

2021 TUS-CPS Data User Webinar Series

How to Conduct Analyses Using the TUS-CPS 1992-2019 Harmonized Dataset

Carolyn Reyes-Guzman, PhD, MPH

NCI, Tobacco Control Research Branch

Webinar Agenda

Questions?

Please type into the chat box and moderators will ask them at the end of the presentation

- 1. Background on data harmonization*
- 2. Self-response and replicate weights for variance estimation*
- 3. How to run analyses using harmonized data*
- 4. Applied examples*
 - Disparities in menthol cigarette use*
 - Tobacco use surveillance among small populations*

Disclaimer

The views and opinions expressed are the presenters' and do not necessarily represent the views, official policy, or position of the U.S. Government, U.S. Department of Health and Human Services or any of its affiliated institutions or agencies.

Background on Data Harmonization

TUS-CPS Background

- NCI and FDA co-sponsor this survey of tobacco use administered as part of the US Census Bureau's Current Population Survey (CPS) every 3-4 years since 1992-93
- Most recent publicly released data are from the 2018-2019 wave; 2022-2023 wave currently being planned.
- Previous waves are 1992-1993, 1995-1996, 1998-1999, 2000, 2001-2002, 2003, 2006-2007, 2010-2011, and 2014-2015
- CDC was a co-sponsor with NCI from 2001-2002 through 2006-2007

What is TUS useful for?

- TUS data can be used by researchers to:
 - Monitor tobacco control progress and assess long-term cross-sectional population trends;
 - Track tobacco health disparities;
 - Evaluate tobacco control programs; and
 - Examine national, state and county*-level data.
- Data from the TUS can be linked to other CPS supplements, including detailed labor force, occupational, and economic and health insurance data (ASEC or March Supplement) that can be related to tobacco use, as well as to mortality data from the National Death Index.

*where available in these smaller geographical unit areas

Effort behind data harmonization

- Researchers who previously wanted to examine long-term trends had to track variable names for a specific question of interest over several survey waves
- Harmonized file:
 - one variable name across all waves, using “flag” variable to track survey year
 - Adult self-respondents, no proxy responses
 - variables with >1 wave of data

Inclusion/exclusion criteria for harmonization

- Variables to harmonize as-is:
 - remained consistent over time, harmonized in their current state
- Variables to harmonize after adjustments:
 - slight wording changes over time or those for which universe had slight variations across survey waves were first adjusted
- Variables to drop:
 - items with significant wording, structure, or universe variations
- Data dictionary + crosswalk available on TUS-CPS site
- Harmonized file user guide to be released Fall 2021 (with examples of SAS/SUDAAN code)

Harmonized Variables

- Core variables from CPS (sociodemographic, geographic, occupational, economic)
- Cigarette use – including menthol cigarettes since 2003
- Workplace and home smoking restrictions
- Attitudes toward smoke-free policies in indoor work areas/public places
- Smoke-free attitudes for multi-unit housing*
- Advice by physician/dentist
- Health perceptions/beliefs (harm reduction)
- Smoking history, cessation, former smokers
- Use of other tobacco products: cigar, pipe, smokeless tobacco use, e-cigarettes
- Attempts to quit smoking by switching to e-cigarettes*
- Flavored tobacco products

*Only include 2014-2015 and 2018-2019 TUS waves

Self-Response and Replicate Weights

Dealing With TUS-CPS Weights

- The harmonized dataset includes a full-sample weight for self respondents (SmplWgt)
- “Full sample weights are created to compensate for differential selection probabilities, nonresponse, and under-coverage of the target population of U.S. adults”¹
- But there is a second type of weights available for the TUS-CPS too –
 - “Replicate weights, which can accommodate various types of statistical analyses, are created to more accurately estimate standard errors by accounting for the complex survey design”¹
- Depending on the goals of your analysis, you may need to use replicate weights

[2018-2019 Tobacco Use Supplement to the Current Population Survey \(TUS-CPS\) User Guide for Conducting Weighted Analyses](#)

TUS-CPS Replicate Weights

- The number of replicate weights used for TUS-CPS has changed over time, and, accordingly, replicate weights are provided in three separate files:
 - 1992-93: 48 replicate weights
 - 1995-96 through 2003: 80 replicate weights
 - 2006-07 through 2018-19: 160 replicate weights

Linking TUS-CPS Harmonized File to Replicate Weights

- To run analyses, first need to merge TUS respondents with their corresponding replicate weights by ID
- “RecordID” variable in harmonized file is unique within survey year and survey month
- Same RecordID variable was created for replicate weight files
- Files merged by SurYear, SurMonth and RecordID
- SAS code to merge files available on TUS-CPS site

Linking TUS-CPS Harmonized File to Replicate Weights (cont.)

```
Data Rep9293;  
  Infile Rep9293;  
  Input @001 SurYear 4.  
        @005 SurMonth 2.  
        @007 RecordID $Char22.  
        @029 SmplWgt 11.4  
        @040 (RepWt001-RepWt048) (11.4);
```

Run;

```
Data Rep9503;  
  Infile Rep9503;  
  Input @001 SurYear 4.  
        @005 SurMonth 2.  
        @007 RecordID $Char22.  
        @029 SmplWgt 11.4  
        @040 (RepWt001-RepWt080) (11.4);
```

Run;

```
Data Rep0619;  
  Infile Rep0619;  
  Input @001 SurYear 4.  
        @005 SurMonth 2.  
        @007 RecordID $Char22.  
        @029 SmplWgt 11.4  
        @040 (RepWt001-RepWt160) (11.4);
```

Run;

```
Data Repls;  
  Set Rep9293  
      Rep9503  
      Rep0619;
```

Run;

```
Proc Sort Data=Repls;  
  By SurYear SurMonth RecordID;  
Run;
```

```
Proc Sort Data=Harmon;  
  By SurYear SurMonth RecordID;  
Run;
```

```
Data Harmon;  
  Merge Harmon (In=In1)  
           Repls (In=In2);  
  By SurYear SurMonth RecordID;  
  Harm_File=In1;  
  Repls_File=In2;
```

Run;

Using TUS-CPS Replicate Weights

- Fay's method (a variation of balanced repeated replication, or BRR) when conducting variance estimation for the TUS-CPS with replicate weights
 - For frequency tables using one wave of data:
 - In SAS: use **Proc SurveyFreq** and **VarMethod=BRR (Fay=0.5)**
 - In SUDAAN: use **Proc Crosstab** and **ADJFay=4**;
 - For multiple waves of data requiring an expansion of replicate weights (e.g., see slide 12), use **Fay=0.75** in SAS; **ADJFay=16** in SUDAAN
 - For multiple waves of data retaining the same number of replicate weights, use **Fay=0.5** in SAS; **ADJFay=4** in SUDAAN

Using Weights in Harmonized Data Analyses

Example: using 1 wave of data

- *Calculate mean cigarettes per day among overall current smokers during the 2014-2015 wave (self-respondents only)*
- From the TUS-CPS website, download in the ZIP file containing the 1992-2019 data (.dat file), a SAS program to read in the data, a SAS program with formats for harmonized variables, three files of replicate weights (.dat files), and three corresponding SAS programs to read in the replicate weight data
- Open the SAS program to read in the self-response replicate weight files and merge them with the survey data (located in the Harmonized TUS-CPS dataset .zip file available on the TUS-CPS website). Modify the Filename, Libname and Include statements to match location and names of the data files and the format program, and run the program.

Example: using 1 wave of data (cont.-1)

- From the dataset, select the TUS-CPS self-respondents who completed the 2014-2015 survey wave.

```
Data harmon1415;  
  Set harmon;  
  If SURWAVE=9; /*2014-2015 Survey Wave*/  
Run;
```

- Divide person-weights and replicate weights by 3 (the number of months of data being combined for analysis).

```
/* RepWt001-RepWt160 are the replicate weights */  
Data harmon1415;  
  Set harmon1415;  
  Array Wgts(160) RepWt001-RepWt160;  
  Do I = 1 to 160;  
    Wgts(I)=Wgts(I)/3;  
  End;  
  NSmplWgt=SmplWgt/3;  
Run;
```

Example: using 1 wave of data (cont.-2)

- Recode variables needed for analysis: current cigarette smoking status (yes/no, called CURRSMK), cigarettes smoked per day (continuous, among current smokers, called CIGPD)

```
Data harmon1415;
  Set harmon1415;

  /* CurrSmk: Current Cigarette Smoking Status */
  If CigStat in (2,3) Then CurrSmk=1; /* Current Cigarette Smoker */
  Else If CigStat in (1,4) Then CurrSmk=0; /* Non-Smoker */
  Else CurrSmk=.;

  /* CigPD: Number of Cigarettes Per Day */
  If CigStat=2 & (0<=CPDD<=99) Then CigPD=CPDD; /* Daily Smokers */
  Else If CigStat=3 & (0<=CPDS<=30) Then CigPD=CPDS; /* Non-Daily Smokers */
  Else CigPD=.;

  Label CurrSmk = "Current Cigarette Smoking Status"
  CigPD = "Number of Cigarettes Per Day";

  Keep RecordID CurrSmk CigPD NSmplWgt RepWt001-RepWt160;
  Format CurrSmk CurrSmkF.;
Run;
```

Example: using 1 wave of data (cont.-3)

- The following SAS code will output weighted means and standard errors for cigarettes smoked per day among current cigarette smoking (CURRSMK=1) adult self-respondents.

```
Proc SurveyMeans Data=harmon1415 VarMethod=BRR (Fay=0.5);  
  Var CigPD;  
  Domain CurrSmk;  
  Weight NSmplWgt;  
  RepWeights RepWt001-RepWt160;  
Run;
```

Example: pooling multiple waves of data

Calculate the prevalence of overall current smoking during the 2003 and 2006-2007 waves (self-response)

- From the full, harmonized dataset (survey data and replicate weights), select the TUS-CPS self-respondents who completed the 2003 or 2006-2007 survey waves.

```
Data harmon030607;  
  Set harmon;  
  If SURWAVE in (6,7); /* 2003, 2006-2007 Survey Waves */  
Run;
```

Example: pooling multiple waves of data (cont.-1)

- Both the person-weights and replicate weights must be divided by 6 (the number of months of data being combined for analysis; 3 months from each survey wave), so that the weights total to the average size of the U.S. population during the 2003 and 2006-2007 time periods.

```
Data harmon030607(Drop=I J RepWt001-RepWt160 SmplWgt);
Set harmon030607;
Array OldR(160) RepWt001-RepWt160;
Array NewR(240) NWgt001-NWgt240;
NSmplWgt=SmplWgt/6;

/* The 2003 dataset has 80 replicate weights and the 2006-07 has 160, so it is necessary to expand or extend the number of replicate
weights to 240 so that all survey months have an equal number of replicate weights. */

If SurWave=6 Then Do;
  Do I = 1 to 80;
    /* The following mathematical adjustment factor extends the number of replicate weights from 80 to 240 so that they are equalized
with the other survey waves. */
    NewR(I)=(1/6) * (SmplWgt+(.866025*(OldR(I)-SmplWgt))); /* .866025 = 1/2 x (Sqrt(240/80) */
  End;
  /* For 2003, 160 replicate weight are generated beyond the first 80 by dividing the main sample weight by the number of surveys. */
  Do I = 81 to 240;
    NewR(I)=SmplWgt/6;
  End;
End;
/* Need to do the reverse for the 2006-07 survey data. */
Else Do;
  Do I = 1 to 80;
    NewR(I)=SmplWgt/6;
  End;
  Do I = 81 to 240;
    J = I - 80;
    NewR(I)=(1/6) * (SmplWgt+(.612372*(OldR(J)-SmplWgt))); /* .612372 = 1/2 x (Sqrt(240/160) */
  End;
End;
Run;
```

Example: pooling multiple waves of data (cont.-2)

- Recode current cigarette smoking status (yes/no, called CURRSMK)

```
Data harmon030607;  
  Set harmon030607;  
  
  /* CurrSmk: Current Cigarette Smoking Status */  
    If CigStat in (2,3) Then CurrSmk=1; /* Current Cigarette Smoker */  
  Else If CigStat in (1,4) Then CurrSmk=0; /* Non-Smoker */  
  Else CurrSmk=.;  
  Label CurrSmk = "Current Cigarette Smoking Status";  
  
Keep RecordID CurrSmk NSmplWgt NWgt001-NWgt240;  
Format CurrSmk CurrSmkF.;  
Run;
```

- Output weighted prevalence estimates for current cigarette smoking (CURRSMK) among adult self-respondents.

```
Proc SurveyFreq Data=harmon030607 VarMethod=BRR (Fay=0.75);  
  Tables CurrSmk /CL;  
  Weight NSmplWgt;  
  Repweights NWgt001-NWgt240;  
Run;
```

Example: pooling the entire harmonized dataset

- *Estimate the adjusted odds ratio for the association between demographic factors and smoking status among all TUS-CPS adult self-respondents from 1992-2019*
- Both the person-weights and replicate weights must be divided by 29 (the number of months of data being combined for analysis; 3 months from each survey wave), so that the weights total to the average size of the U.S. population during the 1992-2019.

Example: pooling the entire harmonized dataset (cont.-1)

```
Data harmon_recode(Drop=I J RepWt001-RepWt160 SmplWgt);
Set harmon_recode;
Array OldR(160) RepWt001-RepWt160;
Array NewR(288) NWgt001-NWgt288;
NSmplWgt=SmplWgt/29;
```

/* The 1992-93 data has 48 replicate weights, the 1995-96 through 2003 data has 80 RPs and the 2006-07 onwards has 160 RPs, so it is necessary to expand or extend the number of replicate weights to 288 so that all survey months have an equal number of replicate weights. */

```
If SurWave=1 Then Do;
  Do I = 1 to 48;
    /* The following mathematical adjustment
factor extends the number of replicate weights from 48 to
288 so that they are equalized with the other survey
waves. */
    NewR(I)=(1/29) * (SmplWgt+(1.224745*(OldR(I)-
SmplWgt))); /* 1.224745 = 1/2 x (sqrt(288/48)) */
    End;
    /* For 1992, 48 replicate weight are
generated beyond the first 48 by dividing the main sample
weight by the number of surveys. */
    Do I = 49 to 288;
      NewR(I)=SmplWgt/29;
    End;
  End;
```

```
/* 1995-2003 */
Else If SurWave in (2,3,4,5,6) then do;
  Do I = 1 to 48;
    NewR(I) = SmplWgt/29;
  End;
  Do I = 49 to 128;
    J = I-48;
    NewR(I)=(1/29) *
(SmplWgt+(0.948683*(OldR(J)-SmplWgt)));
/* 0.948683 = 1/2 x (Sqrt(288/80)) */
  End;
  Do I = 129 to 288;
    NewR(I) = SmplWgt/29;
  End;
End;
```

```
Else If SurWave in (7,8,9,10) then do;
  Do I = 1 to 128;
    NewR(I)=SmplWgt/29;
  End;
  Do I = 129 to 288;
    J = I - 128;
    NewR(I)=(1/29) *
(SmplWgt+(0.670820*(OldR(J)-SmplWgt)));
/* 0.670820 = 1/2 x (Sqrt(288/160)) */
  End;
End;
```

```
Run;
```

Acknowledgements



- Margaret Mayer (NCI)
- Kristen Mangold (NCI)
- Todd Gibson (IMS/NCI Contract)
- Nalini Corcy (BLH/NCI Contract)
- Dana Chomenko (BLH/NCI Contract)
- TUS-CPS Team (NCI DCCPS, NCI DCEG, and NIH NIMHD)
- FDA CTP

TUS-CPS Website - <https://cancercontrol.cancer.gov/brp/tcrb/tus-cps>