Cyberseminar Transcript

Date: April 1, 2019

Series: VIReC Database and Methods Seminar

Session: Assessing Race and Ethnicity in VA Data

Presenter: Maria Mor, PhD

*This is an unedited transcript of this session. As such, it may contain omissions or errors due to sound quality or misinterpretation. For clarification or verification of any points in the transcript, please refer to the audio version posted at* [http://www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm](file:///C:\Users\Kathy\Desktop\VA%20trans%20-%20in%20process%20&%20DONE\l)

Hira: Hi, everyone, and welcome to Database and Methods, a Cyberseminar series hosted by VIReC, the VA Information Resource Center. Thank you to CIDER for providing technical and promotional support. So Database and Methods is one of VIReC’s core Cyberseminar series, and it really focuses on helping VA researchers access and use VA databases.

This next slide shows the series schedule for the year so far. Sessions are typically held on the first Monday of every month at 1 PM Eastern, except for the next session in May, which will be held at 12 PM Eastern. You can find more information about this series and other VIReC Cyberseminars on VIReC’s website, and you can view past sessions on HSR&D’s VIReC Cyberseminar archive.

A quick reminder to anyone just joining us. Slides are available to download. This is a sample screenshot an email you should have received today before the session, and in it you will find a link to download the slides.

Today’s session is titled, Assessing Race and Ethnicity in VA Data, and it will be presented by Dr. Maria Mor. Maria is co-director of the Biostatistics, Informatics, and Computing Core for the Pittsburgh site of the VA Center for Health Equity Research and Promotion. As a collaborative statistician there, she works with investigators on a variety of health services research projects focused on understanding and improving the quality and equity of health and healthcare for vulnerable populations of Veterans. Thanks for joining us today, Maria.

Dr. Maria Mor: Thank you for that introduction. So for this session, we’re going to talk about the VA race and ethnicity data, and at the end of the session you should be able to know where to locate race and ethnicity in both VA and Medicare data, assess the quality of VA race and ethnicity data, and also create SQL code to help you in extracting those data elements.

First, I’ll provide a brief introduction, discuss locating race and ethnicity in the VA data, then in Medicare data, briefly discuss the quality of the VA race and ethnicity data, provide some simple examples using SQL, and then also some recommendations on how to address the data quality issues in your particular projects, and then where to go for more help.

Before we get started, I do have two poll questions. The first one is: What is your role in research and/or quality improvement? Are you an investigator, PI, or co-PI; a data manager, analyst, or programmer; a project coordinator; or other. And if you respond other, please describe via the Q&A function.

Heidi: And those responses are coming in. We’ll give everyone a few more moments to respond before we close the poll and go through the results. And it looks like we’ve slowed down here, so I’m going to close this out.

And what we’re seeing is 20% of the audience saying investigator, PI, or co-I; 37% data manager, analyst, or programmer; 23% project coordinator; and 20% other. And in that other category we have doctoral student, statistician, physician, and compliance. Thank you, everyone.

Dr. Maria Mor: And then one just second quick poll question before we continue. Have you ever used VA race and ethnicity data? Yes or no.

Heidi: And again, we’ll give everyone just a few moments to respond before we close the poll out. And it looks like we’ve slowed down here, so I’m going to close this. And what we’re seeing is 49% of the audience saying yes, and 51% saying no. Thank you, everyone.

Dr. Maria Mor: All right. So racial and ethnic disparities in health and healthcare persist in both the U.S. and in VA. So within U.S., the U.S. as a whole, disparities and measures for access and quality vary by racial ethnic groups. We have seen improved trends with disparities in some measures, with improvement predominantly for Blacks and some for Asians and Hispanics, but most disparities have not changed or have worsened for other racial or other ethnic groups. Within VA, racial and ethnic disparities persist even though financial barriers to receiving care are minimized within the VA compared to the private sector, and although quality has improved, we still see significant within facility disparities observed in some clinical outcomes.

So more research is needed to detect, understand, and address disparities in health and healthcare, and for that research, accurate race and ethnicity data are essential. The problems that we have with race and ethnicity data in VA, I would say the number one problem is incomplete data. We don’t have data on everyone. We can also have inaccuracies in the data, or also inconsistent data over time.

The overall racial and ethnic distribution of Veterans as a whole across the country is largely white. About 77% are White, about 12% Black, 7% Hispanic, and less than 2% each fall in the categories of Asian, two or more races, or American Indian or Alaskan Native. However, use of VA healthcare differs by race and service connection. So the distribution that we’ll see amongst those who actually utilize the VA and our roles is a little bit different, because Asian Veterans are less likely to utilize VA services, whereas Black, American Indian, and Alaskan Natives are more likely to use VA services. So we should expect to see slightly higher proportions of those groups.

We do have standard methodology for the collection of VA race and ethnicity data. Ethnicity is to be collected and it consists of Spanish, Hispanic, or Latino would all be examples of Hispanic ethnicity. And for race, more than one race may be selected amongst five standard race categories, which we’ll see again throughout this presentation. And those groups are: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or other Pacific Islander, and White. There also could be an option for the patient to state that they do not know what their race is. And our current reporting methods are to use a two-question format, so ethnicity and race would be separate questions, and also to try to obtain the data through self-report, which is directly from the patients themselves.

So for the acquisition of the race and ethnicity data that we have in VA, the data are acquired. Our preferred method is directly from the patient, that would be self-report. We can also obtain information from a proxy, such as a family member who comes in with a Veteran or also completes the paperwork for them. Or there’s also the potential the data is completed at the time that a Veteran comes in to the VA. An enrollment clerk or a coordinator could be the one that enters that information based on their perception of the Veteran. The data are acquired at the time of application for health benefits, or, there’s also the opportunity to obtain the information every time the patient comes in for an inpatient stay or an outpatient visit to a facility, and the data are entered directly into VistA.

So next we’ll turn to where we find the race and ethnicity data in VA. And this leads to the last poll question, is: What sources of VA race and ethnicity data have you used? And please check all that apply. So never used race/ethnicity data; CDW; OMOP; MedSAS files; Department of Defense data such as through VADIR or DaVINCI, those are sharing agreements that we have; or other VA data sources.

Heidi: And I actually ran out of options for the other VA data sources.

Dr. Maria Mor: Oh, okay.

Heidi: But if you do use one of those other VA data sources, just don’t check it off, just send it in to the Q&A screen and let us know what else to use, and we would love to hear that. So I will—responses are coming in on the survey here. We’ll let it churn for a few more moments here and we will go through the results in just a few seconds. Yeah, it only lets me put five options for poll questions. It’s very limiting.

Dr. Maria Mor: Oh, okay.

Heidi: It’s overly restrictive. I feel like there should be a lot more options available, but they don’t agree with me. Okay, it looks like we’ve slowed down here, so I’m going to close this out.

And what we’re seeing is 39% have never used race or ethnicity data; 56% have used CDW; 7% OMOP, 16% MedSAS files, 7% DoD, and we did get a write-in on VistA. Thank you, everyone.

Dr. Maria Mor: Okay. All right. That is helpful. So the first thing I’m going to talk about, the first source I’m going to talk about are going to be the MedSAS data. Partially because the data that we have collected in the MedSAS data is actually a little more standardized than what we can see amongst some of our CDW data. So it just helps us to understand a little bit some of the data that we have collected and the manor in which it’s collected.

So prior to fiscal year 2003, and you might hear me refer to this as old data collection methods, race and ethnicity were captured jointly in a single variable, and in the MedSAS data, that stores the variable called race. And a single value allowed for the joint capture of race and ethnicity. After fiscal year 2003, we have newer data collection methods, and these are going to match the categories that I read to you from our handbook that we now have, for the collection of these data. So we have multiple race values and they are captured in multiple variables in MedSAS called RACE1 through RACE7. The structure of the data in our other data sources is a little bit different. And then we have a single value for ethnicity captured in a variable called ETHNIC. In MedSAS, the way the data are standardized is all of the race variables in ETHNIC have a length of two characters. The first character denotes the race or ethnicity, and the second character has the method of data collection. Now, this only applies to the newer data collection methods. From our older data we did not also include the information about how the data were ascertained. And within the MedSAS file we have data on race at each inpatient stay or each outpatient visit or encounter.

So within the Medical SAS data, the combined values for race and ethnicity will provide information on both only if the individual is White or Black. So we have the categories of Hispanic/White and Hispanic/Black. The categories of Black and White are assumed to be non-Hispanic Black and non-Hispanic White. For American Indian, there is no option to offer specified ethnicity, and then we also have one category for Asian. This is also the category that we would include the Native Hawaiian and Pacific Islanders, which are a separate category under the newer data collection methods.

And then for RACE1-RACE7, we would have our five standard race categories, with the additional options for the patient to refuse to answer the question or to state that the answer is unknown.

And similarly, with ethnicities, the option to be Hispanic or Latino, not Hispanic or Latino, declined to answer, and unknown.

And then finally, the common data element that we have for capturing the data collection method would allow for observer, so that would be say an enrollment clerk or coordinator, a proxy, such as a family member, self-identified, or for the patient to state that they are unknown as to their race or ethnicity.

Within the Corporate Data Warehouse, it’s a national repository of data from the VistA Patient File. It contains data from October 1999 to present. The structure of the data is a little bit different from the SAS data. The SAS data we had one record for each essentially encounter with the healthcare system. With the CDW we have one demographic record for each VA station a Veteran has visited. We have standard and nonstandard race values. And the racial data are available in the PatSub.PatientRace table. We have two sets of variables; Race, which contains the newer collection standards; and LegacyRace, which contains data under the older collection standards. If you wanted all available race data on an individual, you would have to use both of these sets of variables. In addition, in the MedSAS, we saw that we had multiple variables in order to capture multiple races. So for example, if a Veteran identified as being both White and American Indian, there would be two different variables that would contain that information. With the CDW, the data is what’s a long format. So if a Veteran came into a station and identified multiple races, then they would have multiple records in that demographic record. One for each race that they identified.

When talking about the CDW race data, it’s important to note that there have been some changes to the way the data are stored. In particular, the way that what is now the LegacyRace, these older race data values are stored, have actually changed multiple times over the course of the history of the CDW. So as of January 2018, all of the available documentation that we had that was specific to race and ethnicity—and you’ll see me refer to them throughout this presentation—were not based on the current data structure. With the advent of the Patient 3.0 Domain, the data structure has changed again. If you look in the VIReC Factbook for that domain, it will document the current data structure for the race and ethnicity data. And there’s also—in time, there have been some business rules for extraction that may have also led to some differences. And once that happens, again, if you’re using the data now, you won’t see the impact of that. but there could have been a potential in the past if you had pulled some data and then re-extracted it over the same time period, you might get some different results.

So in particular, if you are looking at some of these older documentations for race and ethnicity, they may refer to the older race collection methods as being stored in other locations. For the Patient 2.0 Domain, that information was contained in the Patient or S.Patient tables, and there was a RaceSID variable that would link to Dim.Race to obtain the race, the old race values for those individuals. And then prior to that, the data had actually—also was back in the PatientRace table. Currently, that information is in the LegacyRace variable, or the LegacyRace SID, but previously, all values—the old values and the new values—were contained in the variable Race. However, because there was no data collection method associated with the older race values, the collection method was null for those values. And that was how you could differentiate between the two.

So all race data are contained in the PatSub.PatientRace table. Data are at the level of the Patient/STA3N level, and it contains the most recent data available for the patient. If the patient identified more than one race, there will be multiple records at the patient STA3N level. That doesn’t mean that we have conflicting information on the patient. If it’s at that same patient/station combination, it means the patient identified more than one race. Not that they came in one day and they were—you know, identified as White and another day they were identified as American Indian. It means that they self-identified as belonging to both of those racial categories. If there’s newer data, it will overwrite the older data, not augment it. The collection method contains the method of data collection for race. And then LegacyRace contains the data from the older data collection values. It does not allow for multiple races, so there’s only a single value for LegacyRace at the Patient/STA3N level, however, because there can be multiple records if the patient identified more than one race, there may be multiple—for each one of them, that value of LegacyRace would be duplicated. So you could have duplicate values from LegacyRace, because it’s being added to a table that already allows for duplicate values. However, most patients have values of missing, indicating that there is no older race data available for them.

So within the CDW, we have non-standard race values. This is predominantly an issue with the older LegacyRace data, both in terms of what was standard values at the time that those data elements were collected are not the standards that we have now. And in addition, there were some inconsistencies in how the text to describe those races were used throughout the system. So for example, we can see in this first column here, this group, they’ve abbreviated Amer instead of American. Here they’ve written out American, but they might have a slash instead of “or” between American Indian and Alaskan Native, or Alaskan Native. So these slight differences in text do not map to the exact same standard text, but we can easily identify that this person belongs to the American Indian or Alaskan Native category. And similar with the other examples that are provided. And these, I believe the—these are all from the Best Practices Guide: Race Data Quality Report, and here’s the link. Because I’m going to show you some more examples from that.

So, as I said, the older race values very rarely used standard values. The newer race values are very rarely non-standard. Almost all of them are standard with the exception of really one site is responsible for a category that’s a non-Hispanic White. But other than that, most of the values are standard.

But if you are using the older race data, you will need to identify the standard race that goes with the text for each of these values, and note that five of them do not map to current standard values. These include combinations that include Asian and Pacific Islander, or Pacific Islander within the group. And then we have also Mexican American, and unknown. As of the time of that guide book in 2012, nearly 5% of data fell into one of those five categories, and looking at the data now, almost 20% of non-missing LegacyRace values fall into one of these categories, and the vast majority of them are listed as unknown. So there’s nothing that we could do ever to categorize those individuals, but there are 3% that indicate Asian or Pacific Islander race. So if you’re using the data and you are using it in a way such that Asian and Pacific Islanders will be in the same category, then those data would still be usable to you, even though they don’t map to standard values.

And then for multiple race values in the CDW. There was a CDW Race Data and Multiple Races Data Quality Report, and when they looked at the data, they found that almost 2% of patients who linked to a standard race had more than one standard race, and that was back in 2013. They determined it’s not possible to identify the most recent record for a patient. So they had a recommendation for multiple values. And when we are talking about here conflicting information, again, we’re talking about across sites. If you have multiple races identified within a site, it means the individual is multi-racial. But their recommendation for multiple values was to only use self-identified races if they were recorded, and then if you didn’t have self-identified race, you would use all recorded races for the individual. You could also have the potential for conflict within self-identified as well, if they had self-identified a different racial group at different sites.

So for the ethnicity data in the CDW, it’s found in two tables. Our newer data are found in the PatSub.PatientEthnicity table, and they contain the values denoting Hispanic or Latino, or not Hispanic or Latino. And then if you’re looking for data that’s collected under the old race collection method, it’s going to be in the PatientRace table, because it will be included with the LegacyRace variable, or rarely, with that current race variable, because we do know that there are some indications of the non-Hispanic White. So, categorizations of race that include indication of either Hispanic ethnicity, such as Hispanic White or Hispanic Black, or Non-Hispanic ethnicity, such as Non-Hispanic White or Non-Hispanic Black would also indicate ethnicity. However, not all of these values actually do indicate whether or not the patient was Non-Hispanic or Hispanic. For example, Asian. We would never collect ethnicity in addition to Asian race under the old collection methods, and even for Black race, that may be indicated without noting whether or not it’s Hispanic or Non-Hispanic.

We also have data available to us from within OMOP, and I see that some people are using those data elements. So again, the underlying data source is still going to be based upon our CDW data, but the data are going to be placed into a common data model to map and standardize the data, and business rules will have already been applied in order to take the data and create a value for race and ethnicity for each person. These data are contained in the OMOPV5.Person table. And they contain one standard value for race and ethnicity for each PERSON\_ID, which is the identifier used within the OMOP data. And it can link back to the CDW identifiers, and they do note in the documentation that sometimes there can be some additional issues with those links, so if you have any issues, I would refer back to the OMOP documentation. I do note that they exclude non-Veterans, test patients, and possible test patients.

So the race data in the OMOP is really based upon two documents that I believe I already referred to. There is the race Race Data and Multiple Races Report, and also this VIReC Researcher’s Notebook, which is, Using SQL to Sort Out Race in CDW. So the source data will be the underlying CDW tables that we’ve discussed. I believe when I looked at the documentation last, it referred back to the way the data were stored under the Patient2.0 Domain, where the LegacyRace data was contained in the SPatient table. They allow six categories for race. So the five standard race categories, plus unknown.

And the business rules that they apply in OMOP are essentially to identify the records as being self-report or non-self-report, and they count the distinct values. And they select the most frequently occurring self-reported race. If there is no frequently—most frequently occurring self-reported race, such as no data, or there’s a tie, then they’ll select the most frequently non- self-reported race. And then there are various rules that they will also follow if basically those steps do not result in a value for race and ethnicity.

And for the ethnicity, they follow a similar trend. The three categories are Hispanic or Latino, not Hispanic or Latino, or unknown. Then they use, again, the same logic. They are going to give emphasis to the newer data collection methods and the self-reported data, and then only use data from the older data collection methods, when they don’t have data from the newer methods available.

And then I just wanted to also mention that there’s also now race data available in DaVINCI, which is the joint DoD and VA data repository. There are two variables available. There is a race variable, and then there’s a combined race/ethnicity variable available. I have provided links to look at the information that they have in VINCI Central for using these data. There’s actually not a whole lot there. I will say in that documentation, they do refer to that there could be potential issues with how reliable these data elements are, but I also did not see there very comprehensive information on how to obtain access to the data. So the second link I have provided is to a Cyberseminar that was given by Scott Duvall, which then outlined that process for obtaining access.

So we can also have data that’s available to us on race and ethnicity from Medicare and Medicaid. The easiest source for us to access is from the VA Vital Status File. It does require that you have to sign an additional form in terms of your agreement to using the data. The VA Vital Status File, if you are familiar with it, is really two files. It joins data from multiple data sources, and there can be conflicts in the linking variables that they use, the Social Security Number, date of birth, and gender. And so there’s a Master File, which contains potentially multiple records per Social Security Number that contains these different combinations, where not all of the data elements agree. Maybe the Social Security Number agreed but the date of birth is off slightly. There is also a Mini File that is only one record per Social Security Number, but the Medicare Race Data is only available in the Master File, so you want to be aware that you will request access to the correct file. Data are also—if you go through the process of applying to use the VA Medicare and Medicaid Data—data for Medicare are available in the Denominator file for Medicare. It contains a variable race, which is the same as what we have in the VA Vital Status File. And there’s also an imputed race variable that we will discuss called RTI\_RACE. For the VA Medicaid Data, there’s data available from the enrollment file, and what I’ve listed here is there is a summary combination race and ethnicity variable, but there’s also individual variables available as well.

So for Medicare, it’s a potentially useful source of data for Veterans who are enrolled in Medicare. Which generally means that they are age 65 and older. So among that group that are age 65 and older, more than 95% tend to have data available through Medicare on race and ethnicity. For the younger Veterans, only about 20% have data available through Medicare, and these are predominantly people who enrolled due to disability or because they have been diagnosed with end-stage renal disease. The data do come from the Social Security Administration, and they are obtained generally at the time of application for a Social Security Number. It’s usually from either the individual or their family member, and because many Social Security Numbers are obtained for children, it’s usually going to be a family member that’s obtaining that information. There are some important distinctions from our current VA race and ethnicity data. In particular, Hispanic is a separate race category, and there is no option to select multiple races.

Until 1980, there were only four categories that were collected, which included an unknown. So really three usable categories, and this included this other group, which in 1980 was split out to include the Asian and Pacific Islander groups, Hispanic, and American Indian or Alaskan Native. But for many of our Veterans, we may only have data that’s collected based on these original categories.

And then as I mentioned before, there’s also an imputed race variable. Research Triangle Institute created and implemented an algorithm to increase the accuracy of the race variable, especially for Hispanic and Asian individuals, which as we saw previously, may have only been coded as other. Their algorithm uses the first name, last name, preferred language, and place of residence, and they improved the sensitivity in the racial codes and it increased from 30% to 77% for Hispanics, and from 55% to 80% for Asian and Pacific Islanders.

In terms of the data quality, as we’ve previously mentioned, data for many enrollees is only limited to these original four categories, and that it’s very difficult to tease out both race and ethnicity, because Hispanic is a separate racial category and there is no option for multiple race reporting. There have been initiatives to improve the data quality, including updates from the

Indian Health Services, and also a 1997 survey of enrollees who were classified as other, or unknown, or who had a Spanish surname. And then also this imputed race variable that we discussed.

So Medicaid contains an overall combination of race and ethnicity variable, and we can see this information on both race and ethnicity and multiple races. So we have Hispanic or Latino, no race information, Hispanic or Latino one or more races, or even if they are Non-Hispanic more than one race.

If this information isn’t sufficient for use, they do have the individual codes, so you could obtain ethnicity separately from race and then multiple race codes for those who identified as having more than one race.

An additional issue with the Medicaid data is that it does lag behind both VA and Medicare data. So Medicare data tends to lag a year or two behind the VA data, maybe about a year or so, and then for the Medicaid data, it’s going to lag another year, probably closer to two, I think, behind the Medicare data. We also have fewer enrollees in Medicaid, so we have more like about 10% of Veterans who are enrolled in Medicaid, which is considerably lower than what we saw in Medicare. And there have also been some changes to the data collection over time as well with the Medicaid data.

So finally, we’re going to talk a little bit about the quality of the VA race and ethnicity data, before we get into some examples. So in terms of talking about the completeness of the data, if you’re using the MedSAS data—and this trend will be very similar in the CDW. When we look prior to fiscal year 2003, which is the older data collection methods, less than 60% of patients had usable race and ethnicity data. So that means they had values that weren’t missing, unknown, or declined. Beginning in fiscal year 2003, with the change in the data collection method, because we had brand new variables, they didn’t just bring over the old values. The completeness of the data was actually a little bit lower, around 50%. And then from fiscal year 2015 onward, the completeness of the data has been improved steadily, and so now it’s greater than 90%. There is variability in the completeness between the inpatient and outpatient files, so if you do use MedSAS, you always want to use both the inpatient and outpatient data to capture race and ethnicity, even if you are only focusing on utilization in one of those settings.

Similarly for the CDW, if we look over time, based on utilization, this is from the CDW Race Data and Multiple Races Data Quality Report, I believe from 2012. We found that those who only had utilization during the older time period—so for example, 1999 was the furthest back they went—only 39% had usable race available. Whereas, by the time you got to 2012, almost 85% had a useable standard race among those who were utilizing services in 2012.

When we look at the data more recently, I have a different definition of utilization. I was not able to recreate what they used in that handbook, but using a simple definition of utilization based on outpatient visits. For those who utilized VA within fiscal year 2018, 92% had a standard usable race data available using the newer race data collection methods. Almost 1% with the new data were coded as multi-racial, and only 0.3% had conflicting values. That means they had different races coded at different sites for the same individual. When we attempt to use the older data collection methods to fill in the data, less than 1% of Veterans only had race data from the older race data methods. So they didn’t have the new data, but they had the old data. And among those, 1.3% had conflicting values.

So there’s also a CDW ethnicity data handbook that also looked at the completeness of ethnicity data at the time of fiscal year 2012. Among all patients who had ethnicity recorded, about 61% had a valid value or usable value. Amongst those who had a healthcare activity in fiscal year 2012, that improved up to 88%. Seventy-eight percent with one standard category identified were self-identified, so most of the data as of that time of fiscal year 2012 was coming from the newer data collection methods that was self-identified, and then 1% had conflicting ethnicity categories.

In terms of what the data looks like from the CDW more recently, it’s going to look more similar to the race data, so we’re going to see more than 90% who currently have data available for those with more recent utilization.

Their recommendations for using the CDW ethnicity data were essentially the recommendations that were underlined in what we saw for the OMOP business rules. So they are going to give emphasis, they are going to use that self-reported ethnicity data if it’s available. Otherwise, they would use new data from the new data collection method that wasn’t self-report, and only use data from the older data collection methods when no other data were available.

And then finally I’m going to present results from a study from Kevin Stroupe and colleagues that really compared our VA data to non-VA data sources, with two main focuses. The first aim was to estimate the extent to which missing usable race data in the VA MedSAS, which was the data source at that point in time, could be reduced by using non-VA data sources, which were Medicare and DoD data. And then also to evaluate the agreement between the VA self-reported race and the Medicare and DoD data. So these really get at two crucial issues—is one, if I don’t have the data within VA, can I get it from one of these other sources, and two, if I’m obtaining data from the other sources, am I getting similar types of data, or do the data look differently? Their particular cohort was a 10% representative sample of VA patients who obtained services during fiscal years 2004 to 2005. This is nearly half a million patients. And this was shortly after our transition to the newer data collection methods.

And based on data that we saw during that timeframe, they had about 52% who were missing usable race from VA data sources. And they stratified their results by age. And the reason for this was two-fold. So one, we know from the Medicare data that those under the age of 65 are not going to use Medicare to the same extent as those over the age of 65. But in addition, with the DoD data, because of the timeframe over which those data were available, they were really only available for this younger age group. So within the older age group, greater than or equal to 65 years of age, again 53% were missing VA usable race data, but for those who are missing the VA data, 95% had usable Medicare data. So we have almost complete coverage using Medicare. However, the story is different among those who are under the age of 65. Again, about half are missing VA data, but of those only 18% had usable Medicare data, 30% had usable DoD data, and because there’s a little bit of overlap, some people had both sources. Fifty-two percent had usable data from Medicare or DoD. So an important take home message here is although we don’t expect such a large amount of missing data if we’re looking at a recent cohort within VA, we may expect similar percentages among those who are missing. About 95% of the elderly will probably have Medicare data available, whereas maybe for those under the age of 65, maybe only half of them might have data available from these two sources.

An important second aim was to look at the concordance between the non-VA data sources with the self-reported race from VA. And so what they found was agreement was good for both White and African Americans for both data sources. So agreement was greater than 90%, up to 99%. However, the agreement was poor for non-African American minorities, 27 to 55% across categories and data sources. And for Hispanics, when we look specifically at the Medicare data, most of them were categorized by race, not ethnicity. So about 64% were classified as White and only 25% were coded as Hispanic. So this really speaks to the issue of identifying Hispanics using that data. And then for the Asian/Pacific Islanders, and other minorities, they had to be classed into a single category for comparison, because we didn’t have that level of detail either available in the other sources, or because agreement was so poor that even in pooling, when data were available in finer categories, in order to get agreement that was a little bit better, you would have to pool.

All right. I’m going to show some brief examples, simple examples that will show us a little bit about some basic frequency of what the data look like and how you might access them in SQL. I am going to refer you to some other sources for using SQL in CDW, and with the race data in particular. So this first one is Getting Started with Using CDW, which includes examples for SQL. The Best Practices Guide Race Data, this provides several examples for multiple tasks that use race and ethnicity data. Again, the way the data look now are not going to match what’s in that guide. But they do go and do more complex things with the data than what I’m able to show in this presentation. And then also another good use is this Using SQL to Sort Out Race in CDW. And again, the structure of the data is going to be a little bit different, but there’s also going to be examples that are fleshed out more. And then just using the server that I used, because other servers may have different archives of the data with different structure or different content.

So the first simple example is just looking at the PatientRace table, is just to summarize what are the values that we see for race within this table. So one very important thing to note, so when we look at the race table, is that we have very few values that are actually denoted as missing. So we have a total of six. So for most people, if they’re missing the PatientRace table, instead of having a value in this table that says missing, they’re just simply not going to have a record, and that’s how you’re going to identify those missing values, by the fact they’re not going to link to the table. And then here we do see we have this category White, not of Hispanic origin, which is a non-standard way of phrasing this group.

And so if you want to map the standard race values, there are a couple of different ways that you can do so. You can create a lookup table that maps between the non-standard and the standard races, and that is the example that I’m going to show you here, in this guide. You can also just directly program that. And program—you know, if you find the certain pieces of text that you’re interested in, here’s how you’re going to re-code that value. And that’s the example that’s provided in Using SQL to Sort Out Race. And so that example is provided there fully. If you do use a lookup table, this particular example is based on page 10 of the Race Data Best Practices Guide. There were some additional entries that I found in later data that weren’t available at the time of that guide, and these would be coded under their scheme as unable to map. Which would be this unknown/missing, and the Asian/Pacific Islander category. However, if you’re creating a lookup table, you may want to use different categories than what they have there. That’s another nice feature of the lookup table. So for example, you may want to code Asian/Pacific Islander into a useful category rather than unable to map.

And so using the code that they had, I just want to point out this lookup table is a temporary table. You provide information on the non-standard race value that’s coming in, the standard race value that’s coming out, and then you actually—you would type in or copy and paste each of these values. I just want to point out that they did have a value here of null, that’s a text value, and some of the—because you cannot link on an actual underlying null value. In some of the examples of the older data structure, it was useful to have a text value of null, which you could then link upon.

But once we create this table, we can simply use this to convert our data to standard values. We join our PatientRace table to the temporary lookup table based on the race from the PatientRace table and the inbound race from the lookup table, and then we can report on the standard race, and now the six values of missing are coded as unable to map, and we no longer have that group of White/non-Hispanic Whites. They are now included in the group of Whites.

For PatientEthnicity, again, we can look and see what our underlying data look like. Again, we see very few missing or unknown values. Most individuals are coded as Not Hispanic or Latino. And we do have some Declined to Answer or Unknown responses. But these are different from just simply missing, because that value was actually entered into the system that it was Unknown or Declined.

And then we can also look at the data collection method, so it’s going to be similar between race and ethnicity. What I’ve presented here is the collection method from the PatientRace table, and the key thing to note is the vast majority of these are coded as the default value, which is self-identification. We have observed clerks entering data in the past. This was not something that either they changed or they necessarily even knew how to change. So just because it’s coded as self-identification does not necessarily mean it was truly self-identified. So although a lot of our older handbooks really emphasize using this self-identified race over other race values, this is if you’re using the newer race data collection methods, you may choose to ignore the collection method, understanding that it’s really the default value in there and it is not necessarily as meaningful as we would like.

When we use the LegacyRace variable, I just want to remind you, it’s in the PatientRace table, but because the patient race, they could have selected more than one race from the newer data collection methods, we can end up having duplicate values from the LegacyRace, so when I’m using this data, I am going to select the distinct combinations of patients in LegacyRace, the PatientSID. And I just have the most frequently occurring results here, but we can see these are non-standard values, the vast majority of which are missing. And if we want to see what they look like once we standardize, we then link to our lookup table, the RaceTranslationTable. Again, most values are missing. We have White, Black, African American. And very few—you know, Asian or Native Hawaiian or Pacific Islander, just because these data are really missing for most individuals.

And then just a simple example, if you wanted to combine the data from the newer data collection methods and the older data collection methods, I’ve provided an example where I just create a dummy sample cohort using the top 10 PatientSIDs from the patient table. When you join them using a union null, you just need to make sure that you have the same data setup. They don’t have to be the same names, but they have to be the same data types for each of these elements that we’re collecting. So for example, PatientRace and LegacyRace have different names, but they contain the same data type and the same information. And then I have a collection method for the newer data, but not for the older data. So I still have to read in the variable here, and so in this case, I just put in the null value. You could also put in the meaningful value to let you know that’s from the older data race collection methods.

So finally, recommendations to address data quality. Just as a quick summary, when we have multiple sources of race and ethnicity data, we really want to emphasize the data from the newer collection data collection methods, if they’re available. We only use the older data if newer data are not available. And unless you have a cohort say that’s based on utilization from maybe a longer time period ago, there may not be enough information in that LegacyRace variable to make it necessarily worthwhile even trying to capture that information. But if you do use it, remember that you want to use LegacyRace to get race and ethnicity from CDW, or the variable RACE in MedSAS. If conflicting values are still present, you may prioritize values from specific sites, if it’s relevant for your project, depending on how you identify the sites, you know, primary sites utilized by the patients or the preferred site for the patient. But if you don’t have a method to prioritize the sites, you could use all recorded values as suggested in the handbook. And when you’re using MedSAS, you do want to use patient race and ethnicity from both the inpatient and outpatient files.

If you’re using non-VA data, it can help you with filling in missing data. You want to be careful of any potential bias, such as in Medicare, we saw it’s the older patients or those who are disabled that tend to have the outside data sources. And you also need to be careful about classifying data from non-Black minorities, because the agreement can be poor.

Just remember for Medicare, if you're using the Vital Status File, it has multiple records per person, so you want to match on date of birth, gender, and Social Security Number, and that the data is not particularly good for identifying Hispanics. And if this is a concern, you may want to use the RTI\_RACE variable, or if you have a concern in identifying Asians.

Where to go for more help? So, VIReC has nicely put together a race and ethnicity page that contains a variety of links, probably everything I link on my presentation is located here.

We also have general links for VA data sources through VIReC.

There are options for getting questions, the VHA/HSRData Listserv is a really good source to go. Before you post questions there, I would always highly recommend that you first do a search to see if somebody has already addressed the question that you have. Otherwise, you can post a question. You can directly contact the VIReC help desk.

Here is the contact information for both myself and VIReC.

And that is it. Do we have any questions?

Hira: Thank you, Maria. Yes, we do have a few questions for you, so I will just go through them. The first question: Can you briefly explain the Choice data files, any race data that cross-reference this dataset, and your thoughts on how or if this dataset will change vis a vis the VA Mission Act?

Dr. Maria Mor: I am trying to think. I do not particularly recall seeing the race data in the Choice data files, but are you talking about like the pit domain tables? But I would have to double check that again. So I cannot speak to that. I can take another look at those data elements, but when I was looking for them, it was not with a focus on race and ethnicity. But I didn’t recall seeing that in there. And I do not know how the data are going to be structured for the Mission Act.

Hira: Okay. We can refer this person to either email yourself or VIReC’s help desk for more information. We’ll try to find the answer for you.

Dr. Maria Mor: So, I would suggest the VIReC help desk. I know—I mean, HERC has done a lot of work with the Choice data, but if it needs to go somewhere else, I think the VIReC help desk is really going to help get the person to the correct place.

Hira: All right. Thank you. The next question: Did you say that race is interpreted as conflicting when from different stations and interpreted as multiracial when different races are within the same station?

Dr. Maria Mor: Yeah. So, for the newer data collection methods, where you can select more than one race, if they have more than one race in there, that means at a single instance when they came in, they selected more than one race. It doesn’t mean again that they—you know, that they were in there as White, they came in another day and said, oh wait, you know I’m actually Asian. They don’t just add on. It’s supposed to be a new record and it’s supposed to overwrite. So if they have multiple races in there, that should be because the patient reported multiple races at that same point in time.

Hira: Thank you. Last question, here: Is the race variable captured at the CBOC level? If so, where are these data stored?

Dr. Maria Mor: So, the race—so when a patient comes into a visit at a CBOC, the data is captured. But as I said, the data or the demographic record is at that PatientSTA3N level, and so the data are going to be captured in the demographic record for the parent station. So just because the data is in there as a parent station level, doesn’t mean it was collected at the parent station. So a person could only go to the CBOC and maybe data only collected there, but just like everything else that’s collected on that individual at the CBOC, it’s recorded in the CDW at the STA3N level.

Hira: All right. Thank you so much, Maria. Those were all the questions we have from the audience. If you have additional questions, the audience—you can contact the presenter directly and also email the VIReC help desk at [VIReC@VA.gov](mailto:VIReC@VA.gov). You can tune in for the next session in VIReC’s Database and Methods Cyberseminar on Monday, May 6th at 12:00 PM Eastern, instead of our usual 1:00 PM slot. Dr. George Hauser and Ankur Bhargava will be here to present on VetPop: A New Data Source with Socioeconomic Variables. We hope to see you there.

[ END OF AUDIO ]