**Tobacco Use Supplement Current Population Survey Webinar Transcript**

MR. WILLIS: … Supplement to the Current Population Survey. You are in the right place. So what we're going to do today is it's a variation of a theme on the past. We've done this live, meaning not virtual but in the same room. We've had two past events where we've talked about new developments to the Tobacco Use Supplement. Especially given that travel is not quite as easy as it once was, we're trying a webinar format.

By the way, I'm Gordon Willis. I'm a survey methodologist at the National Cancer Institute. I'd like to start by acknowledging the brains behind the operation; that's Anne Hartman of NCI. I'm just the mouthpiece, at least to start, and my job is to introduce us and take no more than ten minutes. So I think I should get going here. We are scheduled for two hours in total, for those who are up for staying with us.

As background, Anne will go into this further, but I thought I would just mention that the TUSCPS is, effectively, a really good, interesting survey, a rich resource for tobacco surveillance and research. So we have a very large sample. We have good population representation if you look at response rate as a measure of that. We can track some trends over time and even include new trends where we have new questionnaire content.

Because this is such a rich survey, we feel it's part of our mission to make this available to the extramural research community. We realize, however, there are a couple of challenges to that. First is making people aware of new TUSCPS developments, because there are some rather complicated linkages to other data. There are some potential creative uses, but you have to know how to make use of various data sets and bring them together and so on and so forth. So we're going to talk about that somewhat.

The other common challenge or problem or barrier with federal surveys is making them easily analyzable. That's not always a trivial factor, especially given the complex sampling design that's used.

So the webinar objectives are to address those two issues—communicating what we're doing and making data easily available.

This is our agenda for today, and a few of our rules. We'll have four speakers taking, in total, hopefully an hour and 20 minutes and no more. First up, we'll have Anne Hartman of the Cancer Institute talking about TUS linkages and unique aspects—again, what's new and interesting. Then we have Sean Altekruse, a statistician at NCI, talking about the

relationship between the TUS and the National Longitudinal Mortality Study outcome data. After that, we have Benmei Liu from NCI, who is calling in. She couldn't be here because she's just brought her second child onto this Earth, so we're happy that she's still dedicated enough that she wants to delivery her talk with a very new newborn. And, finally, batting cleanup here is Todd Gibson, who will talk about the nuts and bolts of tips and tricks for handling TUSCPS data for those who are interested in getting down in the trenches.

During the presentations, these four talks I just mentioned, we will be digitally audio recording everything. Listeners will be put on mute, not because we don't like to hear from you—don't take offense. It's because, otherwise, there are too many, in my experience, cats meowing in the background and things of that nature. So we're going to try to make it quiet except for the speaker for everyone. However, we do want feedback and questions, so what we're going to try is to encourage listeners to ask questions in the chat area to the right of the screen, which basically you're doing already in telling us whether you can hear me or not. In your question, please indicate which speaker you're addressing, or if you're asking a general question, that's okay, too. But if you'd like it targeted to a particular speaker, let us know that. You can identify yourself to the degree that you'd like to. If you just want to use a screen name, that's fine, too, and we'll use that—with one exception. If you call yourself Valdimore, we'll have to refer to you as "he or she whose name cannot be spoken," but, otherwise, we'll use what you tell us. NCI staff will be collecting these questions for the follow-up discussion, Part II section.

So, Part II is our 40 minutes—hopefully, there will be enough time here to answer critical questions. And there will be some questions. We will have to be—we'll have to filter through these. We'll address questions as we can as they're addressed to individual speakers. These speakers will try to answer your questions and, if you'd like, as the question answerer, you can type in a follow-up, either a thank you or a complaint or whatever—anything you want to follow up with as your question is being answered or immediately after.

We'll use discretion, again, to try to get through in 40 minutes. With five minutes to go, though, we'll wrap up.

So I believe that's what I was supposed to say. And that brings up Anne Hartman, our first substantive speaker, if I can call her that. So I'm going to move out of the way and let Anne take over control.

**TUS LINKAGES AND UNIQUE ASPECTS—ANNE HARTMAN**

MS. HARTMAN: Thanks, Gordon, and, Gordon, you're not just the mouthpiece. You're a very important part of the team. Without your [inaudible] guidance and cognitive skills, the TUS questionnaire development would lead to a very insufficient survey. So thank you.

MR. WILLIS: Thank you, Anne. Don't use up your time.

MS. HARTMAN: Oh, right! That's right, you did say that. Okay, thank you. And I want to thank the audience, too. And according to the registration, we seem to have a very diverse group, so I hope we can fulfill the promise we have—government colleagues and colleagues from academia and public health and also even some industry representatives.

Okay, so TUS linkages and unique aspects—okay. Let's see, okay, just some background. The NCI sponsors the Tobacco Use Supplement to the Census Bureau's, the Bureau of Labor Statistics Current Population Survey. So I'd also like to thank them for tolerating us for over 20 years. And TUS has been administered from 1992 to 2011, roughly about every three years.

Data are available for public use, and some of the areas that these data can be used in are to monitor tobacco control progress, conduct tobacco-related research, evaluate tobacco control programs, and examine health disparities. It's a key source of state, some local, as well as national-level data on cigarette and other tobacco use, including emerging products, related social norms and attitudes, and intervention and policy.

Okay, it's a large, nationally representative address-based household sample and the sample comes from the civilian, noninstitutionalized population ages 15+. And since 2007, it's been 18+. It yields about 240,000 individual respondents for each survey cycle, mostly self-reports yielding about 180,000 self-respondents. Sixty-four percent of the interviews are conducted by telephone, and 36%, in person.

Current and past use of tobacco—some of the topics, actually, are current and past use of tobacco; cigarettes; cigars, all types—and, most recently in our 2010-11 survey, we asked some questions about use of flavored cigars; pipes—most recently, we added hooka separately from regular pipes; smokeless tobacco—we've asked about snoose, chew, dip—most recently, we've asked a combined question on smokeless that includes all of these and snoose. Then, we introduced in 2003 an Emerging Products section, some harm-reduction-type products, and, most recently, Dissolvables and E-cigarettes. Menthol cigarette use is asked since 2003, and we've

asked questions on workplace and home smoke-free policies since the beginning in 1992. We also ask about attitudes toward smoke-free policies in several public places. And, most recently, we've added cars and casinos.

In 2003 forward, we added questions about cost of cigarettes and purchase location. We've always asked about physician and dentist advice to quit smoking, interest and attempts to quit, and in 2003 and 2010-11, we asked about treatment methods to curb tobacco use.

Some unique aspects that I'd like to really highlight here—because the TUS is a supplement to the Current Population Survey, which many of you may know but some of you may not be aware of—the Current Population Survey is the survey that the employment figures that you hear about every month are based on. So it gives us the ability to delve into economic and occupation patterns.

Two recent examples—one was by the Congressional Budget Office in the June 2012 report on raising the cigarette excise tax—the impact of doing this. And another example is an examination of the mortality cost to smokers.

A second unique aspect—because of the large sample size, we're able to examine tobacco-related health disparities and get fairly narrow. And one tool that I encourage people to look at in conjunction with the TUS and other national surveys is the NCI Health Disparities Calculator. And the website is here and you can learn more about that if you go to that website.

We also—another unique aspect is that the data are suitable for longitudinal analysis, and that's because the CPS itself, the basic CPS, is a panel survey. They go to households four months in a row, and then the next year the same four months. So you have the unique opportunity to follow up people. And we actually took advantage of this when we did our 2001-2001 series on top of our 2003 series; we actually had a February 2002, February 2003 overlap, and there is information about that on our website. It's given here.

And, also, most exciting is we recently, as part of the 2010-2011 series, we have followed the May 2010 respondents again in May 2011. So that's like a half month overlap. And those data will be available by the end of this year. And what's nice about that is that the Census is creating weights and everything will be merged, so you'll be able to use that file directly.

Okay, some other unique aspects—the TUS allows detailed data to be used in modeling, either indirectly linking to other ecological summary data; for example, by state or

media markets. And one really recent example in 2010 was Sherry Emory and her colleagues looked at the effects of smoking-related television advertising on adult smoking and intentions to quit, linking television ratings for top U.S. media markets to the TUSCPS data. So not only can you get state information and some substate information, but with that unique aspect, you can combine across states, in this case the media markets.

And then there are special direct linkages to disease outcomes data with the National Longitudinal Mortality Study and, actually, my colleague Sean Altekruse will be discussing that in more detail. There is also direct linkage to other CPS supplements. Just like we periodically supplement the CPS with the Tobacco Use Survey, others do that as well. And I'll get into a few examples in a moment. And here are some websites that can give you a sense of the supplements to the CPS and even some abstracts.

Okay, I'm just going to briefly mention this National Longitudinal Mortality Study to just say that this gives us the opportunity to link to outcome data, which is really exciting. And that's because the Tobacco Use Supplement can be linked to other data, and at the heart of this National Longitudinal Mortality Study is the March supplement to the CPS, the Annual Social and Economic Supplement. And this project links to National Death Index data, NCI SEER cancer registry data, tobacco use I said, and also some Medicare data. And Sean will be talking more about that shortly.

The other unique linkages to outcome and other CPS supplement data. Some examples are linkage to, as I've just mentioned, the Annual Social and Economic Supplement, and that's the March supplement, and that provides detailed economic, occupational, social, and health insurance data. And an example of the use of these data was to examine the effect of Medicaid coverage on tobacco dependence treatments, on quitting attempts, and intention to quit by Medicaid smokers by Liu in a *Public Health* article in 2010. And I will show you something about that shortly. I also have the website for the March supplement.

Another real exciting aspect is linkage to the American Time Use Survey, ATUS, or sometimes people also call it the TUS. It's sponsored by the Bureau of Labor Statistics and it provides estimates of the amount of time that Americans spend in various primary activities for a given day of the week based on a 24-hour activity diary. The sample is drawn from the CPS samples one month after they have completed their panel participation. This would be useful for tracking the pattern of smoking during a 24-hour period to the extent that smoking is a primary

activity. It would also provide information about patterns of daily activity for smokers, former smokers, and never smokers. One example that I will show in a minute is one by Song in 2012. And, again, here is that website for the ATUS.

The final one I'm going to talk about is linkage to the Voting and Registration Supplement that can provide data about a person's ability to influence policy. A very elegant study was done by Hirsch and her colleagues and was published in the *Economic Inquiry* in July 2004, where she examined the influence of TUS attitudes to smoking in public places, with state smoke-free laws and participation in voting. And there is some information about this Supplement.

Okay, and this is just the illustration of the Medicaid example that I mentioned before looking at covered copayments and various information.

And you can't probably see this very well. This is just more as a reminder for me. But this is from the American Time Use Survey, and this is an illustration of looking at time in different activities among different smokers—current smokers, former smokers, and never smokers—looking at how much time is involved in education, sports, watching television, and also physical activity.

So this expands, then—you see the scope of questions. The TUS really asks about tobacco use in policy-related—but you can look into other areas as well.

Another unique aspect is that in 2003 we started asking questions about price. What I like to think of is the real price of cigarettes. That means after coupons, what's the price the person is facing? While my economist colleagues tell me you really can make use of individual data in an analysis, what's really neat is, because of our sample size, you can aggregate this price information to the state level and then have a measure at the state level. And this—on the left, we have an illustration of actually comparing the tax burden on tobacco price data with our TUS supplement data and differences by states. A lot of this refers to price reduction strategies that people use. And so there is some elegant analysis that you can do with that.

And on the right, I just want to remind you that there's a lot of other demographic data that relate to immigration status as well as health disparities, and some people have looked at this information much more extensively.

And as I mentioned before about the panel aspect, that in February 2002 and February 2003 we had the ability to do a follow-up. And this is an analysis of looking at smokeless tobacco use and cigarette use using the February 2002 and February 2003 by Zoo and colleagues. And it was one of the first analyses that could be conducted in the U.S. trying to see if we get similar patterns to the snoose experience in Sweden. And it seemed quite different.

And then the one on the right is just a reminder, again, that you can look at health disparities and policy. Our colleagues Dennis Trinidad and others have published some work on this.

I'm not going to go into detail here, but some of you may know about the Robert Wood Johnson Foundation Chart Book, where a lot of information by state was put together in 2009. And it has an active database and it gets updated and TUSCPS was featured on this. You can ask questions like testing the hard-core hypothesis about smoking. There are also maps to look at home rules by state. And, again, the impact team website is very useful.

And I'm just going to briefly go, because I'm running short on time, but especially those people who are interested in state data and state public health staff, you may know about the State Cancer Profiles website, but not only can you get information about cancer site—cancer information—but now we also have some risk factor information there, including information on tobacco and some screening. And this is the website. I'm just showing you a little bit of what it looks like. So you get some information. This was on home smoking rules among smokers. And you can also get some really nice, elegant maps. And this was looking at home rules among those smoking—respondents who currently smoke—the home rules in their household.

Okay, it's been used and can be used for modeling, again in conjunction with other data. It was used to model smoking prevalence with menthol [inaudible] by David Levi and colleagues in the *American Journal of Public Health* in 2011.

And I mentioned before the mortality cost of smoking. This is actually a neat example that not only uses the detailed data on occupation and earnings, but also actually merges—and I think, Todd Gibson—he'll be happy that I found an example when he gets to talk. They actually merged the TUS and the CPS that's already in as part of the TUS dealing with other CPS basic data from the panel to get information that's not collected in every month. So that's kind of neat.

And I wanted, then, to tell you that you can find information on our website. And for publications, past workshops, presentations, and general information about the TUS, this is the website. We just, just changed the website so we had to scurry to make some changes in the slides, and I'm sure you may get some emails from me, many of you with our old website. But this is what it looks like. Information over here are on the different things. About halfway through, I want to call your attention to a Publication and Reports database where it's searchable. And, currently, we have almost 200 reports that have utilized the TUS and are available that you can search on title, topic, and author, and also the where the abstracts are available you can also pull those up.

And, finally, we've published—not we, actually personally—not the NCI personally, but the extramural community largely, and that's why this is such a valuable resource, and many peer-reviewed journals and manuscripts and reports—we have just an example of some, including NCI monographs, Surgeon General's reports, a whole slew of tobacco and other addiction journals like *Addiction, American Public Health,* even *Nicotine and Tobacco Control,* and various economic and health policy venues, too.

So, with that, I will say thank you for your attention and I hope this has whet your appetite for use of this data source, which is publicly available.

I will now introduce my colleague, Sean Altekruse, who will be talking next on TUSCPS and the National Longitudinal Mortality Study, using epidemiologists in the Surveillance Research Program at our Division of Cancer Control and Population Sciences at NCI. And he heads the NCI efforts in this multi-Institute/agency collaboration. Okay, Sean, thanks.

**TUSCPS AND THE NATIONAL LONGITUDINAL MORTALITY STUDY—SEAN ALTEKRUSE**

MR. ALTEKRUSE: Thank you, Anne. Thank you for that nice introduction and also for everything you do with TUS. It's amazing! We all really appreciate it.

So I'm going to give a brief, about ten-minute overview of a component of the TUS that you may not be aware of and, basically, it enables us to look at outcome information; namely, cause of death and cancer diagnoses as well as medical treatment, comorbidities, and the cost of care as they relate to the use of tobacco products.

All right, so Anne did a very nice job of giving an overview of the linkage between the National Longitudinal Mortality Survey and TUS. The core component of both is really the Current Population Survey. TUS is a supplement to that. In the National Longitudinal Mortality Study, we link current Current Population Survey data to the National Death Index. And the last ten years have been real banner years for TUS, so we, with the current match, have a considerable amount of data on tobacco use in the National Longitudinal Mortality Study. We also have been making an effort to update the Surveillance, Epidemiology and End Results, or SEER data to the National Longitudinal Mortality Study every two years. So at present, we have data through 2009, and as of April 2014 we will add two additional years. So not listed in the bullets of linkages is the Center for Medicare and Medicaid Services, or the Medicare linkage. But we are in the process of updating that match as well. It will be done almost immediately in fiscal year 2014. We ran a little short on time and we would have to start all over again, so a lot of the ducks are in a row, and immediately in October we will start that. I anticipate that the CMS data will be available in January of 2014.

So the additional part of my presentation is going to be to talk about what exists now, what will exist with the current linkage that is now under way, and what will be available in 2014, and then also to talk briefly about some of the studies that are under way using the tobacco use supplement facet of the National Longitudinal Mortality Study. Finally, I included a slide about how to apply to access these data.

So, the first subject is really the past, present, and future of the NLMS-TUS linkage. At the present time, that linkage is only available through 2002, so the numbers are very small. And then when it's—so this shows—it's really two panels, two tables on one slide. And our entry-level question is: Have you smoked greater than or equal to 100 cigarettes in your lifetime? And with the National Longitudinal Mortality Survey data through 2002, we have about 400,000 responses with tobacco use data. You can see the number of cases alive and the number dead, and by their tobacco use history. And then, also, in descending order of the cause of death, some of the causes of death that are listed. So heart disease is first. You can see that the number, the sheer number of deaths among ever smokers is higher than among nonsmokers. The next number is cancer deaths, and then that's about double; respiratory-disease related deaths also a little more than double. And then "other" and "not otherwise specified" cause of death, which are about equal.

The right-hand panel shows the number of responses at present with the 1985 to 2002 cohort. So there were 411,106 respondents with tobacco use information. However, today, as we speak, the Census Bureau is taking the CPS supplements through 2011 to the match to the Death Index, and that process is going to take about two weeks to accomplish. So in the very, very near future, we estimate that we'll have an additional quarter of a million CPS responses for tobacco use through 2011, giving us over 600,000 total responses. And then, looking out into the future, we think that with the planned TUS surveys and a sufficient follow-up, we will reach a quarter of a million plus tobacco use responses linked to cause of death, cancer incidence, and Medicare data.

Now, the SEER component, the cancer incidence overlap, is only people in the SEER registries, which is a little more than 28 percent of the U.S. population with a cancer diagnosis. So the numbers do start to decrease, but over time, we are accumulating a good number of cancer incidence responses, and that includes the information on demographics as well as the stage at diagnosis, treatments, and survival time. So it's a very rich resource for those of you who are familiar with SEER, and the CMS data give us much more granular information on treatment, essentially from the health insurance claims, as well as comorbidities and billing information on the cost of care. So this resource, the tobacco use linkage to the National Longitudinal Mortality Survey, is really becoming a rich resource.

Now, I mentioned that there are some studies under way using this, and this is a short list of them: Smokeless Tobacco Use in Cancer Incidence; Cancer Incidence among Ever vs Never Users of Tobacco; The Role of Stress among Smokers in Lung Cancer Incidence; A Study of Smoking and Suicide; and then various FDA Center for Tobacco Product Analyses that are currently under way.

So how to apply for data? There are two different routes to access NLMS data, which are sensitive data because they contain personal identifiers. One is to work with an NLMS sponsor to have these analyses done at the Census Bureau. The other is to work through the Census Regional Research Data Centers, and they're located primarily in the two coastal areas at major academic and government institutions. Either route requires you to become a special sworn status employee, or this second route does.

So that really summarizes my talk and I thank you for your time. I've put my email address on there.

The next speaker is Dr. Benmei Liu, who is joining us from Shady Grove today, and she is a colleague in my Branch, a Ph.D. statistician with the Surveillance Research Program.

**STATISTICAL ASPECTS AND BEST PRACTICES FOR ANALYZING TUSCPS DATA—BENMAI LIU**

DR. LIU: Thank you, Sean. It's a very good talk. Can you hear me okay?

UNIDENTIFIED: Yes, we can hear you.

DR. LIU: Okay, good. For the next 15 minutes, I'll talk about the statistical aspects and the best practices for analyzing TUSCPS data, so the topic is different from Sean's and Anne's talks.

I'll start with explaining why we need standard errors in analyzing survey data. I will then introduce two different approaches to compute standard errors for TUSCPS estimates. At the end, I'll talk about how to construct replicated weights on March data sets.

In survey statistics, standard errors are commonly used to indicate the accuracy of survey estimates. They are mainly used to assess the sampling error, the error that results from taking one sample instead of examining the whole population. Standard errors can also be used to construct confidence intervals for survey statistics—for survey estimates. As most of you know, the confidence interval gives our estimated range of values which is likely to include our unknown population parameter, the estimated range being calculated from a given set of sample data.

Standard errors can be used to conduct statistical hypothesis tests, as many analyses as are needed. I will provide a couple of simple examples illustrating how to construct confidence intervals and perform hypothesis tests in the next couple of slides.

As I mentioned earlier, a confidence interval is a range about a given estimate that has a specified probability of containing the average results of all possible samples. Now let's—y-hat denotes the survey estimate for the outcome of interest; for example, the smoking prevalence for males 18+. The confidence interval of y-hat is—as many of you are already familiar with this formula, it's y-hat plus minus [t or thee?] times the standard error of y-hat. Here, [t or thee?] is the statistic assuming a normal distribution; for example, [t or thee?] equals to 1.96 if we want a 95 percent confidence interval; [t or thee?] equals to 1.65 if we want a 90 percent confidence interval.

Based on the most recent TUSCPS data, the 2010-2011 cycle, the current smoking prevalence of males 18+ equals to 18 percent. The associated standard error equals to .19 percent. Then the 95 percent confidence interval of y-hat is 17.7 percent to 18.4 percent based on this confidence interval formula.

The group Chi test is commonly used to test whether there is statistically significant difference between two group means. The Chi statistic is defined as the difference between the two group means divided by the standard error of the difference of the two means. If the absolute value of the Chi statistic is bigger than the [t or thee?] value, then the difference is statistically different at a certain alpha level. For example, the male current smoking prevalence equals to 18 percent with a standard error of .19 percent, as I just showed you. And the female current smoking prevalence equals to 14.2 percent with a standard error of .15 percent. The difference between the two group means is 3.8 percent and the standard error of the difference is .2 percent. So based on this Chi statistics formula, the Chi value equals to 18.9, which is bigger than the [t or thee?] value, the critical value, 1.96 at the alpha equals .05 level. So the associated [t or thee?] value from this Chi test almost equals to zero. That means that the difference between the two group means is statistically significant at the alpha equals .05 level. So you have to define that critical alpha value first before you conduct the hypothesis testing.

Now let's switch to how to estimate standard errors for TUSCPS data. Two main approaches can be used: the generalized variance functions method and the replication method. I'll go over the generalized variance functions method quickly. I think for most applications, this method is not that popularly used. We would rather spend more time on the replication method that—people use it more often.

A generalized variance function is a simple model that expresses the variance as a function of the expected value of the survey estimates. The parameters of the generalized variance function are estimated using [direct?] variances. These generalized variance parameters provide a relatively easy method to obtain approximate standard errors for numerous characteristics. But the generalized variance functions method can only be used to estimate the variance for means, [inaudible], percentages, and their differences, and cannot be used for complex estimates such as regression coefficients. For this method, you need the parameters, two parameters, a and b, and they are provided by the TUSCPS document. So for more details, you

can look at the following link. Attachment 16 of each CPS report provides a lot of details on this method. I will not provide too much detail here.

Replication methods are commonly used to estimate the variance of standard errors of survey estimates. The basic idea is to select subsamples repeatedly from the whole sample to calculate the statistics of interest for each of the subsamples and then to use the variability among those subsamples or replicate statistics to estimate the variance of the whole-sample statistics. There are different ways to create subsamples from the full sample. The subsamples are called replicates and the statistics calculated from these replicates are called replicate estimates.

Three replication methods have been developed in survey statistics: the jackknife method, the balanced repeated replication method, and the bootstrap method. The jackknife method is probably the most commonly used replication method in survey statistics because it can handle basically all kinds of complex designs. Depending on the structure of the design, JK1, JK2, and JKN can be used. TUSCPS didn't use this method. I won't provide a lot of detail here.

The balanced repeated replication method was originally developed for stratifying designs involving the sampling of two PSUs per stratum. The basic idea is to drop one-half of the PSUs for each replicate. Each replicate half-sample estimate is formed by selecting one of the two [of our?] units from each stratum based on a Hadamard matrix to form the weight for each replicate. The weights for the selected [of our?] units are multiplied by a factor of 2.

Fay's method is a special adjustment to the regular BRR variance estimation formula. Instead of multiplying the weights by 2, Fay's method is like a—puts a weighted average of the two of our units by a given factor between 0 to 1. So the procedure of Fay's method is the same as a full regular BRR. The primary advantage is when dealing with sparse subsamples, subgroups, there is some evidence that Fay's method gives somewhat better confidence interval coverage than the regular BRR when sparse subgroups exist. TUSCPS uses Fay's method under the BRR approach.

The bootstrap method has been used extensively for variance estimation in surveys. The idea is to generate artificial data sets of the same size and structure as the original data set by repeatedly resampling the PSUs in the observed data. The bootstrap method is more computation-intensive than the other two methods.

BRR is used to generate replicated weights for TUSCPS data, as I just mentioned. The full sample weight and the replicated weights have to be used when you analyze TUSCPS data. Please be aware that the replicated weights are not on the TUSCPS public use file. The 2010-2011 replicated weights are available from the Census Bureau's website given here. For earlier files, they are available from NCI upon request. You need special software such as Sudan, Westfar, Stata, or SAS Prox Survey Package to analyze TUSCPS data.

The replicated method provides a more accurate standard error than the [something F] approach in general.

Here is a general formula for the standard error based on replication methods. Here, R is the total number of replicates. C in this formula is a constant that depends on the replication method, like BRR or jackknife. Y-hat-R is the estimate for the outcome of interest based on the R of the replicate weight. Y-hat-zero here is the estimate for the outcome of interest based on the full sample weight. For TUSCPS, C is equal to 4 over R; 4 is the inverse of the C factor.

This is the standard error formula for TUSCPS. The number of replicated weights equals to 48 for the 1980-based designs. R equals 80 for the 1990-based designs. And R equals to 160 for the 2000-based designs. So it's the same formula as I showed you earlier, but just plugging 4 over R to replace C—the constant, C.

Here is a simple, specific example. Assume we have three replicated weights in total. The estimated y-hat for a given outcome of interest based on the full sample weight is 10. So based on the three replicated weights, y-hat are 8, 11, and 12, respectively. Plugging those numbers into the formula represented earlier, we can get the standard error of y-hat equal to 3.46 after calculating this square root formula.

Now let's see how to implement replications. We need—the first step is to create weights for the full sample. Then, we need to form replicates of the full sample and then create weights for replicates following one specific method such as jackknife or BRR, depending on the design. Then, we attach weights to the survey data sets; finally, we can compute estimates and the standard errors using special software such as Sudan, Westfar, Stata, etc. Usually, the public use files of large-scale surveys provide you the full sample weight and the replicate weights, such as like the TUSCPS. You don't have to create the replicate weights by yourself.

Since we have several cycles of data available, people are interested in analyzing multiple years of data together to enlarge the sample size or look at some trends across time. Since the survey design has been changed several times in the past two decades, we have to adjust the replicated weights to account for the merged data. Within the same sample design, the data from multiple years are correlated. Across different sample designs, the data from multiple years are all correlated. As I mentioned earlier, for the 1980-based cycles, there were 48 replicated weights created for each cycle. For the 1990-based designs, there were 80 replicated weights created. And for the designs starting from 2000, there are 160 replicated weights created for each cycle.

When combining data within the same sample design, one can just concatenate the data together. No special adjustments are needed to adjust the replicated weights. The Fay factor is unchanged. But when combining data across different sample designs, we need to stack the replicates and adjust the replicated weights to account for stacking. We also need to change the Fay factor from 4 to 16.

Todd's presentation, next, will show specific examples on how to write the SAS codes for merging multiple years of data. I will not spend too much time here.

In this talk, we covered why we need standard error in analyzing survey data. I also explained two approaches to compute standard errors for TUSCPS estimates. At the end, I explained briefly how to adjust the replicated weights when combining multiple years of TUSCPS data together. That's all I want to talk about today.

Thank you very much.

**TIPS AND TRICKS FOR HANDLING TUS DATA—TODD GIBSON**

MR. GIBSON: Good afternoon. My name is Todd Gibson. I'm with Information Management Services, and I'll be going over some tips and tricks of handling the TUS data. I'll be going over some basic information on getting started with the data, merging replicate weights, working with multiple years of data, merging an overlap sample, linking to other CPS files and other supplements to the CPS.

The public use files are ASCII text files, and included with the files is technical documentation that has an overview of the Current Population Survey, an overview of the Tobacco Use Supplement, the record layout for the file, the TUS questionnaire, a source and accuracy statement, and we've added to the end of the documentation user notes for any updates

to the file. Another thing to note is for the 2003 and later data, we currently have programs up on our website that take the ASCII text files and creates SAS data sets from them.

There are core and supplement variables. The core variables include state and other geographic information, information on family income, race, origin, gender, age, education, marital status, along with the labor force information and occupation. The Supplement has various variables, including language of interview; the interview method, which is either telephone or in-person; relationship to the proxy for proxy responses; cigarette smoking prevalence; smoking history; menthol use; cost of cigarettes, as Anne mentioned earlier; use of other forms of tobacco; smoking policy at home; smoking policy at the workplace; attitudes toward smoking; and medical and dental advice to quit.

When working with the TUS data, better estimates can be obtained using three months of data collection, and each file has two sets of weights: the nonresponse weights, which are used for analyses that include both self- and proxy respondents; and self-respondent weights for analysis of self-respondents only.

For my first example, I'm going to do a very simple example to do point estimates using the May 2010, August 2010, and January 2011 data. The variables that we'll look at are age, gender—I'm going to include some variables that aren't in the tables like race; ethnicity; education; interview, PRS64—who is actually responding to the Supplement, whether it's a self- or proxy respondent; and for this one we're going to stick with self-respondent, so we'll use self-response weights. Selections that we'll make are PR person type equals 2, which is adult civilian records ages 18 and older who had the Supplement interview and were self-respondents.

Since each survey is weighted by the population and we're combining three surveys, we need to divide the weights by 3, and we'll produce a table of current smoking prevalence rates by gender and age.

The first part of the SAS code is just the three files that we'll be using—the May, August, and January—and some formats that we'll use in the tables. The next part of the code is a macro that can be used for each of the surveys, and we're going to read in month and year of survey. I included the state variable just to show that you could do your estimates by state; age; gender; education; race; and origin; PR, person type; questions A1 and A3—interview status, smoking status; self- or proxy respondents; and then the weight. So we select adult, civilian household members ages 18 and over from the Supplement interview, self-respondents, and

we're going to exclude "don't knows," "refused," and nonresponses. And the weights on the CPS and the TUS all have four implied decimal places, so we divide by 10,000.

The next three statements just read in the data using the macro. And then the next data step—we merge these three data sets together. As mentioned earlier, because we're working with three, we need to divide the weights by 3.

I have some statements to create an age recode for ages 18 to 24, 25 to 44, 45 to 64, and 65+. We associate the formats with the different variables and the labels. And using a simple cross-tabulate in SAS, we'll generate a table of current cigarette smoking status by gender and age group. The variables we're using are gender, age group, and smoking status, and it's weighted using the PR—the self-response weight.

And this shows the table that you generate from the SAS code and shows the gender and the age group for the rows. And, then, across the columns, we have this [inaudible] recode for never and former combined and current combined, which is every-day and some-day smokers, along with the totals.

The second example I'd like to go over is merging the replicate weights, which are very important when we're doing standard errors and confidence intervals. There are two replicate weight files for each survey in 1992 through the 2011 data. One file contains the nonresponse weights, as I mentioned earlier, that are used when you're doing an analysis with self and proxy. And the second set has the self-response weights. And Benmei had touched on this earlier—that, depending on which set of files you're using, there are different numbers of replicate weights. In the '92-'93 files, there're 48 replicates. There are 80 replicate weights in the '95 through 2003, and the 2006 and later use 160 replicate weights.

Also, depending on which data you're using, there are different unique identifiers for merging the replicate weights onto the main survey data. For '92-'93, we use the household ID and the person's line number. For '95 through '99, it's a household ID, a serial suffix, and a person line number. And then for 2001 and later, we have a unique household identifier and a unique person identifier to be used for the merge.

We go through an example merging the replicate weights and calculating the current smoking prevalence using SAS and Sudan, using the May 2010, August 2010, and January 2011 data. To do this, we'll read in the main survey data, then read in the replicate weight data, which have multiple lines per record, sort and merge each of these by question

number and occurrence number, and once again because we're using three surveys, we'll have to divide the weights by 3. Using replicate weights in Sudan, we'll calculate current smoking prevalence, standard errors, and 95 percent confidence intervals. With the options in the cross-tab procedure using design equals BRR, for balanced repeated replication, as Benmei mentioned, and also the adjustment factor of 4 for the Fay adjustment, and generate a table of current smoking prevalence rates by gender.

So the first three files using in this SAS code are the same as the last example showing the May, August, and January. The next three files are the actual replicate weights. And, once again, we have a few formats for use in the table.

So the macro has changed a little bit. The top code is the same as the previous macro, but starting about halfway down the page or a little after halfway down the page, we have the code to read in the replicate weights. And so we read in the replicate weights with question number, occurrence number—the sample weight is the main weight and then the replicate weight is 01 through 160.

The next two sort procedures just sort the main data and the replicate weights, and then we create a data set for each survey merging the main survey and the replicate weights by question number and occurrence number, keeping only the respondents that are in the survey.

The next three statements actually go through and read the data in, and then starting with the data step, we merge those three together and divide our weight by 3—the main weight on the survey and our 160 weights, by 3, and generate our age groups again. And this time, we add a variable for current smoking.

Using the cross-tab procedure in Sudan, we can generate the table. The set environment statement just defaults the number of decimal places to four, and we have our weight statement for our main weight and our replicate weights for our 160—and you'll see the adjustment Fay equals 4 statement at the end of that line.

The variables that we use in the table are the current smoking status, gender, and age group, and we'll generate a table by sex and also by age group.

The table looks something like this. This is the table by gender. I haven't included by age group, but it would look similar except the age groups would be down the left side.

The next example I'd like to go over is an example of merging the replicate weights and calculating current smoking prevalence when we have surveys from two different

time periods. And Benmei had touched on this earlier—what we'll need to do is read in and merge the main survey and the replicate weights like we did in the previous example. But seeing how we're going to work with two survey groups that have different replicate weights, we'll have to construct a new set of replicate weights. Our 2003 surveys had 80 replicate weights and the 2006-'07 had 160. So I've backed up on the time point that I've used; instead of using the newer data, just to show for this example, going from 80 to 160 replicate weights and constructing the new weights. And as Benmei mentioned, we'll have to change our adjustment to 16 from 4 for the Fay adjustment.

This first slide just shows the six different survey data files and then the six replicate weights and some formats.

So I've set up a macro for each time point that reads in the main data and then the replicate weights. So in this one it's for 2003. We'll have 80 replicate weights and, just like before, we merge by question number and occurrence number and keeping only the respondents in the main survey.

Similar to the 2003, there's a macro for the 2006 and 2007, and this time we have, instead of 80 replicate weights we have 160 replicates. Once again, we sort and merge by question number and occurrence number.

So we read in all six sets of data and merge them together, once again created in our age group variable, our current smoking status variable, and also a variable for survey group so we can use it later on.

This next portion of code shows how to construct the 240 replicate weights. Our old weights either have 80 or 160. Our new replicate weights will have 240. Because we're working with six different surveys, we'll have to divide our weights by 6. And we create a new weight by taking one-sixth of the sample weight plus a factor that's calculated by taking the square root of the total number of weights over the number of weights in that survey divided by 2, and multiplying that by the old weight minus the sample weight. So if the survey group is 1, we're creating these new weights in the first 80 weights and then weights 81 through 240 are just the sample weight divided by 6. And then when we move to survey group 2, under the [inaudible], the first 80 weights are just going to be the sample weight divided by 6, and then the new weights are going to be 80 through 240.

Once again using Sudan to generate the percentages and standard errors and confidence intervals, we still use the design equals BRR. And, this time, instead of having 160 weights like we did in the 2010-2011 data, we have 240 for the combined 2003, 2006, and '07, and our adjustment factor is set at 16. Once again, we'll create a table by gender and by age group.

And this is the table by gender. Once again, I didn't include the age group table here.

The next thing I want to talk about is merging the overlap sample. Anne had touched on this, and it's a unique CPS panel design feature where each household in the sample is surveyed for four consecutive months, which is panels 1 through 4, and then for four consecutive months later, which are panels 5 through 8. So when looking at the February 2002 and February 2003, we can match panels 1, 2, and 3 to panels 5, 6, and 7.

Matching variables for matching the February 2002 and February 2003 overlap include the household identifier, the month, and sample. So it was months 1, 2, and 3 in 2002 get matched with months 5, 6, and 7. There were no TUS items for panels 4 and 8 in 2002 and 2003, so we're only matching 1 through 3 and 5 through 7. Other variables that are needed for the match are the sample identifier, the serial suffix, a household number, the person's line number, and then we looked at gender and person's age and made sure that the age was within plus or minus 1for the 2002 age.

So when we did the match, we get almost 22, 600 self and proxy that match and almost 16,000 self alone. When we do the match, the reasons for mismatches include the entire household or individuals moved to another location, or individual or household nonresponse.

As mentioned earlier, we can merge the TUS with other basic CPS data and CPS supplement data. And some of the ones that we can merge with are the March [ASEC?] data, the American Time Use Survey, voting registration, computer and Internet use, and food service securities information. One use where we've used this in the past is, starting in January 2003, the Occupation and Industry categories that were coded were different to the prior data, pre-2003. The questions that were asked weren't modified, but the actual information gathered was classified according to new standards and definitions. So we had to come up with a way to group these new coding schemes into occupation groups that we've used in the past, which were white collar and blue collar, service, and other. So what we did is we merged the February 2002 CPS

data with February 2002 Bureau of Labor Statistics monthly extract file. And this February 2002 BLS file had the new code. So we were able to look at the old codes on the February 2002 CPS with the new codes on the BLS file.

That concludes my part of the talk. Thank you for your time. And I'll turn it over to Gordon.

**QUESTION-AND-ANSWER SESSION**

MR. WILLIS: Okay, thank you, speakers, very much. And now we're technically into our question-and-answer session. And the way we're going to, at least in theory, do this is to be flexible. But the plan is to address each of the questions that were posed to the speakers, by speaker. We're going to cluster them according to speaker. And so, in a second, I hope, Janna is going to bring me over a compiled list of all the questions—not that there were that many—and we'll load them up here on this computer, and then we'll have them all nicely arranged. Just give us a minute to take care of this and we'll be back with you.

Okay, we're back with you now. The way we're going to do this—we're going to start off with the questions that were asked of Sean, which is actually our second real talk. The way we're going to do this is I think we'll have the person who is addressed the question try answering it. If someone else on the panel thinks they can a better job, you're free to answer it. But then, after your question is answered, feel free to type into the chat area with additional requests for clarification, whatever. We'll try to work with this in real time, so at the end of the questions for that speaker, we can get to those as well before we go to the next speaker. I think there aren't that many questions.

But to get going on this, the first question comes from Any McQueen for Sean. What sort of resources or arrangements are needed to use option 1 out of the two use options that you mentioned; that is, having the NLMS analyst do the analyses rather than having to become a special sworn employee and analyze the data yourself?

MR. ALTEKRUSE: Thank you for the question, Amy. It's a real simple process. There is a form that you complete, and it's generally a two-page concept really summarizing the variables that you need, the hypothesis, maybe some background from the literature. And that's submitted to the sponsor agency of NLMS on the steering committee. So if your question relates to cancer, it would go to NCI. If it relates to heart disease, it would go to the National Heart, Lung, Blood Institute. If it's related to aging, it would be to the National Institute on Aging. If

it's more of a health statistics question, it would go to NCHS. I put on the last slide of my presentation my email address, and if you would like a copy of that form, you can send me a request for it. The completed form goes to the steering committee for review, and it's generally not a vetting process but a development process. There may be some back and forth to try to refine the question.

MR. WILLIS: Okay, our second question for Sean comes from Cindy Chang. Are individuals whose data come from proxy interviews in TUS included in the NLMS; that is, in addition to the self-respondents?

MR. ALTEKRUSE: So the short answer to that is, yes, they are. The match with NLMS to TUS is all responses, and those are marked so you can tell which is a proxy vs a self-response.

MR. WILLIS: Okay, and the final question for Sean—I feel like I'm on a game show here—from [Sa Shang?] is about the data linkage. Do you have step-by-step instructions for users and how to link the TUSCPS to a variety of data sets, or have the data already been linked so that all you need to do is apply for the access?

MR. ALTEKRUSE: So that's a great question. And the short answer to that is that the data are already linked. And, in fact, the reason for that is because they contain information behind the firewall at the Census Bureau that you are never going to be privy to—things that could disclose the individual's identity. So the linkage is done by the Census Bureau and the variables that are needed for your analytic data set are made available to you. Whether that work is done at the Census Bureau by a Census employee, or whether it's done by you as a special sworn status employee of the Census Bureau—the special sworn status component requires that you complete two web-based module protection and have fingerprints done and that sort of thing. But that enables you to access the data at any one of the remote data centers that the Census Bureau hosts around the United States.

MR. WILLIS: Okay, do we have any follow-up comments to address here related to Sean's questions or can we move on to the next speaker?

Okay, so we're going to move on to questions related to Benmei, who talked about a number of complex statistical things here. So, Benmei, presumably, you're there and you can answer these, but even if not, the way this works is we have a little bit of an interactive component here where Benmei answered a couple of these questions already on line. So what

I'm going to do is ask the question. Then, I'm going to read Benmei's answer. And then, Benmei, if you'd like, you can weigh in further on your answer.

So the first question for Benmei is from Paul Harrell. Can you repeat software can be utilized for the data sets? To clarify, I'm referring to the use of weights. I though you said Stata and SAS could be used but do not see those listed on the slide. And Benmei's answer was: It can be Sudan, Westfar, Stata, SAS Prox Survey—a number of packages that can be used in analyzing TUSCPS data.

Benmei, anything to add on that?

DR. LIU: Yeah, another available software is R, which is a free software. They also have a survey package under the R software, which can be used to analyze complex survey data. That's another option. I think Sudan—and I'm more used to Sudan and SAS Prox Survey package. I think those are enough. But some people prefer using Stata or other software.

MR. WILLIS: So, Benmei, what software would you say is used the most to analyze TUS data? Do you know what's the easiest to work with?

DR. LIU: Yeah, I think NCI always prefers Sudan. Like Todd's presentation showed most of the Sudan code, like merging those data sets, analyzing—that you had the CPS outcomes.

MR. GIBSON: Sorry to jump in here, Benmei, but, yeah, that's correct. And one of the reasons why—we started with Westfar and then we went to Sudan. And as far as the SAS Prox Survey package, it wasn't available for use with replicate weights until recently. So because we've been doing this for years, it's—we've kind of been using Sudan the most, and it's Sudan that's callable through SAS. So we set everything up in SAS but SAS then calls Sudan to do the procedure.

DR. LIU: And another comment: Westfar is a free downloadable software, where with Sudan you have to purchase the lessons, also SAS. Westfar has a lot of limited availability for analyzing the data sets.

MR. WILLIS: It does at least seem that more statistical packages are respecting survey data now than used to be the case, so that's a nice trend. At least we have more options.

And did we just get a comment? I saw something flash by. Okay, so I'll move on then to [Carribe Nandi's?] question, again for Benmei. According to the formula for replicate weights, these were calculated for specific outcomes. So how can other users use these for their

own specific outcomes, which may be different than the outcomes that you used. And Benmei has already said: The formula is the same. In my example, I used proportion outcome. The same formula can be used for totals, means, regression coefficients, etc.

So, Benmei, I take it to mean that it's not exactly the same formula, is it, but one that's appropriate for means or regression coefficients from a basic statistics textbook—is that what you're saying?

DR. LIU: Yeah, if you want to calculate the standard errors, you need to replicate the weights by [half?], basically you plug in the same formula, like you calculate the outcome of interest using the full sample weight first. And then you compute an estimate using each replicate for the same outcome you're estimating. And then you plug in those estimates based on the full sample weight and the estimates based on each replicated weight and then plugging in the same formula as I showed earlier. But when you use those software—Sudan or Westfar—you do not need to calculate them manually; the software will calculate them for you.

MR. WILLIS: Okay, thank you.

DR. LIU: Like TUSCPS has 160 replicate weights, and to calculate to the standard error manually is a lot of work—involves a lot of work. So and the usual approach is to just use the software to calculate for you.

UNIDENTIFIED: Okay, Benmei, we have a follow-up: Would the replicate weights change for different outcomes?

DR. LIU: No, the replicated weights are calculated for the data set. It's the same for every outcome.

MR. WILLIS: Okay, I hope that was clear; assuming it was, we're going to move on to David Timberlake's question for Benmei. In the absence of design characteristics—for example, PSU—how can we estimate variance within multilevel analyses of state-level data? And Benmei said: Good question! PSU information is not available for TUS for confidentiality purposes. Some analyses are restricted. You cannot use PSU information for your analysis.

It seems to me that, Benmei, you are basically agreeing with the premise of David's question, which is: Since there are not PSUs or that kind of information, for the reasons you say, presumably, Benmei, how in the absence of that information can we estimate variance in multilevel analyses of state-level data?

DR. LIU: Yes, the PSU is—since the information is not available, I'm not sure the multilevel analyses you are doing—if your original purpose is to calculate the median PSU variance as a second level under the state level, I'm not sure—I don't think you can do that because PSU—the information is not available.

MR. WILLIS: So is that the answer—you can't, basically?

MR WILLIS: Yeah.

MR. ALTEKRUSE: I have a follow-up to it.

MR. WILLIS: Okay, go ahead, Sean.

MR. ALTEKRUSE: So, Benmei, is there information available at the county or the census tract level for TUS to allow that multilevel?

DR. LIU: The TUS and CPS only release the county identifiers for counties with a population size like 200,000 and more. So for the rest—for the smaller counties, they are all concatenated into one category. So, basically, you cannot utilize county-level information. The project I'm working with and collaborating with the Census Bureau—I'm working on producing more error estimates at the county level for several outcomes from TUSCPS. I have no county identifier access, so we have to collaborate with the Census Bureau to do the other—to produce estimates at the county level. NCI doesn't have access to the confidential county identifiers, not to mention the PSUs.

MS. HARTFORD: I'd like to add that while we don't have county-level identifiers, there are other geographic variables that are on the substate level. Certain cities are identified and, actually, some metropolitan areas or clusters of counties. And the example I gave you for—sort of 75 percent of major media markets were able to be identified, and that, of course, also crossed state lines with subgroups of states. So there may be some things that you can do, depending on what the location is and the question, without having to go and become a sworn Census employee to utilize everything on the county level.

Plus, let me take this opportunity that we're actually working—Benmei is working with the Census Bureau to look at some small area estimations using TUSCPS.

MR. WILLIS: Okay, it looks like a hot topic. Any other comments from the panel? If not, in fact, there's one more question that could be for Benmei. I know it's not for me. How did you arrive at ADJ Fay equals 16? That must be something named after Bob Fay, I

presume. You know you've made it in the world of statistics when you have a statistical variable in an analysis named after you. I'll tell Bob that.

How did you arrive at ADJ Fay equals 16 for the combined 2006-'07? Is that a question for Benmei? Do you want to take that one?

DR. LIU: Yeah, I can take that one. That was that derived based on the formula. I have on my sheets all the formulas for deriving the factor of 16. We want to—like after merging the data sets together, we want the variance to stay the same as before. So, basically, we're using the original formula before merging, and then using the new replicated weight after merging. And we want those two variances to stay the same. So after some mathematical calculation, we calculated the Fay factor as 16 after the merging instead of 4. That's the short answer.

MR. WILLIS: Okay, any more comments from the panel on that one? We've answered it?

Okay, then I think that's it for Benmei. Are there any more follow-up questions that have come in over the pipe here for Benmei?

All right, that moves us to a single question, I believe, for Todd, which is: How do we know that fore May 2010 and May 2011 data which panels were longitudinal?

MR. GIBSON: For the May 2010, panels 1 through 4 are the panels that would be used, and then in May 2011 these panels are 5 through 8, because it goes through the first four and then nine months later the same four, but they are 5 through 8 then.

MS. HARTMAN: I'd like to add that Census will be taking care of that problem for you because what we will release to the public will not be, like in the past, a separate May 2011 file where you all have the hassle of trying to get guidance for us on how to produce weights and how to merge. So there will be one file and it will actually be a longitudinal file which will contain the baseline information from May 2010 along with the follow-up information from May 2011. So I believe this should be clear, and you may not even need to know about what panels. The panel is a variable in the file.

MR. GIBSON: Right, the example that I gave for February 2002 and 2003 is how you would do it if you had to do it. But for May 2010 and 2011, the file for the public will already be merged and have the variables for both surveys.

UNIDENTIFIED: And we have a follow-up to that. Is the longitudinal data already available?

MS. HARTMAN: No, I mentioned in my talk that we hope to have that available by the end of the year, and we will keep people posted. We'll put some information on our website.

MR. WILLIS: Okay, well, it's nice to hear that we're doing something here to make the longitudinal data more easily analyzable. It's almost like you guys are saying, "We're from the government and we're here to help you."

DR. LIU: Yes.

MR. WILLIS: At least trying. Anything else related to Todd's talk to follow up on?

Okay, which brings us to a few general questions, which I suppose fall to Anne Hartman but can go to whatever panel member feels empowered to address the issue at hand. First, for the May 2010 to May 2011 longitudinal data, what are the appropriate weights to use for analyses? This is a follow-up—the same issue.

MS. HARTMAN: Yes, so what I was saying is that the Census Bureau is actually constructing special weights for the longitudinal data, so that will be provided already. And that, of course, will depend on—I believe there should be similar—in terms of like nonresponse weights, that include proxies and self-response weights for self-only examination.

MR. WILLIS: And another question from the same questioner: What sampling weights should be used for analyses after we link TUS data to some other national data, like the ones you mentioned in your talk, Anne?

MS. HARTMAN: That's a good question, and it probably depends on what months and what panel you are looking at. Basically, a general response would be that you might do something similar to what we have in our overlap report for how to look at the February 2002 and February 2003 data and how to create weights. But, also, one example is with the March supplement—that's the sort of basis for NLMS. Sometimes, you can actually see that, for instance, the May and January TUS data fit and kind of complete the March. And so, in some instances, that may actually mean you may be able to use the March weight. So it's got to be tailored to the individual circumstance.

MR. WILLIS: Well, this really seems like it's getting at the crux of the whole issue of making the data analyzable, especially once they're linked. So how much—Anne, how much guidance do we provide to people, say, on the website? Is there information that helps

people get through this—to answer this question of: If I do merge different data sets, what sampling weights are appropriate and how do I come up with them? I mean how is it that they're supposed to do this?

MS. HARTMAN: To the extent of some of the specific examples, we have some information in the technical documentation for specific—for instance, like the 2002-2003 use of longitudinal data, I guess our biggest thing was the overlap report that we have on there. And, also, there's like many, many ways that people can merge the data. And so we're not necessarily even that familiar with some of the supplements that people have used. But just talking to a few of those sponsors, they're eager to do this type of thing, too. And I think they'd be happy to work with you, and also the Census Bureau. And we would, if we can, so ask questions.

MR. WILLIS: Okay, anything else anyone wants to add from the panel on that? If not, we're—do we have any other input recently that's come over the chat area? Yeah, before I go, the last question here, which is from Wanda: Is there an explanation as to why the current smoking rate is less than that from BRFSS for the U.S.?

MS. HARTMAN: I'm assuming you're talking about the total U.S., and usually what's published with BRFSS, when they provide a total U.S., they actually are using like a median state estimate as opposed to nationally representative estimate. So you would expect that to be different. Also, there are different—obviously, different methods and survey contexts, so estimates you would expect to be somewhat different. And, finally, the BRFSS is structured to be kind of state-specific and the methodology and how surveys made in each of the states are conducted also differ. So that contributes to that, too.

MR. WILLIS: And, Anne, how do the TUS estimates for prevalence of smoking relate to those from other national surveys like the Health Interview Survey?

MS. HARTMAN: We typically are a little lower than the NHIS, national estimates for smoking prevalence. And that is largely explained by differences in methodology. A number of years ago, NHIS actually looked at mode of interview, where both actually household probability samples were not random-digit-dialed or anything, but there are a percentage of respondents in both the NHIS individual information and, of course, the TUS that are answered by telephone as opposed to in person. And a number of years ago, actually, NCHS found that for most behaviors there was not any difference. Smoking was one of the few health behaviors where there was a slightly lower estimate for telephone interview vs in person. And,

amazingly, we published some methodologic papers and looked at mode, and we found very similar estimates. So when you adjust for that, that pretty much explains about 90 percent of the difference. And our patterns are similar; they fit like a glove.

MR. WILLIS: Thank you, Anne. This is Gordon again. To me, this speaks to the general issue, not only for tobacco, of why it is that prevalence estimates differ between surveys. And the general answer that we've always come to for that is that there are so-called house effects related to surveys. There are a multitude of differences that wind up producing somewhat different estimates, but as a methodologist, I'm always really surprised and reassured, actually, as to how close these normally are. So whether you look at TUS or the Health Interview Survey, the estimates are not much different. People want really high precision, but if we're off by a percentage point or two or different, then that's probably not too bad for an endeavor that relies, after all, on human self-report of their behavior.

Okay, enough of my prosthelytizing. Are there any more comments, questions, anything else that's come in? And did I miss any questions from your nice list?

If not, that brings us to our closing here. The only things that I have to say in closing—maybe Anne wants to add a few more things since she's our major host. One—we do plan to send a post-questionnaire, a little Internet-based survey, out to registered participants as an evaluation of our webinar. We'll use your responses to that in large part as a basis for deciding whether we think it's worth doing a webinar like this again. So tell us what you think about how worthwhile this activity was. We're also going to make the recording available as soon as we can and have that on our website and announced in an email to participants, I believe. Janna, do you want to correct me or add anything more to this?

MS. EISENSTEIN: I don't think we've decided yet where we're going to house the recording and the slides, but they will be posted shortly and we'll let everyone know where they're posted.

MR. WILLIS: Okay, so at least we'll let you know when we do.

Anne, anything you'd like to add before we close?

MS. HARTMAN: You did mention evaluations.

MR. WILLIS: Yeah, I did mention that we're going to send out an evaluation questionnaire to our registrants so we can, on that basis, decide how useful the webinar was.

MS. HARTMAN: Actually, then, with that, I'd like to thank the audience for taking the time. And, also, this is our first attempt at doing a webinar, so I'm happy you were patient with us. And I'd also like to thank not only the speakers here but those that are behind the scenes for making this possible: Julia Strasser and Janna Eisenstein [ph] and also Susan Scott, who was responsible for communication; and certainly our host here, NOVA Research Company, who also was involved in the website development and all and of course actually conducting the webinar and recording; and, of course, like I said before, also the Census Bureau and Bureau of Labor Statistics for allowing us to be involved for over 20 years with the CPS. Thanks.

MR. WILLIS: Presumably, our participants have someone's email address centrally here so that if they think of more questions, they can send them after the fact.

MS. EISENSTEIN: Yes, I believe everyone has Julia Strasser's email address.

MS. HARTMAN: Yeah, actually, I had a slide—I guess we can't.

MR. WILLIS: Anne Hartman's presentation?

MS. HARTMAN: Yeah. There's our website—from our website, there is a general "contact us," and I know a number of you have used that, because we've gotten questions from that. I don't know if we can bring up …

MS. EISENSTEIN: Yeah, we'll bring it up.

MS. HARTMAN: Good, so it's the one before this, I think. There. Oh, next slide, okay. Yes, it's http://appliedresearch.cancer.gov/studies/tus-cps … oh, no, I'm sorry. That's our general website, I'm sorry. For the contact it's: http://appliedresearch.cancer.gov/about/contact.html.

MR. WILLIS: Okay, that's good. You may have just saved a few of us from a whole lot of emails, especially Janna.

If there's nothing else, I'd like to thank Anne, our panel, our participants, and we hope to do something like this again in the future. So thank you very much.