b>	<b>1.99</b>	<b>4.59</b>
35-49	23.32	1.21	32.51	1.60	<b>9.19</b>	<b>2.01</b>	<b>4.58</b>
50-64	20.58	1.23	30.01	1.35	<b>9.43</b>	<b>1.83</b>	<b>5.16</b>
65-79	5.20	0.61	11.98	0.94	<b>6.78</b>	<b>1.12</b>	<b>6.05</b>
80+	0.63	0.51	1.12	0.56	<b>0.49</b>	<b>0.76</b>	<b>0.65</b>
<b>Education level</b>							
Less than high school	6.47	1.36	6.40	1.06	<b>-0.07</b>	<b>1.72</b>	<b>-0.04</b>
High school graduate	11.99	0.93	19.94	1.57	<b>7.95</b>	<b>1.82</b>	<b>4.36</b>
Some college	23.93	1.33	34.67	1.94	<b>10.74</b>	<b>2.35</b>	<b>4.57</b>
College graduate or more	36.00	1.27	46.47	1.61	<b>10.47</b>	<b>2.05</b>	<b>5.11</b>
<b>Race</b>							
NH White	23.11	0.75	33.26	1.06	<b>10.15</b>	<b>1.30</b>	<b>7.82</b>
NH Black	13.61	1.70	23.25	3.37	<b>9.64</b>	<b>3.77</b>	<b>2.55</b>
Hispanic	7.15	1.00	11.21	2.01	<b>4.06</b>	<b>2.25</b>	<b>1.81</b>
NH Other	22.12	2.37	28.22	3.67	<b>6.10</b>	<b>4.37</b>	<b>1.40</b>
<b>Gender</b>							
Male	16.70	0.81	25.27	1.44	<b>8.57</b>	<b>1.65</b>	<b>5.19</b>
Female	22.42	0.87	31.01	0.89	<b>8.59</b>	<b>1.24</b>	<b>6.90</b>
<b>Income</b>							
< \$25k	10.06	0.92	18.00	1.52	<b>7.94</b>	<b>1.78</b>	<b>4.47</b>
\$25k - <\$50k	16.59	1.21	25.55	1.85	<b>8.96</b>	<b>2.21</b>	<b>4.05</b>
\$50k - <\$75k	27.32	1.63	30.40	1.95	<b>3.08</b>	<b>2.54</b>	<b>1.21</b>
\$75k +	36.32	1.76	44.59	2.12	<b>8.27</b>	<b>2.76</b>	<b>3.00</b>

# Percentage change by age and education





# Analyses Using Combined 2003 and 2005 Data



# Using a Combined Data Set



- ❖ Need to combine both data sets
  - See Appendix A for syntax
  - Make sure variables have same name, formats, and interpretation
- ❖ Need to construct final sample weights and replicate weights
  - 1 final sample weight (for estimated national point estimates)
  - 100 replicate weights (for appropriate variance estimates since each HINTS survey has 50 replicate weights)

# Construction of Statistical Weights for Combined Data File



	<b>Final Sample Weights</b>	<b>Replicate Weights 1-50</b>	<b>Replicate Weights 51-100</b>
HINTS 2003	2003 Final Weight (fwgt)	2003 Replicate Weights (fwgt1-fwgt50)	2003 Final Weight (fwgt)
HINTS 2005	2005 Final Weight (fwgt)	2005 Final Weight (fwgt)	2005 Replicate Weights (fwgt1-fwgt50)
Combined Data	Final Weight (nfwgt)	Final Replicate Weights (nfwgt1-nfwgt50)	Final Replicate Weights (nfwgt51-nfwgt100)



# Construction of Final Sample and Replicate Weights for Combined Data File: SAS Syntax

```
***Set new weight variables for the combined dataset;  
array origwgts[50] fwgt1-fwgt50;  
array newwgts[100] nfwgt1-nfwgt100;  
nfwgt=fwgt;  
do i = 1 to 50;  
if hintsyear=1 then do;***2003;  
    newwgts[i]      = origwgts[i];  
    newwgts[i+50] = fwgt;  
end;  
else if hintsyear=2 then do;***2005;  
    newwgts[i]      = fwgt;  
    newwgts[i+50] = origwgts[i];  
end;  
end;  
drop fwgt--fwgt50 i;  
label nfwgt="Final full-sample weight";  
attrib nfwgt1-nfwgt100 label="Final sample replicate weight";
```

Note: Information also found in Appendix A

# Design Statements for Combined Data



```
proc procedurename data=combined design=jackknife;  
weight nfwgt;  
jackwgt1-nfwgt100 /adjjack=.98;
```

## Notes:

- 1) nfwgt= Final sample weight for estimated national point estimates
- 2) nfwgt1 to nfwgt100= Replicate weights for variance estimates
- 3) Information also found in Appendix B

# Testing for Differences Using a Combined Dataset



```
proc descript data=combined design=jackknife;
weight nfwgt;
jackwgt1-nfwgt100 /adjjack=.98;
class hintsyear cd07internetforcancer/nofreq;
var cd07internetforcancer;
catlevel 1;
diffvar hintsyear=(2 1)/name="Change from 2003 to 2005";
print nsum percent sepercent lowpct uppct t_pct
p_pct/style=nchs;
run;
```

Note: Information also found in Appendix C

# Using Combined Dataset to Test for Differences by Demographics: Syntax



```
proc descript data=combined design=jackknife;
weight nfwgt;
jackwghts nfwgt1-nfwgt100 /adjjack=.98;
class hintsyear agegroup education gender income raceethnicity
cd07internetforcancer/nofreq;
var cd07internetforcancer;
catlevel 1;
diffvar hintsyear=(2 1)/name="Change from 2003 to 2005 By
Demographic Subgroup";
tables agegroup education gender income raceethnicity;
print nsum percent sepercent lowpct uppct t_pct p_pct/style=nchs;
run;
```

Note: 1) Outcome variable is a dummy coded (0,1); 2)  
Information also found in Appendix C

# Estimating Change While Controlling for Covariates With Combined Data



- ❖ Can only be done with combined data
- ❖ Across all subjects
- ❖ By demographic subgroup
  - Demonstrate using education
- ❖ Use a regression approach:
  - Multiple regression for continuous outcomes
  - Logistic regression for dichotomous outcomes
  - Use survey year as a covariate
  - See Appendix D

# Testing for Changes Across Years Controlling for Covariates-Syntax



```
proc rlogist data=combined design=jackknife;  
weight nfwgt;  
jackwghts nfwgt1-nfwgt100 /adjjack=.98;  
class hintsyear education agegroup gender/nofreq;  
model cd07internetforcancer=hintsyear education agegroup  
gender;  
reflev hintsyear=1;  
run;
```

1. Controlling for education, agegroup, and gender;
2. Assumes the effect of the covariates is the same across years
3. Can also test for differential change of covariates by including an interaction term in the model statement (e.g., hintsyear\*education)
4. See Appendix D

# Testing for Changes by Demographic Subgroup Controlling for Covariates



Test for differences across levels of **education**. Start with lowest level (Less Than High School) controlling for age and gender (note SUBPOPN statement)

```
proc rlogist data=combined design=jackknife;  
weight nfwgt;  
jackwghts nfwgt1-nfwgt100 /adjjack=.98;  
subpopn education=1; /*Less than High School*/  
class hintsyear agegroup gender/nofreq;  
model cd07internetforcancer=hintsyear agegroup gender;  
reflev hintsyear=1;  
run;
```

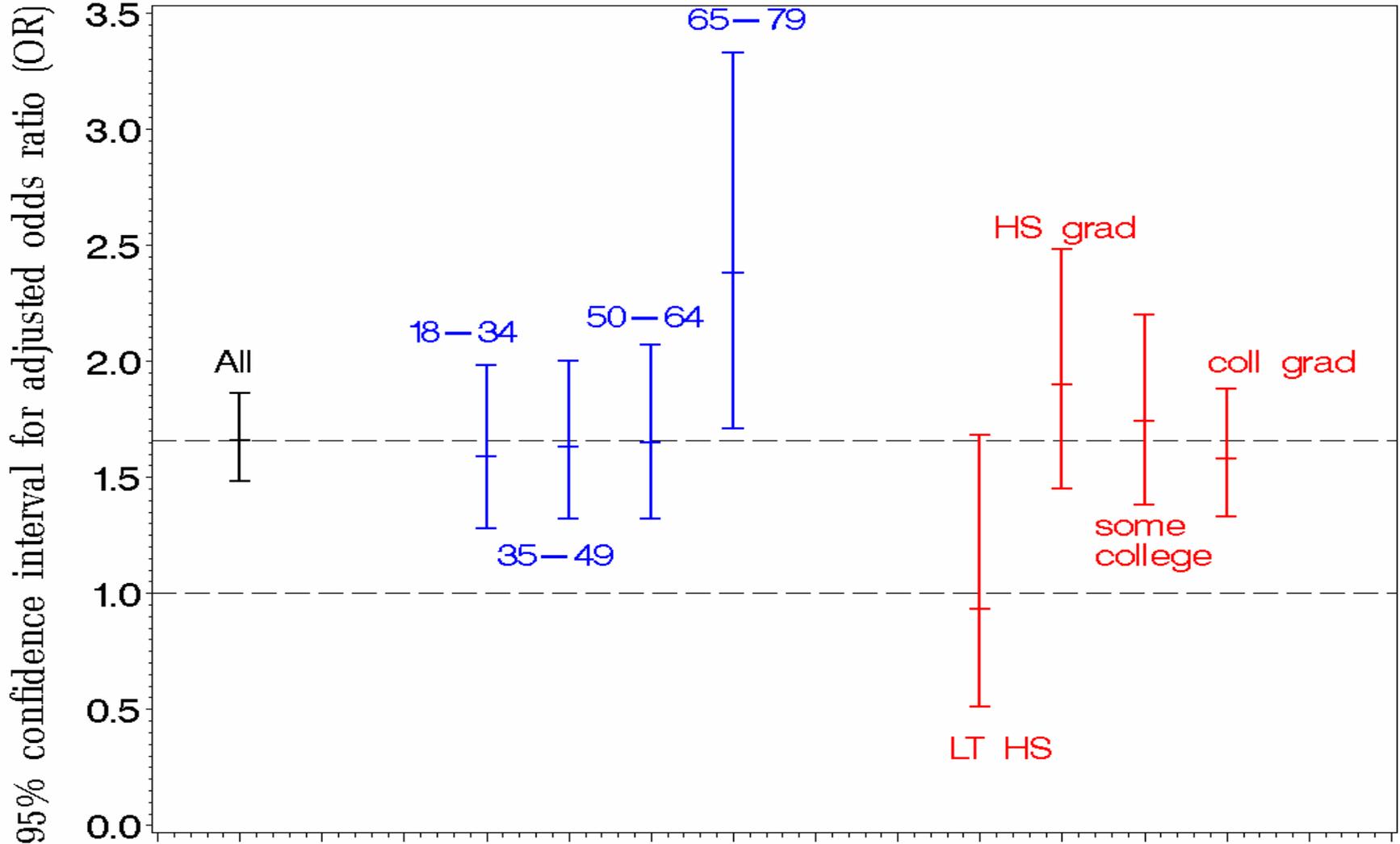
Note: Can also test three other levels of education substituting remaining values; See Appendix D

# Testing for Changes by Levels of Education: Results



	<b>Odds Ratio</b>	<b>Lower Bound 95% CI</b>	<b>Upper Bound 95% CI</b>
<b>Less Than High School</b>			
2003	1.00	1.00	1.00
2005	0.93	0.51	1.68
<b>High School Graduate</b>			
2003	1.00	1.00	1.00
2005	1.90	1.45	2.48
<b>Some College</b>			
2003	1.00	1.00	1.00
2005	1.74	1.38	2.20
<b>College Graduate or More</b>			
2003	1.00	1.00	1.00
2005	1.58	1.33	1.88

# Adjusted Odds Ratios by age and education



# Estimating Weighted Mean Using Combined Data



- ❖ Can be assessed across respondents and by subgroups
- ❖ Will calculate weighted mean across combined data
  - Weights each year proportional to its estimated population

# Calculate Mean % of Respondents Using Combined Data



```
proc descript data=combined design=jackknife;  
weight nfwgt;  
jackwghts nfwgt1-nfwgt100 /adjjack=.98;  
var cd07internetforcancer;  
catlevel 1;  
print nsum percent lowpct uppct/style=nchs;  
run;
```

Note: Will give sample size, mean %, lower and upper 95% CI; See Appendix E

# Calculate Mean % of Respondents by Subgroups



```
proc descript data=combined design=jackknife;  
weight nfwgt;  
  
jackwghts nfwgt1-nfwgt100 /adjjack=.98;  
  
class hintsyear agegroup education gender income  
raceethnicity cd07internetforcancer/nofreq;  
  
var cd07internetforcancer;  
  
catlevel 1;  
  
tables (agegroup education gender income  
raceethnicity)*hintsyear;  
  
print nsum percent lowpct uppct/style=nchs;  
  
run;
```

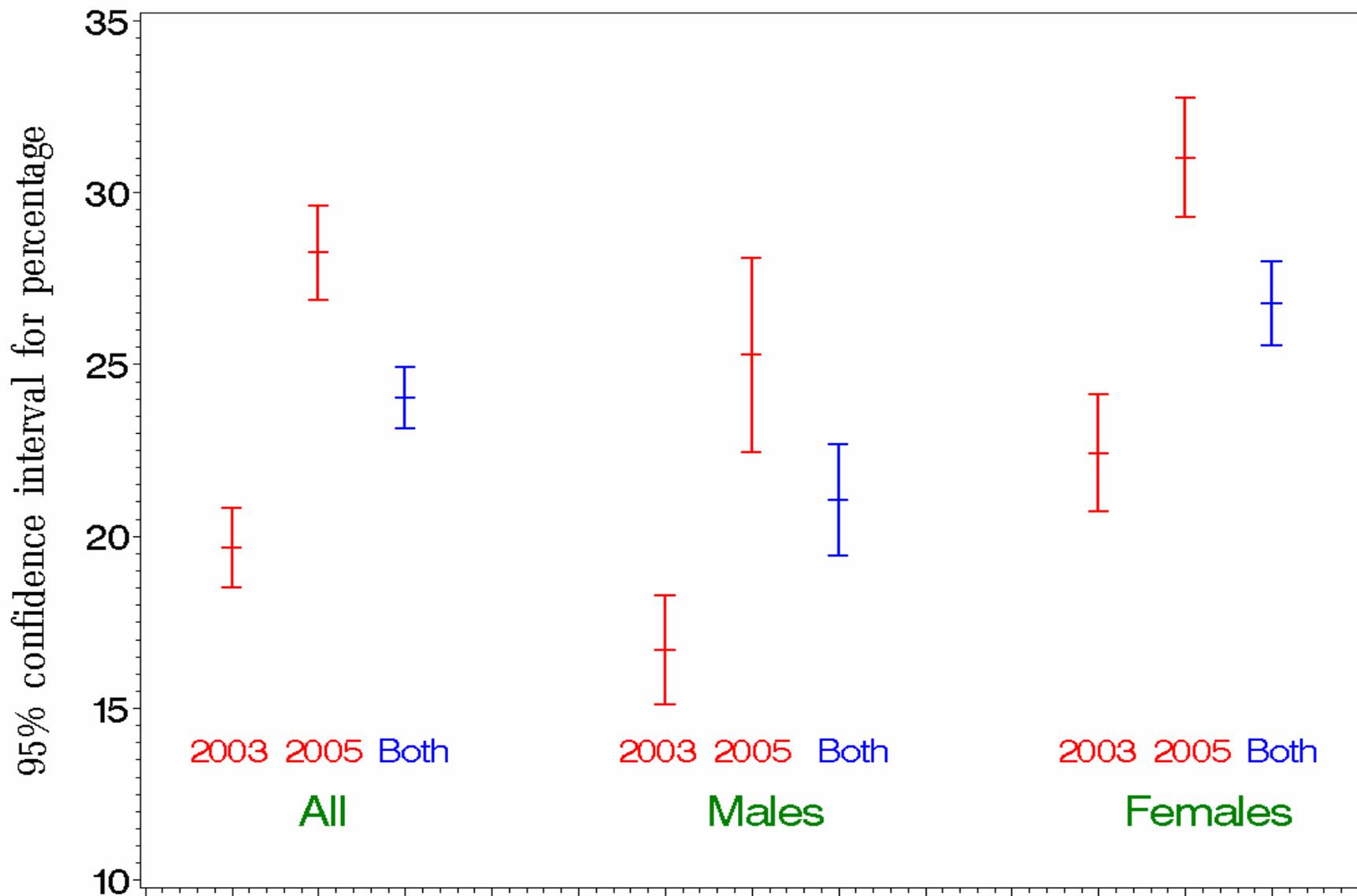
Note: Will give sample size, mean %, lower and upper 95% CI; See Appendix E

# Means From Combined Data



<b>Variables</b>	<b>Weighted Mean</b>	<b>SE</b>
<b>All</b>	<b>24.02</b>	<b>0.46</b>
<b>Age</b>		
18-34	28.08	1.00
35-49	27.90	1.00
50-64	25.50	0.92
65-79	8.61	0.57
80+	0.88	0.38
<b>Education level</b>		
Less than high school	6.44	0.88
High school graduate	15.88	0.90
Some college	29.85	1.22
College graduate or more	41.19	1.02
<b>Race</b>		
NH White	28.18	0.64
NH Black	18.39	1.87
Hispanic	9.31	1.17
NH Other	25.48	2.28
<b>Gender</b>		
Male	21.05	0.83
Female	26.77	0.62
<b>Income</b>		
< \$25k	13.72	0.88
\$25k - <\$50k	20.50	1.07
\$50k - <\$75k	28.99	1.29
\$75k +	40.77	1.38

# Percentage estimates by gender and HINTS data source



# Summary



- ❖ Creating the combined data set is hardest part, but gives more versatility than using separate data sets
  - Do not use combined data to get single-year estimates unless you adjust denominator df
- ❖ If using combined data, make sure variable names, formats, and interpretations are equivalent across years
- ❖ Once you have combined data, analyses are similar to those done with a single data set

# What's Next?



- ❖ Posting combined data and syntax on HINTS Web site in future
- ❖ What would be most useful to you?



## ❖ Contact information:

- Rick Moser: [moserr@mail.nih.gov](mailto:moserr@mail.nih.gov)
- Bill Davis: [davisbi@mail.nih.gov](mailto:davisbi@mail.nih.gov)

