National Cancer Institute
Tobacco Use Supplement to the Current Population Survey (TUS-CPS)
2021 Data User Webinar – Analyses Using the TUS-CPS 1992-2019 Harmonized Dataset

Webinar Transcript

MS. NALINI CORCY: Hello, everyone. Welcome and thank you for joining. My name is Nalini Corcy, and I’m the host for the webinar today. I will be providing technical support along with my colleagues. Welcome to the National Cancer Institute TUS-CPS Data User Webinar. Today's topic is going to be analyses using the TUS-CPS 1992–2019 Harmonized Dataset. Just a couple of housekeeping items for our webinar today. So, firstly, all participants are on mute. Please, at any time during the webinar, feel free to use the chat box, which should be on the right-hand side of your WebEx screen. You can enter questions for the speakers in the chat box and also requests for technical support. For any technical support requests, we will try to respond to you as soon as possible. Questions for the speakers, we will be saving them and addressing them during designated Q&A periods throughout the webinar. We have a closed captioning service available today. I will be dropping a link to that in the chat box shortly, so please look out for that. We are also recording the webinar today, and the recording as well as all slides and materials will be posted online in approximately three weeks’ time. An email will be sent out to all participants.

So, I wanted to briefly introduce our speakers for today. So, first we have Dr. Carolyn Reyes-Guzman. She is a program director with the National Cancer Institute Behavioral Research Program, Tobacco Control Research Branch. Our next speaker for today will be Dr. Elizabeth Seaman. She is a project manager with the CDC Foundation. And finally, we will be hearing from Dr. Kelvin Choi, who is a senior investigator with the National Institute on Minority Health and Health Disparities in the Division of Intramural Research. So with that, I’d like to once again thank everyone, remind everyone to please enter any questions or requests for technical support using the chat box. I will be putting in a link for the closed captioning service. You are welcome to use that if you need. And I think we are ready to have our first speaker, so, Carolyn, if you want to take it away?

DR. CAROLYN REYES-GUZMAN: Great. Thank you. I just need to obtain the share content permission. There we go.
OK. Can you all see my screen?

MS. CORCY: Yes, we can.

DR. REYES-GUZMAN: Great. Thank you. All right. Good afternoon, everyone, and welcome to our first webinar of the series for the 2021 TUS-CPS Data User Webinar series that will be happening from now until November. The focus of today's presentation is on how to conduct analyses using the harmonized dataset with data from 1992 through 2019. So, I am Carolyn Reyes-Guzman. I am a program director at the Tobacco Control Research Branch at NCI, and I also serve as the co-lead for the TUS-CPS study.

So, just a brief agenda on what I’m planning to cover. I’m going to begin with a brief background on what was the impetus behind the harmonization process. I’m going to lead you through the importance of the self-response and replicate weights for variance estimation when using data from the TUS-CPS. Then I’m going to walk you through three simple examples on different scenarios of using the harmonized data. And then the second and third presentations will be focusing on the applied examples of using the harmonized data. So, as Nalini mentioned, just a quick reminder that we will have a designated Q&A session at the end of the third presentation. But in the meantime, if you have any questions, you can always type them into the chat box and Nalini will ask them at the end of the presentation if it's an urgent question or at the end of the third presentation once we go into the Q&A.

So, just wanted to add this disclaimer also that the views and opinions expressed are the presenters' only and do not necessarily represent the views, official policy, or position of the U.S. government.

So briefly then in terms of some background to set this up, so the TUS-CPS is a study that is co-sponsored by the NCI and FDA as part of the annual U.S. Census Bureau's Current Population Survey. The TUS-CPS has been happening every three to four years since 1992. And the most recently publicly released data are from the 2018-19 wave. We are currently underway with planning together with the Census Bureau for the next cycle in 2022-23. And these are the previous waves with TUS data, so we have data from '92, '95, '98, 2000, 2001, 2003, 2006, 2010, and 2014. And each of these contain – for the most part, each of these contain three time points, so three months of data. And also to mention that the CDC was a co-sponsor together with NCI from 2001 through the 2006-07 cycles.

So, what is TUS data useful for? So, I wanted to just bring this back as to there's several surveys that collect data on tobacco use behaviors, tobacco use patterns, but what is unique about the TUS? So, this dataset can be used by researchers to monitor tobacco-control progress and assess long-term cross-sectional population trends, so just reminding you all that this is a cross-sectional survey. It's also useful to track tobacco health disparities or tobacco use disparities, to evaluate tobacco control programs, and also, one of the uniqueness of TUS is in its capacity to examine smaller unit, geographical unit-level data. So, apart from national and state-level data, in some instances, there's also county-level data that can be analyzed. And the other interesting thing about TUS is in the linkages to other CPS supplements, including, for example, some of the detailed labor force, occupational, and economic health insurance data from the CPS's March Supplement, also known as the ASEC, as well as other linkages. And there will be a webinar on TUS linkages from the CPS in September. So, just know that that will be coming.

So, in terms of why we at NCI decided to embark on this process of harmonizing the different years, waves of the TUS is that previously researchers wanted to examine long-term trends, had to basically track all the variable names for a specific question of interest over several different waves and sort of keep track of all the differences in the naming of those variables, differences in – potentially slight differences in wording. And so, the idea with this harmonized file was then to create a single variable name that would remain consistent across all the waves, which would then permit the investigator to use a quote-unquote "flag variable" to keep track of the survey year. The harmonized file is restricted to only the adult self-respondents. So, there's no data on proxy responses due to the potential reliability issues of those proxy responses. And so, the inclusion in the harmonized file meant that questions with at least two time points were to be included in the file. So, here is a little more detail about what the inclusion-exclusion criteria for harmonization was as we went through this process. The variables that were easily harmonized were those that were deemed as is, so to speak. Those were questions that remained consistent over time and were basically harmonized in their current state. Then there were questions that needed some slight adjustments. For example, there were slight tweaks in wording over time or there were some instances in which the universe of respondents has slightly changed, and so we as a team had to make decisions on whether those questions could be harmonized, and for the most part they were. And then there were a few instances where we had to drop certain variables from the harmonized dataset because those items included significant wording changes, structures, or very significant variation in the respondent universe. So, with all that, we created back in 2016 – that was the first harmonized file with data through the ‘14-‘15 cycle, and we’ve now finalized and will be publishing on the TUS site the updated harmonized data with ’18-’19 data as well as the replicate weights and the data dictionary and crosswalk. So, all of that will be posted, as Nalini said, in the next few weeks. We’ll send an email out to announce that that’s ready. And we are also in the process of developing a harmonized file user guide that will be released to our website in the fall with very detailed examples using SAS and SUDAAN codes for you to be able to run more detailed analyses. So, these are the topics that are covered in the harmonized dataset.

We included the core variables from the Current Population Survey. These are all the sociodemographic, geographic, occupational, and economic variables. Then there’s the questions on cigarette use including menthol cigarettes from 2003, the questions on workplace and home smoking restrictions, on attitudes toward smoke-free policies in indoor work areas or public places, and a new harmonized item or set of items includes questions on smoke-free attitudes for multi-unit housing. So, you may know that this was a series of questions that was asked for the first time in the 2014-15 cycle. Now with the ‘18-‘19 data, we have at least two time points, so that fulfills the criterion for inclusion in the harmonized data. We also have questions on physician or dentist advice to quit, on health perceptions and beliefs, sort of the harm-reduction questions, as you may have seen in the past studies. And questions on smoking history, cessation, and a very detailed section on former smokers, including especially for those interested in former smokers who may now be nondaily cigarette smokers, but you’d like to learn about what their previous behaviors were in terms of whether they were daily smokers who then transitioned to nondaily. That former smoking section has those questions. There's also a very long section on use of other tobacco products, cigars, pipes, smokeless, and e-cigarettes. And then the other new item in the updated harmonized dataset is this little section on attempts to quit smoking by switching to e-cigarettes. And finally, the flavored tobacco product section. So, a very detailed set of topics in this harmonized dataset.

So, now I want to walk you through the importance of using the self-response and replicate weights with TUS data. So, there’s two sets of issues that we’re sort of dealing with here. The first is that the harmonized dataset includes a full sample weight for self-respondents. This is the variable name in parentheses, sample weight. And so, this little blurb from the user guide that we are working on or have been working on is that the full sample weights were created to compensate for different selection probabilities, nonresponse, and under-coverage of the target population of U.S. adults. But there's also a second type of weights that we need to think about for TUS. And these are the replicate weights. And so, these 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. So, basically, the importance of these replicate weights are for various estimation processes as you’re considering more advanced analyses. So, depending on the goals of your research project, you may need to think about using replicate weights. So, here's a bit more detail about the challenges, so to speak, in dealing with replicate weights. So, the number of replicate weights used for the TUS-CPS has changed over the different cycles. And so, therefore, the replicate weights have been broken out into three separate files because of those differences. So, when you look at the 1992-93 dataset, there were only 48 replicate weights. That was then expanded to 80 replicate weights in the 1995 through 2003 cycles. And now more recently since 2006 through the current cycle, there are 160 replicate weights. So, the idea then is that the harmonized data, and of course we will be providing very detailed documentation to walk you through this, but the idea for today is to help you navigate and provide you the opportunity, if you have any questions at this time, to kind of walk through this in a bit more detail.

But the idea then is before you go on embarking on running an analysis with harmonized data from TUS, you have to think about whether you would need to use the replicate weights and how to merge that. And so, to do that, you then first need to merge the data from the respondents in the harmonized files with their correspondent replicate weight. And so, what we've done is we've created a RecordID variable in their harmonized file that is unique to every respondent within each survey year and each survey month. And that same RecordID variable was then created for the replicate weight files. And so, those two datasets basically get merged according to these three variables: survey year, survey month, and the RecordID. And so, we are providing you with the SAS code on how to combine these two parts.

And this is not the entire SAS program, but more importantly what I wanted to walk you through is sort of how this process works. So, you see at the top left, you see the dataset from '92-'93, and you see at the bottom that there's 48 replicate weights in the last line of that segment. And then you take the '95 through 2003 dataset, and you now see that there's 80 replicate weights. And finally for the third piece, the 2006 through the 2019 data, you now see that that's grown to 160 replicate weights. So then how do we combine all this? We basically stacked those three files into one single replicate weight file. We sort that dataset. We also sort our own already downloaded harmonized file. And then we combine the two pieces, the harmonized data and the replicate weights by the three variables that we said. And so, that is our analytic dataset that we can then use. So, you're going to call that Harmon. So, one more thing that I wanted to mention about replicate weights that you need to think about when you're running weighted procedures, either in SAS or SUDAAN, is the adjustment factors. So, within TUS, we used Fay's method, which is a variation of the balanced repeated replication method, or BRR, that you may have seen with other surveys, when conducting variance estimation for the TUS-CPS with replicate weights. And so, one example that I wanted to give you here is when you are running a weighted frequency procedure using a single wave of data, if that is done in SAS, obviously you would use the Proc SurveyFreq, and you would set your variance method to BRR and your Fay adjustment factor to 0.5. If you're using SUDAAN software package, then you would use the Proc Crosstab, and you would set the adjustment Fay factor to 4. But the caveat here is if you're analyzing multiple waves of data, then those adjustment factors get slightly tweaked. And for the SAS version of SurveyFreq, the Fay coefficient becomes 0.75 and for SAS that becomes 16. So, just laying all this out so when you see the code, you're not scratching your head as to how this came about. So now let's talk about then how do we actually apply those weights in the harmonized data analysis? So, we're going to take a very simple example. These are very simplified examples that I've included. When we post to the detailed user guide, we'll provide much more complex examples, sometimes comparing two or three different variables, but the idea here is to just keep it simple so that we could walk through it in an easier manner.

So, for example, let's say we were interested in calculating the mean cigarettes per day among overall current smokers in the 2014-2015 cycle, which would include self-respondents only because we are using the harmonized data page. And so, these will be our first steps, download the data from the website, or download the zip file which would contain the data, the SAS program, the formats, the replicate weights, and so on. Then we would download the program also, that would be the zip file, which would be the SAS program that I just walked you through to merge the replicate weights to the harmonized data. And then you would modify as needed your file name, locations, your library names, your include statement so that everything maps to where it's stored on your computer. And so, very, very simple walk-through here. But if then you're interested in a single time point, you're using the Harmon dataset. In your set statement, you're creating a new dataset called Harmon1415, and you're only outputting the observations for that cycle, so that would be the SURWAVE is the variable that determines what we said the flag to see what year we're interested in. And you would set that to 9 because that is the value corresponding to the 2014-15 cycle. But then we have to go through a process of adjusting for the self-response weight and the replicate weights. And we have to divide those by 3 because remember each time point contains three months of data, right? And so, if we don't divide that by 3, we end up with three times what the U.S. population would look like. And so, obviously that would be a very overinflated number. So, the idea here is that we first have to go through an adjustment process of dividing the person weight and the replicate weights by three. So, that's what you see in this little calculation here. And so then we can – once we've done that, our weights are now correct, we can go through the process of creating our constructed variables for cigarette smoking status, cigarettes smoked daily, and create our usable datasets, which we then can use to output a weighted mean with SAS with the Proc SurveyMeans and to the appropriate standard errors because we are using the correct number of replicate weights.

OK, so this is the first example of a single time point. I now want to walk you through what happens when you collect more than one time point. Say you're now interested in two time points, not the entire dataset, just two time points. And so, now we're choosing to calculate the prevalence of overall current smoking during the 2003 and 2006 waves. And so, now we are subsetting a different number of respondents from our entire dataset, right? And we're saying that [inaudible] has to only output the observations for SURWAVE in 6 and 7; 6 and 7 corresponds to the years 2003 and 2006. So, this is our now analytic dataset for two time points. And apologies if this is a little bit small, but I wanted to fit it all in one page, so I will walk you through it. So, previously, we had divided the factor by 3. We now have to divide it by 6 because we have two years of data, and each of those two years has three months of time points. So, three months from each survey. So, again, we have to go through that process of adjusting the weights, both the self-response and the replica weights, to the average size of the U.S. population for those time periods. And so, that's what we're doing here. You see it with this first line here of the sample weight, we're dividing it by 6 so that our new sample weight is correct. And the comment that I've included for your reference here is that now we're facing a situation where the 2003 dataset has 80 replicate weights, but the 2006 through '07 dataset has 160 replicate weights. So, we have to go through a process of expanding or extending the number of replicate weights to 240 so that all survey months have an equal number of replicate weights. And so, that's what we're going to do. These adjustment factors will take care of making sure that everything is equalized over all the time points that we're using. And so, hopefully you see the pattern here that you have to count how many cycles you're using, take into account that each cycle has three time points, and so create those conversion factors based on that. And so, we're going through this series of "do" loops to be sure that we are correctly extending the number of replicate weights to in this case 240 because we have the 2003 data with the 80 replicate weights, and you can refer back to the previous slide that I showed you to see, OK, the cycle fell in this segment of the dataset, so there's 80 replicate weights, but this other dataset now falls into the expanded replicate weights, and so we have to sort of make an adjustment for all that. So, the idea here was to just show you the thought process that you have to go through. Of course, we will support you as much as we can with sample code, and you can always email us with questions if we can be of help also. So, again, we go through once more that similar process of recording our variables, and now we're creating an outputted weighted prevalence estimate for current smoking among self-respondents, and we very nicely have included our VarMethod of BRR. We're now using two multiple – I mean, two data points, right? So, we're now using the adjustment factor where we have multiple points of data. We have our corrected sample weight, right? We divided it by 6 because we have six time points basically with all the months, and three months in each of the two cycles. And we now have correctly adjusted to create 240 replicate weights so that everything is equal.

OK, so the last example is now what happens when we want to use the entire harmonized dataset. So now, we're talking about 1992 through 2019. This is just an example research question. I haven't actually added the code, the sample code for running models, statistic model, but that's something that we will include in the user guide. I think just in the interest of time and to get to the other presentations, I didn't really want to go through that at this time. But the idea here is that when you're using the entire dataset, so now we end up with 29 time points that we have to control for. So, this is what the – and we're going from top left to bottom right, so at the very top, we now have 288 replicate weights. We've divided our sample weight by 29 because the entire harmonized dataset has 29 time points, and the comment in green basically just shows you that the first TUS cycle in '92 has 48 replicate weights, the 1995 through 2003 had 80 replicate weights, and then 2006 onward had 160. So, you add all those up, we end up with 288 replicate weights. And so, these are the adjustment processes that we have to go through for each of the replicate weights. So, once again, this will be all explained in detail in the user guide. But the idea was to walk you through it so you'd have a chance to go through this step by step, and we will be here to support you if you have questions. And I will gladly take questions at the end of the presentation if you have them. I did want to take just a moment to thank all of the contributors to our TUS team, to our TUS work, and I've included the hyperlink at the bottom for you to check out our website and sign up for our mailing list if you haven't. And with that, I will transition to Dr. Seaman, who will be going through the second presentation. Thank you.

DR. ELIZABETH SEAMAN: Thank you so much. Very excited to be here this afternoon. OK, so the "share" button just popped up. OK, so hi, everyone. Thank you so much. I'm really excited to get to talk about our TUS-CPS menthol analysis as an example of using the harmonized dataset and replicate weights. This work was conducted when I was a fellow with NCI. I have since transitioned out to a role with the CDC Foundation managing their e-cig monitoring project, but all of this work was completed when I was a fellow, and my views and opinions do not represent either organization or anyone else other than myself. But I think I'm going to start with a little explanation of our research question and sort of the way we wanted to approach it, and then share a really exciting summary of our findings so far with you all. So, this paper is currently in the manuscript process, and it started as a group of collaborators from NCI and FDA, and I know a few of us have moved since then, so even more organizations are being included. But the work is in the manuscript process, so you'll see in the results I have a few kind of confidential watermarks, and I ask that you don't share these results until hopefully very soon we can disseminate when they are in a manuscript. We'll keep you all updated.

OK, so let's jump right in. So, our group was very interested in understanding how menthol cigarette smoking had changed among current smokers between 2003 and 2014 to '15. And so, the reason we chose that was to maximize the amount of data we could use in the harmonized dataset. 2003 was the first year where the TUS-CPS asked questions about menthol cigarette use, and '14-'15 was the most recent when we started this project, which does sort of date me in the analysis a little bit. So, that gave us four waves of data. And I think the main thing we were looking at was not only affecting the overall weighted prevalence of menthol cigarette smoking in that time, but also being able to adjust our estimates for demographic factors and sort of the power that could come from comparing the weighted, unadjusted and weighted adjusted prevalence estimates. So, very excited to talk about this with you.

So, one of the first things we ran into is that the TUS-CPS changed the way – changed the item they used to assess menthol use in this time period. So, in 2003 and 2006-07, the question was, is your usual cigarette brand menthol or non-menthol? And respondents could say menthol, non-menthol, or no usual type. But then in 2010-2011, the item changed to, do you usually smoke menthol or non-menthol cigarettes? And again, same answer choices. So, I think you can see the construction in the item really was the same, we just dealt with a little bit of changes in the wording. So, we had two waves in our analysis of the older item and two waves with the newer item. Luckily, this is exactly the kind of issue or challenge that the harmonized dataset is designed to help with. So, you see in the harmonized dataset we have one clean, easy-to-use variable, where these two similar items have already been adjusted into one variable, so that definitely made our job easier. So, Carolyn did an excellent job giving an overview and describing the replicate weights and how to use and adjustment, and we did run into this problem or this challenge in our analysis. 2003 only had 80 replicate weights, and then our last three waves, which was in ’06-’07, ’10-’11, and ’14-’15 each had 160 replicate weights. So, because we want to conduct analyses across the years, we needed to create those 240 adjusted replicate weights that Carolyn showed a great example of in her slides.

So, an overview of the way we approach this and sort of the steps we took, so first we harmonized – we formatted the harmonized dataset, and we just output the data we needed. So, just like Carolyn showed you, we just output the waves and the variables that we knew we were going to need. And then we formatted and merged the replicate weights needed, which again we’ve seen some great example code for. We made them for each of the different surveys that we wanted to include. Then we merged the harmonized data and the replicate weights, adjusted the replicate weights, and then we were able to analyze the data. So, as you can see, there’s a few steps to sort of lead up to analyzing the data. But the TUS-CPS is such a rich dataset with so many respondents and so many interesting research questions and the answers that the initial steps in the setup is definitely worth it.

So, here's just a little bit of sample code for merging the harmonized datasets and the weighted datasets. We had created an ID variable to sort of roll survey year and survey month and respondent ID all into one. So, I did this, but I think like with anything with data analysis, there’s multiple ways to approach the same thing. So, I think Carolyn’s code gave a really great example of using the year and the month and the ID. I just put it into one variable, but you can see we sort both datasets by the ID or the variable we’re going to merge on, which in this case is ID, and then we merged by ID if A and B. So, if someone had data in the harmonized dataset and they had replicate weights, we made sure to include them. And this is just a really good way to check and make sure everything’s kind of in the right place, like if you end up with a different number here than you expected, I definitely had to go back a few times and check my code. So, here is – it looks very similar to what Carolyn shared, and this is the way that we adjusted our replicate weights. There’s a few things I want to call out. Again, we’re creating those same 240 weights like Carolyn showed. Because we had four waves of data, each with three months, you’ll notice that we divided our overall sample weight by 12, four waves and three months each. So, that's easy math. And then I think the rest are really closely mirrors the first example. So, for 2003 with 80 weights, you can see that weights 1 through 80 are really the adjusted original replicate weights, and weights 81 to 240 is the self-respondent weight divided by the number of data collection periods. And then for all the subsequent surveys, the first 80 replicate weights are the self-response weights divided by the number of data collection points. And last 160 are the adjusted original replicate weights. So, many thanks to the whole TUS-CPS team for being so helpful in explaining this to me many times.

So, here's a little bit of sample code. And we did have to cut it off because it got to be too much, but just sort of for how we organize our variables and how we did a little bit of recoding. So, those menthol use questions only go to smokers, so we had to do a little bit of coding where we took the cigarette status of were they a former or current cigarette smoker or were they a current smoker, and made that into a kind of indicator variable to see if someone who was a current smoker or who stopped, and then we used both CIGTYPE and CIGSTATS to create this sort of four-level menthol use variable, where we have current smokers whose usual type is menthol, smokers whose current type is not menthol, current smokers with no usual type, and people who are not current smokers, either they're never or they're former. And that was really helpful for when we wanted to look at menthol smoking in the population in all the respondents. But then we also developed a smaller menthol use among current users with a usual type variable. And this is where most of our analyses focused, that we really were interested in diving into current users who self-identify as those with a usual type, and that type is menthol, and seeing how their behavior had changed in the study period. So, we just kind of ended up with a two-level as well as an option where we have a usual or current smoker with a usual type that is menthol and a current smoker with use that is not menthol. And then I think the harmonized dataset is fantastic because all of these kind of socio-demographic or covariants are included in – they're the most granular form they can be, which is amazing, so you're able to sort of recode them in ways that are helpful to you. So, for age, you can see we wanted to do a five-level age category, and I think these are pretty standard categories, but just it's nice to be able to take the harmonized dataset and sort of roll them up into different categories versus being forced to work with such broad categories or sort of not having things line up. And so, we did a very similar thing with the rest of our covariates. We used race and educational attainment and a few others, but this is just a sample of sort of how we got into that.

OK, and here's kind of the main piece of code we used. So, we were able to produce adjusted prevalence estimates by using this proc rlogist, and you can see that we used the overall self-response weight. I should have changed the [inaudible] but same thing. We used our 240 replicate weights. And then as Carolyn said in her presentation, because we are using harmonized data across multiple waves, we used a Fay adjustment of 16 in SUDAAN. And then so you can see we used the predicted marginal statement to output the prevalence estimate adjusted for every output model. So, sex, age, race, employment status, metropolitan status, region, and educational attainment. So, this is our main or our overall menthol cigarette smoking among current smokers, our main finding. So, then we used really similar lines of code to do some things analogous to look at adjusted prevalence estimates of menthol cigarette smoking among current smokers with a usual type for each of our demographic groups of interest. So, we actually got adjusted prevalence estimates for women, adjusting for every other demographic variable, versus men, adjusting for every other demographic variable. But this is sort of the heart and soul of the main code we used for our [inaudible].

OK, so what did we find? Here you can see the weighted prevalence estimate of menthol cigarette use among current users with a regular type adjusted for sex, age, race, employment status, metropolitan status, region, and educational attainment was right around 27.9 percent in 2003. So, of current smokers with a usual type, 28 percent were regular menthol smokers. And we can see that it stayed about the same in 2006-07 right at 28 percent. By 2010-2011, it had jumped to 31.1 percent and by 2014-15, it was up to 33.3 percent. So, while we know that overall cigarette smoking is decreasing in the population, it's really interesting when we look among the subusers of established current smokers with a usual type, we're actually seeing a slight increase in those who report that they're using menthol use. And I think our group had some really great discussions around does this reflect more new smokers are becoming menthol smokers than non-menthol smokers, does this mean most of the people who are quitting are not menthol smokers. The TUS-CPS did the cross-sectional survey at different points in time, so I think there's a lot more work to be done in this field and kind of explore why we see amid declining cigarette smoking rates increasing prevalence of menthol cigarette smoking among current users. And so, here is a preview of some of our findings when we looked across demographic groups. So, adjusting for all other covariants, you can see that females were more likely to report being a menthol smoker than males, and over time, this difference increases. The number of females that report being a current menthol smoker is increasing in a much higher way than the males are. And looking across age groups, we see something similar, where I think the most pronounced growth in menthol cigarette use is among the youngest respondents. We see that for 18- to 24-year-olds it's pretty dramatic, but even for 25- to 34-year-olds there's a bit of growth that more and more are reporting that they have their usual type, and that usual type is menthol, which is very interesting.

And then looking across race and ethnicity, which I know will probably not be surprising, but we can see that non-Hispanic Black smokers are much more likely to report being a menthol smoker, and we can see that it's grown over time. So, a few quick take-home messages from our work is in the face of declining cigarette smoking, we see this increase in menthol smoking among current smokers. The difference is that there really were some very pronounced differences in menthol use among certain demographic subgroups, especially young adults, women, and non-Hispanic Black smokers, who have higher prevalence of menthol use compared to their peers. And we can see that non-Hispanic Black smokers have almost double the prevalence of menthol cigarette smoking compared to any other racial or ethnic group. So, we know none of these trends occur in a vacuum, and it appears that a lot of the progress in declining cigarette smoking has maybe come from gains in menthol cigarette use, which is interesting. So, we really hope this study will be making its way to a journal near you very soon. We hope that future work can help us better understand the best ways to help menthol smokers successfully quit and to reduce menthol smoking initiation. Thank you so much. I will turn it over to Dr. Choi for a discussion of his work.

DR. KELVIN CHOI: Hello. Finally, I found the unmute button. Just a heads up that I'm in an area that has a rainstorm, so my Internet may cut out. And if I suddenly disappear, somebody can share my slides and just advance and people can just see for themselves. Other than that, I should be able to stay online and go through the visit with you all. Thanks for joining us today. And Carolyn has greatly explained how the harmonization effort took place and what we did in making sure the data is user-friendly as much as we can. And then Elizabeth gave a very interesting and good example in terms of how to use the data, looking at some trends over time. One of the powerful way of using the TUS data is really to look at small populations, being that the TUS data have so many participants in each wave and that you can combine waves over time to look at an even bigger sample size. You can actually use it to look at some of the smaller populations and to do analysis on topics that we previously weren't able to do. So, my bad-looking self, the one on the video who is not so good looking. All right, so, a standard disclaimer, basically whatever I say is my opinion and my thoughts, and do not try to call the director of NIH and complain about it. If you have questions, comments, or concerns, you can always email me. My email will be shared with you all at the end of the presentation. And the other thing is that the word "tobacco" in this presentation refers to commercially manufactured, marketed, and distributed tobacco products and to respect our Indigenous brothers and sisters who use tobacco for ceremonial and healing purposes.

All right, and Carolyn has gone through a list of variables that are available in tobacco use supplements, which measured detailed tobacco use behavior. And if you go through the survey and go through the harmonized data, you find a lot of these. For example, this one, have you ever smoked at least 100 cigarettes in your lifetime, one of the standard questions to ask about established smoking. If you have ever looked into the current populations survey technical documentation, which is my favorite document to look at, you'll find that there's actually multiple attachments there, and one of which is attachment six, which shows you the basics CPS record laid out. And in there, you find a list of variables that is actually delivered together with the tobacco use supplement data, not all of which are harmonized into the harmonized dataset, and we're still working through some of those and try to bring more of those core variables into the harmonized dataset. But what we have there now is that we have family income, education attained, metropolitan status, and for those who are interested in rural-urban differences, that will be a variable that you could actually use. Of course, there is race/ethnicity. But the other thing that we have that other surveys don't usually have is the country of birth of the respondent, the mother, and the father. And you know there's another whole host of labor force-related measures that comes with the core measure of the CPS. So, I'm going to talk about how we look at the respondent country of birth and use that variable to answer some of the research questions that we previously haven't been able to answer. The power of the dataset is that we can pull people, participants across waves, as I mentioned before. So, in a single wave, there may be a small number of people that meet a particular definition of a small population, but over time, that population can get bigger as you pull people from across survey waves. So, in this particular instance, we're interested in tobacco use behaviors among U.S. Blacks, depending on what region they were born, whether they were born in the United States or they were born in other regions of the world. And we want to see how tobacco use behavior differs by the global region of origin just within the U.S. Black population. The work was actually led by my former postdoc fellow, and her picture is here, Launick Saint-Fort. So, just a little background before we dive into the analysis. So, 3.8 million of U.S. black individuals are actually foreign-born, living in United States, which is about 8.7 percent of the U.S. Black population in 2013. And they don't just come from a few countries. As a matter of fact, the foreign-born black individuals came from a diverse list of countries. And tobacco use behavior may vary within these U.S. black populations by global origin. It depends on the home country tobacco use prevalence and marketing and the culture. And so, that question has not been previously examined. So, in this particular analysis, we want to explore heterogeneity in tobacco use behaviors among U.S. Black individuals by global region of origin.

So, first thing we need to look at is the survey waves that actually have data that we want. If you have looked into the harmonization documentation, you find that race, the measure of race actually changed over time. Prior to 2003, we had the five categories of race: Black, White, American Indian, Alaskan native, Asian-Pacific Islander, and other. And then in 2003 and on, we had a much more in-depth measure of race in the survey. So, keep that in mind. We want to look at only those who self-identify as Black only. So, we limited our survey waves to only 2006, 2007, 2010, 2011, 2014, and 2015. When pulling these data across waves, we also wanted to be mindful that we don't pull across data that's too collected in too wide of a range of years so that may be a hidden secular trend that we are missing if we just pull them all together. So, we only pulled the data from three different waves. The other thing is the definition of menthol cigarettes, which I'll also just mention that the message changed a little bit, but the good thing is we are able to harmonize that concept without much trouble, so we can use that measure. And for this particular answer, we looked at those who self-identified as Black only. Now, this is the fun part is to define the global region of origin. If you go through the document, you find appendix B of the harmonized data documentation. And this is a truncated list of the countries that are actually in that document. You see some countries actually have a consistent code over time. The United States is always 57. American Samoa is always 60. Guam is always 66, which makes your life much easier. Some, however, do change over time. The Czech Republic, for example, in the earlier years, it was 155, and then later it's 148. So, in dealing with these, we have to make sure that they actually line up, and we're using the same code when we create the global region of origin variable. So, what we ended up doing is to classify people into U.S.-born. That includes those born Guam, Puerto Rico, and U.S. Virgin Islands, and other U.S. island areas. And that's the biggest group, of course. And then we have people who were born in African countries, both North Africa and sub-Saharan Africa. And that we have about 2,000 people. And then we also classified folks who were born in the West Indies, and those there's about 2,000 people. And then we have about 200 who self-reported being born in Europe. And we excluded other regions because, even though the total is about 1,000 people, the other regions have a much smaller sample size per region, and given heterogeneity across those regions, we don't feel like we can just put them all together and call them “other.” That wouldn’t be doing it any justice.

So, in terms of tobacco use measures, I’m not going to go in depth into it, but we look at current cigarette smoking, current cigar smoking among the established smokers who reported smoking more than 100 cigarettes in their lifetime. We looked at whether they became former smokers and also whether they started smoking regularly as minors. And we also looked among the current smokers time to first cigarette less than 30 minutes after waking, and also regular use of menthol cigarettes.

In terms of analysis, Elizabeth and Carolyn have greatly covered the weighting process when using multiple waves and all that, so I’m not going to repeat that. If you still have questions, you can email us or you can just rewind to the previous part of the presentation and give it another listen. In terms of the actual analysis that we do, we use the multivariable logistic regression model, look at each of the tobacco use measure as a separate outcome, and adjusting for demographics in a survey waves.

All right, now to the fun part. Let's see the results.

So, this is looking at current tobacco use by global region of origin among just the U.S. Black only individuals. And as you can see here, U.S.-born Blacks have the highest rate of current cigarette smoking, 17.4 percent. Next to it, about the same, is Europe-born Black, which is 17.7 percent. When you look at African-born and West Indies-born, even though they are Black, they have much lower prevalence of smoking, 4.7 percent and 4.9 percent, and these two groups have significantly lower prevalence of current smoking when you compare to the U.S.-born Black individuals. When we look at cigar use, we have a similar pattern, although the prevalence of users is much lower in the U.S. Black population. It’s about 3 percent current use in the U.S.-born Blacks and about 3 percent for the Europe-born Blacks, but it’s 1 percent or less for the African-born and West Indian-born Black. And again, those who were born in Africa and born in West Indies also show they have significantly lower prevalence of current cigar use compared to those who were born in the United States. When we look at established smokers, the first thing that we look at is how likely that they could become former smokers. And we saw that among the U.S.-born Blacks, about 40 percent of them who ever became established smokers subsequently became former smokers. And that number was actually lower in the Europe-Blacks but higher in African-born and West Indies-born Black individuals, so much so that those who were born in Africa or born in the West Indies are significantly more likely than those U.S.-born Blacks to become former smokers if they ever became an established smoker. And then we also looked at starting smoking as a minor, which we know is the earliest smoking initiation is related to having a harder time to quit smoking, so that may be some of the explanation why we see that quitting smoking is lower in U.S.-born Blacks versus those who were foreign-born. And here we see that not all foreign-born Blacks actually have the same prevalence of starting smoking as a minor. There’s no difference between U.S.-born versus Europe-born Blacks, but those who are African-born are actually less likely to start smoking as a minor. Again, no difference between West Indies-born and the U.S.-born Blacks. So, starting as a minor, less likely to start as a minor, it may be favorable for Africa-born Blacks to become former smokers, but not the other groups. When we looked among current smokers and examined times of first cigarette, which is a measure of dependence, we also found those who are African-born are significantly less likely to smoke their first cigarette within the first 30 minutes of waking in the United States. It signaled that these African-born individuals are probably less dependent on cigarettes or nicotine and therefore may have an easier time to quit smoking. Finally, we looked at regular menthol cigarette use. And here, it shows that the African-born and the West Indies-born Black individuals as compared to the

U.S.-born Black individuals are actually less likely, significantly less likely to use menthol cigarettes regularly. And we do know that menthol cigarette use is associated with a harder time of quitting smoking, so that may also be the reason why the U.S.-born Black individuals are particularly less likely to become former smokers, had they ever become established smokers. So, in conclusion, the bottom line is not all Black individuals are the same. It's not a monolithic group. Even though a vast majority of them were born in the United States, there are individuals in the Black population who were foreign-born and have quite different tobacco use behavior. So, we have to take this into account when we think about surveillance and interventions in that regard. Another thing to take home is that TUS-CPS is really a powerful data source you can use to pull the data across waves and create a sufficient sample size to answer research questions that you have a hard time finding other datasets that have a big enough number size to do those analyses. So, this is a really nice way to, as Carolyn mentioned, to study health disparities and tobacco-use disparity because the sample size is there, and you can actually do a fairly well-powered analysis in those situations. So, thank you for listening, and again, my email is at the bottom of this slide. If you have any questions, you want to complain, or whatever, you can email me. I'm going to stop sharing.

MS. CORCY: Thank you so much to Drs. Carolyn, Elizabeth, and Kelvin for your great presentations. At this time, we wanted to open the floor for our Q&A session. Just before we jump into that, I did want to mention again for those of you who may have joined a little late that we are recording this session and we will be posting all the recording and slides online in about three weeks. Everyone who registered for the webinar will receive that email notification, and we will also be including links in that email to where the materials are posted online. So, at this time, I want to open up the floor for any questions for the researchers. You can feel free to type into the chat box or you can use the "raise the hand" feature. So, if you take a look at your WebEx, you should see a participants panel. If not, please open it. There's a button on the bottom right to open the participants panel. And at the bottom right of the panel, you will see a little hand icon. If you click it once, it will raise your hand, letting me know that you wish to be unmuted so you can ask your question. And if you click it a second time, it will lower your hand. So, if anyone has any questions, please let us know.

OK, so I see one question in the chat box that states: "Very important research. Can you explain methods used in harmonized dataset to enable one variable with more levels in latter years as opposed or as compared to early years?"

DR. REYES-GUZMAN: Yeah, I can take that question. So, I think that was a great example of the race variable that you saw. And the way we, for the time being, we've gotten around that question is, when the categorization changed, that served as our cutpoint. So, from the 1992 through I believe it was the cycle before 2003, the categorization changes, so that became sort of one version of the variable, and then after the following time point, then it became sort of a different version of the variable. So, I think that's been kind of our workaround for the time being. We are planning on tackling this question further with imputations later this fall, but I think for the time being, this is kind of the best that we can do. Kelvin, I don't know if there's anything else you want to add to with respect to that.

DR. CHOI: Yeah, I'll add a little bit. I think the thought behind harmonization is that there are two types of questions that we need to deal with when it comes to harmonization. One is the questions remain the same but the options change. And in those situations, we try to retain as much information as possible. And that's why instead of creating one race/ethnicity variable with fewer categories to match it to the three, 2003 question options, we ought to have the two sets available so that people who are interested in actually a more fine-grained race/ethnicity variable, so actually they can use that more fine-grained measure to look at, for example, multi-race individuals. And so, we ought to have a more – to retain as much data as possible, and that's the thought going into this. There are times that we cannot. The questions may have changed, and then the options may have changed in a way that we would have to make the harmonized variable usable. Sometimes we'll have to make decision on maybe reduce a few options so that the answer's actually consistent over time. Those are tough questions that we spend weeks and months to mull over because we do want to provide as many data to the users as possible.

DR. REYES-GUZMAN: Yeah, thank you, Kelvin. And I would say that in those types of scenarios that Kelvin was mentioning, what we ended up doing is we did a very thorough job of documenting our thought process. So, if you actually take a look at the crosswalk of our harmonized datasets, then those question contains very detailed explanations of the type of compromises, when those situations came up, the type of compromises that we had to make. So, everything is well-documented so that there's a lot of transparency in terms of how those decisions were made. Yes. Great question.

MS. CORCY: So, we actually have a follow-up from the person who asked that question, and I don't know if you can answer it now or not, Carolyn and Kelvin. So, the question is, "So does the one with more levels contain missing values for those prior years' observations?"

DR. REYES-GUZMAN: That's a good question, and maybe I will ask if Todd Gibson is still on. Can you take this question, Todd?

DR. CHOI: Not necessarily.

DR. REYES-GUZMAN: Oh, Kelvin, OK, if you know the answer to that, please chime in.

DR. CHOI: Not necessarily missing data. So, for example, race/ethnicity, there's not necessarily missing data. It's just that the options available for the participants are just different, so we don't have missing data in that regard. In some cases, certain universe may not have been – sorry, I said universe. Sometimes the inclusion criteria for a particular question might have changed over time. And in those situations, we usually documented how that changed, and we try to maintain the same inclusions within criteria for those questions. We try to minimize as much missing data as possible. And in the work that we do, and Todd in particular was the one put together the dataset, it's clearly documented where the missing data are coming from. Is it because they don't answer the question? Is it because it's not part of the inclusion or exclusion criteria? So, a lot of that information will be available in the data documentation.

MR. TODD GIBSON: Hi, this is Todd from IMS. If I can, if you don't mind, I can chime in a little bit, specifically about the race. I think, Carolyn and Kelvin, you – wonderful job talking about this and explaining it. As Kelvin said, it's not really that it's missing. It's just for some survey waves, for example, we use race because it's kind of a hot topic. For certain years, there's a specific race variable that goes through the 2002 data that has the five categories. And then once we started dealing with, and we deal with this in other products, where people were allowed to report multiple races, the CPS itself changed also. So, in 2003 when they could report multiple races, we had this separate variable so that we can keep that detail in the later data, even though we didn't have it in the early survey waves. With the harmonized file, we've gone through this twice, the process twice of adding variables and adding data. On our to do list is in the future to, especially for race, is to create an imputed race or a bridged race so that you will have a consistent variable across time. Now, we will keep the original variables, but we'll have this consistent variable that could be used in analysis that bridges those race groups with multiple race into a single race.

DR. REYES-GUZMAN: Thank you, Todd, for chiming in. That's very helpful. And I think in some ways, maybe that's a good segue way to the next question that's in the chat. Nalini, I don't mind jumping into that one. So, I think the question was with respect to county-level estimates. "For researchers interested in county-level estimates, could the harmonized data improve their research or analysis?" So, I think definitely that's one of the strengths of the TUS in terms of being able to drill deeper, the caveat being that county-level estimates are not available for all regions of the U.S. So, I think with that caveat in mind, you may not find – even using the harmonized variables, you may not find it for all regions of the U.S. So, keeping that in mind, I think, if you have specific sort of larger metropolitan areas in mind as your research question, as part of your research question, then definitely the harmonized data would be helpful. So, yeah, I don't know if there's anything that anyone wants to add to that response.

DR. CHOI: Carolyn, are we going to have one of the upcoming workshops that talk about small area estimations?

DR. REYES-GUZMAN: Yes, there is, actually. I believe, if I'm not mistaken, that may be later this fall. But there is going to be a whole seminar/webinar on small-area estimation with two of our statisticians at NCI, so it will be Benmei Liu in the Surveillance Research Program and Anne Hartman in the Tobacco Control Research Branch who will be tackling examples of smaller estimations using TUS data. So, if you are interested in that, please sign up for that.

MR. GIBSON: That will be October 14.

DR. REYES-GUZMAN: Say that one more time, Todd?

MR. GIBSON: Sorry, that would be the webinar on October 14.

DR. REYES-GUZMAN: On October 14. Perfect. Thank you. Yes. Perfect. OK, Nalini, I'll toss the microphone back to you.

MS. CORCY: Thank you. So, we do have one other question. "Do you know if you can import TUS-CPS into GIS platforms?"

DR. REYES-GUZMAN: I believe the answer to that is yes. It's not something that I'm familiar with, but it's something that I can look into that question with our Surveillance Research Program at NCI. There's a couple of people who are GIS specialists that I could pose that question to if that's something of interest. And if that's the case, if whoever posed that question, can you please follow up with us by email to any of us at NCI, and we're happy to help you navigate through that.

MS. CORCY: Great. We do have one more question. "Any criteria for suppressing estimates?" Are there any criteria?

DR. REYES-GUZMAN: Yeah, that's an interesting question, actually. It's one of the questions that we've been grappling with that we're actually planning to also tackle later this year. I think we are in favor of staying away from supressing estimates for the reliability issues that that raises. And we're actually going to be working on some documentation to put out, so stay tuned. But as of now, we are not making any recommendation on cutoff values.

MS. CORCY: Great. So, those are all the questions I received so far. If anyone else has any, please feel free to type in the chat box, or you can raise your hand, and I will unmute you so you can speak up.

OK, we have another question. “What is the lowest geographic identifier available in the publicly available data? And what are the sample sizes of the annual survey data?”

DR. REYES-GUZMAN: OK, so the smallest geographical identifiable unit is county level, so I don’t think it goes finer than that because I think we’d be going into violations of identifiable information per the Census Bureau regulations, so I believe that’s the smallest unit. And then the second part of the question, sorry, was?

MS. CORCY: “What are the sample sizes of the annual survey data?”

DR. REYES-GUZMAN: Oh, the sample sizes, I think an approximate number, I believe it is about 170,000, give or take – I think some years it’s been up to 180,000, sort of the lower end maybe 160,000, but I think about an average would be 170,000.

MS. CORCY: OK, does anyone else have any questions? We still have a couple of minutes, so again, feel free to ask to be unmuted or type into the chat box. I don't see anything else coming through the chat box or any requests. Oh, OK, just got one. “If we attempt to merge with the food security supplement, for example the December CPS, would that be feasible?”

DR. REYES-GUZMAN: Yes. That would be feasible. That is going to be covered in the – so, not necessarily the Food Insecurity Supplement. There will be other examples, one with the American Time Use survey, another with the March Supplement, the ASEC, and another example of the NLMS, and that will be in one of the subsequent webinars. But I think the process is similar in terms of the calculation that needs to happen is to determine what is the overlap sample between – because of the panel survey methodology that the CPS follows, right? You have to see how far away – so, for every month from the CPS, right, so say CPS happens at month 1, every month after that that you go further away from that U.S. Supplement, you lose a quarter of the sample. So, as you get closer to month 4, you’ve basically lost three-quarters of that overlap sample. So, what you have to figure out is when the Supplement that you’re interested happens and how much you end up with overlap with TUS to determine if you have to recalculate your self-respondent weights, because that would be the issue. If the overlap is – so, for example, one question that we’ve looked at with a project that I’m currently working on with the TUS-linked data to the March Supplement, the overlap is very high, so it’s almost like I think close to 90 percent. So, in the end, we were able to statistically justify that we didn’t have to recalculate the sample weight. We were just able to keep the TUS survey weight as the main weight of analysis. But if you potentially don’t have that same overlap of using a different supplement, then you may end up having to do a recalculation, which is not overly complex. It’s just an additional step. So, I think it would – if you want some help and guidance with that, we can definitely work with Census to help us clarify those questions in terms of the overlap, but we’ll be glad to help you navigate through that.

DR. CHOI: I just want to add to that. There actually has been a publication came out several years ago and specifically looking at tobacco use and food insecurity using the CPS-TUS data. That wasn’t performed linking the harmonized dataset with the Food Insecurity Supplement. That didn’t exist at that time. The harmonized dataset didn’t exist at that time. But they now certainly can be done.

DR. REYES-GUZMAN: Yes. Great. Thank you, Kelvin.

MS. CORCY: OK, last call for any questions for our presenters. OK, if not, I think we can go to our final slide. So, I just want to thank everyone for your participation. Thank you so much for joining us today. And once more, thank you to our speakers, Drs. Carolyn, Elizabeth, and Kelvin. We really value your feedback, so we would love it if everyone could complete a brief survey. I have just dropped the link to the survey in the chat box. We will also be sending a follow-up email, and that will have the link to the survey as well. If you are interested in learning more about the TUS-CPS, please feel free to visit our website. You can also subscribe to our emails, which we send out periodically. And if you have any questions that you think of after the webinar, please feel free to reach out to our team. Our email is at the bottom. Once again, all of the materials will be posted online, and we will be sending an email notification with links once they are available. So again, thank you so much to everyone, and we hope you have a great holiday weekend.

[event concluded]