Session #7: Assessing Race and Ethnicity in VA Data

April 1, 2019

Maria K. Mor, PhD
Center for Health Equity Research and Promotion
VA Pittsburgh Healthcare System
By the end of this session, attendees will be able to:

- Locate race and ethnicity in VA and Medicare data
- Assess the quality of VA race and ethnicity data
- Create SQL code to use race and ethnicity data
Session Outline

• Introduction
• Locating race and ethnicity in VA data
• Locating race and ethnicity in Medicare/Medicaid
• Quality of VA race/ethnicity data
• Examples
• Recommendations to address data quality issues
• Where to go for more help
Poll Question #1

What is your role in research and/or quality improvement?

- Investigator, PI, Co-I
- Data manager, analyst, or programmer
- Project coordinator
- Other – please describe via the Q&A function
Poll Question #2

Have you ever used VA Race/Ethnicity Data?

- Yes
- No
Session Outline

• Introduction
• Locating race and ethnicity in VA data
• Locating race and ethnicity in Medicare/Medicaid
• Quality of VA race/ethnicity data
• Examples
• Recommendations to address data quality issues
• Where to go for more help
Racial/ethnic disparities in health and health care persistent in US and in VHA

In US
• Disparities in measures for access and quality vary by racial/ethnic group
• Improved trends with disparities in some measures improving for Blacks, Asians, and Hispanics, most disparities have not changed or have worsened for other racial/ethnic groups (AHRQ 2018)

In VHA
• Racial/ethnic disparities persist even though financial barriers to receiving care are minimized
• Although quality has improved, significant within-facility disparities observed in clinical outcomes (Trivedi 2011)

More research to detect, understand, and address disparities in health and health care is needed
Problems with Race/Ethnicity Data in VA

Accurate race/ethnicity data are essential to disparities research and research on clinical factors associated with race/ethnicity.

Problems with race/ethnicity data in the VA:

- Incomplete
- Inaccuracies
- Inconsistent over time
Racial/Ethnic Distribution of Veterans

Use of VA health care differs by race and service connection

Asian Veterans **less likely** to use regardless of service connection
Black, and AIAN races **more likely** to use VHA services

National Center for Veterans Analysis and Statistics 2017 Minority Report
(https://www.va.gov/vetdata/docs/SpecialReports/Minority_Veterans_Report.pdf)
## VA Race and Ethnicity Categories

**VHA Handbook 1601A.01 (2009)**

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Spanish</th>
<th>Hispanic</th>
<th>Latino</th>
</tr>
</thead>
<tbody>
<tr>
<td>Race</td>
<td>American Indian or Alaska Native</td>
<td>Asian</td>
<td>Black or African American</td>
</tr>
<tr>
<td>(&gt;1 may be selected)</td>
<td>2 question format: ethnicity, race</td>
<td>Self-reported</td>
<td></td>
</tr>
</tbody>
</table>
# Acquisition of Race/Ethnicity Data in VHA

| How are these data acquired? | Patient (self-report)  
Proxy  
VHA Enrollment Coordinator or Clerk |
|-----------------------------|--------------------------------------------------------------------------|
| When are these data acquired? | VA Form 10-10EZ Application for Health Benefits (on-line, paper, interview)  
Inpatient or outpatient visit to VHA facility |

Data are entered directly into VistA
Session Outline

• Introduction

• **Locating race and ethnicity in VA data**
  • Locating race and ethnicity in Medicare/Medicaid
  • Quality of VA race/ethnicity data

• Examples

• Recommendations to address data quality issues

• Where to go for more help
Poll Question #3

What sources of VA race/ethnicity data have you used? (check all that apply)

- Never used race/ethnicity data
- CDW
- OMOP
- MedSAS files
- DOD (VADIR, DaVINCI)
- Other VA data sources
Race/Ethnicity Variables in MedSAS

Prior to FY2003 (old data collection methods)
• Race and ethnicity captured jointly in the variable RACE
• Single value allowed for race/ethnicity

After FY2003 (new data collection methods)
• Multiple races captured in RACE1-RACE7
• Single value for ethnicity captured in ETHNIC
• RACE1-RACE7 and ETHNIC have a length of 2 characters
  • First character has race or ethnicity
  • Second character has method of data collection

Location
• Inpatient: Main (PM) file, 1976-present
• Outpatient: Visit (SF) and Event (SE) files, 1997/1998- present
Medical SAS Datasets: Race/Ethnicity Values (Pre-2003)

**RACE: Single value for race and ethnicity**

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hispanic, white</td>
</tr>
<tr>
<td>2</td>
<td>Hispanic, black</td>
</tr>
<tr>
<td>3</td>
<td>American Indian</td>
</tr>
<tr>
<td>4</td>
<td>Black</td>
</tr>
<tr>
<td>5</td>
<td>Asian</td>
</tr>
<tr>
<td>6</td>
<td>White</td>
</tr>
<tr>
<td>7 or missing</td>
<td>Unknown</td>
</tr>
</tbody>
</table>
# Medical SAS Datasets: Race Values (Post-2003)

*RACE1-RACE7: Race and method of data collection
First character specifies race*

<table>
<thead>
<tr>
<th>1st Character</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>American Indian Or Alaska Native</td>
</tr>
<tr>
<td>8</td>
<td>Asian</td>
</tr>
<tr>
<td>9</td>
<td>Black or African American</td>
</tr>
<tr>
<td>A</td>
<td>Native Hawaiian or Other Pacific Islander</td>
</tr>
<tr>
<td>B</td>
<td>White</td>
</tr>
<tr>
<td>C</td>
<td>Declined to Answer</td>
</tr>
<tr>
<td>D</td>
<td>Unknown</td>
</tr>
<tr>
<td>(blank)</td>
<td>Missing</td>
</tr>
</tbody>
</table>
## Medical SAS Datasets: Ethnicity Values (Post-2003)

*ETHNIC: Ethnicity and method of data collection*
*The first character captures ethnicity*

<table>
<thead>
<tr>
<th>1st Character</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Declined To Answer</td>
</tr>
<tr>
<td>H</td>
<td>Hispanic or Latino</td>
</tr>
<tr>
<td>N</td>
<td>Not Hispanic or Latino</td>
</tr>
<tr>
<td>U</td>
<td>Unknown</td>
</tr>
<tr>
<td>(blank)</td>
<td>Missing</td>
</tr>
</tbody>
</table>
# Medical SAS Datasets: Race and Ethnicity Source (Post-2003)

**RACE1-RACE7, ETHNIC**  
*The second character specifies method of data collection*

<table>
<thead>
<tr>
<th>2nd Character</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(blank)</td>
<td>Missing</td>
</tr>
<tr>
<td>O</td>
<td>Observer</td>
</tr>
<tr>
<td>P</td>
<td>Proxy</td>
</tr>
<tr>
<td>S</td>
<td>Self-identification</td>
</tr>
<tr>
<td>U</td>
<td>Unknown By Patient</td>
</tr>
</tbody>
</table>
Corporate Data Warehouse (CDW)

- National repository of data from VistA Patient File with race and ethnicity data from October 1999 to present
- Contains 1 demographic record for each VA station a Veteran has visited
- Contains standard and nonstandard race values
- Racial data available PatSub.PatientRace
  - Race (newer collection standards)
  - LegacyRace (older collection standards)
- **Use both variables to obtain all available race data**

CDW Race Table Changes

The structure of the CDW data is subject to periodic changes.

As of January 2018, none of the available CDW documentation for race and ethnicity match the current data structure.

Patient 3.0 Domain VIReC Factbook documents the current data structure

Changes in the business rules for extraction have also led to some differences in the underlying race data stored in CDW.

### CDW documentation may refer to race from older collection methods as being located in other CDW tables.

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
</tr>
</thead>
</table>
| Patient.Patient or SPatient.SPatient tables | RaceSID contained the SID for the patient race  
Would link to CDWWork.Dim.Race to map to race |
| Patsub.PatientRace     | Currently contains the fields LegacyRace and LegacyRaceSID  
Previously, all race values were stored in the variable Race but those from older collection methods had a value of Null for CollectionMethod |

## Race Tables in CDW

All race data are contained in PatSub.PatientRace

Data are at the Patient/STA3N level with the most recent data available for the patient

<table>
<thead>
<tr>
<th>Race</th>
<th>Contains patient race from newer collection methods. Multiple records if more than one race identified.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CollectionMethod</strong></td>
<td>Contains method of data collection for Race</td>
</tr>
<tr>
<td><strong>LegacyRace</strong></td>
<td>Contains patient race from the older collection methods</td>
</tr>
<tr>
<td></td>
<td>- Does not allow for multiple races</td>
</tr>
<tr>
<td></td>
<td>- The same value of LegacyRace will be contained on all records for a single PatientSID if that patient has multiple values of Race recorded.</td>
</tr>
<tr>
<td></td>
<td>- Most patients have values of “<em>Missing</em>”, indicating the presence of no data on LegacyRace.</td>
</tr>
</tbody>
</table>
Non-standard Race Values in CDW

26 of 31 non-standard races can be mapped to 4 standard races

**Examples**

<table>
<thead>
<tr>
<th>Non-standard Race</th>
<th>Standard Race</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amer Indian or Alaskan Native, American Indian, American Indian/ Alaskan Native</td>
<td>American Indian or Alaska Native</td>
</tr>
<tr>
<td>Black; Black Not of Hisp orig; Black, Non Hispanic; Hispanic Black</td>
<td>Black or African American</td>
</tr>
<tr>
<td>White Not of Hisp orig; White, Not Hispanic; Hispanic White; Caucasian;</td>
<td>White</td>
</tr>
<tr>
<td>Pacific Islander</td>
<td>Native Hawaiian or Other Pacific Islander</td>
</tr>
</tbody>
</table>

Non-standard values rarely used in Race (<1%)
Current standard values rarely used in LegacyRace (<1%)
Non-mapped Values in CDW

5 values are not mapped to standard values

4.6% of data fall into 1 of these 5 categories (2012)

<table>
<thead>
<tr>
<th>Non-mapped values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian or Pacific Islander</td>
</tr>
<tr>
<td>Asian Pacific Islander</td>
</tr>
<tr>
<td>Asian/Pacific Islander</td>
</tr>
<tr>
<td>Mexican American</td>
</tr>
<tr>
<td>Unknown</td>
</tr>
</tbody>
</table>

As of March 2019

- 17.6% of non-missing LegacyRace fall into 1 of these categories
- 96.7% of these non-mapped values are Unknown
- 3.0% of non-mapped values indicate Asian or Pacific Islander
Multiple Race Values in CDW

- Approximately 1.7% of patients linked to a standard race have more than 1 standard race (2013)
- Not possible to identify most recent record for a patient
- Recommendation for multiple values
  - Use only self-identified races (if recorded)
  - Use all recorded races for patients without self-identified race

*CDW Race Data and Multiple Races* (Data Quality Report):
http://vaww.vhadataportal.med.va.gov/Portals/0/DataQualityProgram/Reports/CDW_Race_Data_and_Multiple_Races.pdf (VA Intranet only)
Ethnicity in CDW

Ethnicity data found in 2 CDW tables

**PatSub.PatientEthnicity** - *new method*

‘HISPANIC OR LATINO’ / ‘NOT HISPANIC OR LATINO’

**PatSub.PatientRace** (LegacyRace or rarely Race) - *old method*

Hispanic race/ethnicity (e.g., HISPANIC, WHITE; HISPANIC, BLACK)

Non Hispanic race/ethnicity (e.g., WHITE NOT OF HISP ORIG; BLACK NOT OF HISP ORIG)

Not all race/ethnicity values indicate ethnicity (e.g., ASIAN, BLACK)

*CDW Ethnicity Data* (Data Quality Report)

http://vaww.vhadataportal.med.va.gov/Portals/0/DataQualityProgram/Reports/CDW_Ethnicity_D ata.pdf (VA Intranet only)
VINCI OMOP Version 5

- VINCI Observational Medical Outcomes Partnership (OMOP) seeks to use a Common Data Model (CDM) to map and standardize data.
- Data on Race and Ethnicity are contained in the OMOPV5.Person table.
- Contains one standard value for Race and Ethnicity for each PERSON_ID.
  - OMOPV5MAP.PERSON_SPatient_Spatient will link PERSON_ID to other CDW identifiers.
  - See documentation regarding those without PatientICN or other potential linkage issues with patient identifiers.
  - Excludes non-veterans, test patients, and possible test patients.

VINCI_V5_OMOP_DATABASE_DATA_SPECIFICATIONS_01152018: https://www.vapulse.net/docs/DOC-60310

04/2019
# Race in OMOP

OMOP CDM follows VA Data Quality Program’s “Race Data and Multiple Races Report” and VIReC’s Researcher’s Notebook “Using SQL to "Sort Out" Race in CDW”

| Source data                  | Source.SPatient_SPatient (now LegacyRace in Patsub.PatientRace)  
|------------------------------|------------------------------------------------------------------
|                              | Source.Patsub_PatientRace                                       
| **Six categories for race**  | White                                                             
|                              | Black or African American                                        
|                              | Asian                                                             
|                              | American Indian or Alaska Native                                 
|                              | Native Hawaiian or other Pacific Islander                        
|                              | Unknown                                                           

“CDW Race Data and Multiple Races:”  

“VIReC Researcher’s Notebook: Using SQL to “Sort Out” Race in CDW”:  

04/2019
Race Logic in OMOP

1. Identify records as self-report or non-self-report and count distinct values.
2. Select the most frequently occurring self-reported race value.
3. If no self-reported race or counts of self-reported race (not including unknown or null) are equal, then select the most frequent non-self-reported race.
4. If there isn’t a most frequent value, then select the race value found on record at the patient’s preferred institution.
5. If that is null, then select the value edited most recently as determined by ETLBatchID in the SPatient file.
6. If no most frequent or recent non-null value is available, then the value is “UNKNOWN”
Ethnicity in OMOP

OMOP CDM follows the “OMB Standards for Data on Race and Ethnicity” and the VA Data Quality Program’s “CDW Ethnicity Data Report.”

<table>
<thead>
<tr>
<th>3 categories for ethnicity</th>
<th>Hispanic or Latino</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not Hispanic or Latino</td>
</tr>
<tr>
<td></td>
<td>Unknown</td>
</tr>
</tbody>
</table>

OMOP CDM Logic for Ethnicity

OMOP uses only the self-reported information provided under the new collection method, when available

Otherwise Ethnicity is captured from non-self-reported data provided by the new collection method

Ethnicity captured under the old collection methods is used when no data are available from the new recording method

## Race in DaVINCI (Joint DoD and VA data)

<table>
<thead>
<tr>
<th>race_cd</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>White</td>
</tr>
<tr>
<td>M</td>
<td>Asian or Pacific Islander</td>
</tr>
<tr>
<td>N</td>
<td>Black</td>
</tr>
<tr>
<td>R</td>
<td>American Indian or Alaskan Native</td>
</tr>
<tr>
<td>X</td>
<td>Other</td>
</tr>
<tr>
<td>Z</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>race_ethnic_cd</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>American Indian/Alaskan Native</td>
</tr>
<tr>
<td>B</td>
<td>Asian or Pacific Islander</td>
</tr>
<tr>
<td>C</td>
<td>Black, not Hispanic</td>
</tr>
<tr>
<td>D</td>
<td>White, not Hispanic</td>
</tr>
<tr>
<td>E</td>
<td>Hispanic</td>
</tr>
<tr>
<td>X</td>
<td>Other</td>
</tr>
<tr>
<td>Z</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

[https://vaww.vinci.med.va.gov/VinciCentral/DataSources/Index](https://vaww.vinci.med.va.gov/VinciCentral/DataSources/Index)

Session Outline

• Introduction
• Locating race and ethnicity in VA data
• **Locating race and ethnicity in Medicare/Medicaid**
• Quality of VA race/ethnicity data
• Examples
• Recommendations to address data quality issues
• Where to go for more help
Sources of Medicare/Medicaid Race in VA

VA Vital Status File
- **CMS_RACE** (Master File only)
- Master File contains one record for each SSN-date of birth (DOB)-gender combination found in VA data
- Some SSNs have more than one record

VA Medicare Data
- Denominator file from Medicare
- **RACE** (same as **CMS_RACE**)
- **RTI_RACE**

VA Medicaid Data
- Medicaid Personal Summary (Enrollment)
- **EL_RACE_ETHNCY_CD**
Medicare Race/Ethnicity Data

Potentially useful source of data for Veterans enrolled in Medicare, which generally means they are:

- Age 65 and older (>95% of VA elderly)
- Disabled (~20% of VA patients <65 years)
- Diagnosed with end stage renal disease

Derived primarily from Social Security Administration (SSA)

- Obtained at the time of application for SSN and/or replacement card
- Reporting sources: Usually self or family

Distinctions from current VA race/ethnicity data

- ‘Hispanic’ is a race category
- No multiple race reporting
Medicare Race Data from SSA

Until 1980, only 4 categories collected:

- White
- Black
- Other
- Unknown

In 1980, ‘Other’ replaced by:

- Asian, Asian American or Pacific Islander
- Hispanic
- American Indian or Alaskan Native
RTI Race in Medicare

Research Triangle Institute (RTI) created and implemented an algorithm to increase accuracy of race variable, especially for Hispanic and Asian individuals.

- **RTI_RACE** available in Medicare Denominator File
- Algorithm uses first name, last name, preferred language, place of residence
- Improvement in sensitivity of racial codes
  - Increased from 30% to 77% for Hispanic
  - Increased from 55% to 80% for Asian/Pacific Islander
Medicare Race Data Summary

Data quality issues

- Information on most enrollees (those who obtained SSN prior to 1980) limited to original 4 categories
- SSN application form – single question format and no multiple race reporting

Initiatives to improve data quality

- Periodic updates on American Indians and Alaskan Natives from Indian Health Service
- 1997 survey of enrollees classified as ‘Other’, ‘Unknown’, or with Spanish surname, requesting race/ethnicity self-report
- RTI Race Algorithm
# Medicaid Race/Ethnicity

**EL_RACE_ETHNCY_CD**

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>White</td>
</tr>
<tr>
<td>2</td>
<td>Black or African American</td>
</tr>
<tr>
<td>3</td>
<td>American Indian or Alaskan Native</td>
</tr>
<tr>
<td>4</td>
<td>Asian</td>
</tr>
<tr>
<td>5</td>
<td>Hispanic or Latino – No race information available</td>
</tr>
<tr>
<td>6</td>
<td>Native Hawaiian or Other Pacific Islander</td>
</tr>
<tr>
<td>7</td>
<td>Hispanic or Latino and one or more races</td>
</tr>
<tr>
<td>8</td>
<td>More than one race</td>
</tr>
<tr>
<td>9</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

04/2019
Medicaid Race/Ethnicity Variables Summary

Summary variable
EL_RACE_ETHNCY_CD

Individual variables
ETHNICITY_CODE
RACE_CODE_1 – RACE_CODE_5

Can identify multiple races and/or race and ethnicity
Medicaid Race/Ethnicity Data Issues

- Availability lags behind both VA and Medicare
- Fewer enrollees than Medicare (~10%)
- Data collection changes over time
  - October 1998 many changes/additions
Session Outline

• Introduction
• Locating race and ethnicity in VA data
• Locating race and ethnicity in Medicare/Medicaid
• **Quality of VA race/ethnicity data**
• Examples
• Recommendations to address data quality issues
• Where to go for more help
Medical SAS Datasets:
Completeness of Race and Ethnicity Data

Prior to FY2003
<60% of patients had usable* race/ethnicity

FY2003
Completeness of data was about 50%

FY2015
Completeness of data was >90%

Completeness varies between inpatient and outpatient files.

Always use both the inpatient and outpatient data to capture race/ethnicity in the MedSAS files.

* A usable race value is any value that is not ‘missing’ or ‘unknown’ or ‘declined’
## CDW Completeness of Race Data

Percent of patients with a standard race in the CDW varies by year of most recent healthcare activity

<table>
<thead>
<tr>
<th>FY</th>
<th>Standard Race, %</th>
<th>FY</th>
<th>Standard Race, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999*</td>
<td>39.0</td>
<td>2006</td>
<td>63.0</td>
</tr>
<tr>
<td>2000</td>
<td>42.6</td>
<td>2007</td>
<td>65.9</td>
</tr>
<tr>
<td>2001</td>
<td>43.5</td>
<td>2008</td>
<td>66.6</td>
</tr>
<tr>
<td>2002</td>
<td>44.1</td>
<td>2009</td>
<td>67.2</td>
</tr>
<tr>
<td>2003</td>
<td>48.2</td>
<td>2010</td>
<td>68.5</td>
</tr>
<tr>
<td>2004</td>
<td>53.8</td>
<td>2011</td>
<td>70.2</td>
</tr>
<tr>
<td>2005</td>
<td>58.7</td>
<td>2012</td>
<td>84.6</td>
</tr>
</tbody>
</table>

*No activity after FY1999

**CDW Race Data and Multiple Races** (Data Quality Report):
CDW Completeness of Race Data FY2018

New collection methods

92% of Veterans have standard usable race data available from these new methods

Almost 1% with new data are coded as multiracial

0.3% have conflicting values

Old collection methods

<1% of Veterans only have older race data

1.3% of those have conflicting values

Unique Veterans with ≥ 1 outpatient visit (NoncountClinicFlag = ‘N’) in FY2018
CDW Completeness of Ethnicity Data

61% of all patients have ethnicity recorded

88% with healthcare activity in FY 2012

78% with one standard category are self-identified

1% have conflicting ethnicity categories

Recommendations for Using CDW Ethnicity Data

1. If available, use ethnicity captured through self-identification

2. Otherwise, use ethnicity captured through new recording method (Patsub.PatientEthnicity)

3. Use older collection methods (Patsub.PatientRace, LegacyRace, or Race) when no other data are available
Comparison to Non-VA Data Sources

Aims

1. To estimate the extent to which missing “usable” race data in VA MedSAS files can be reduced by using non-VA data sources (Medicare and DoD)
2. To evaluate the agreement between VA self-reported race data in MedSAS files and Medicare and DoD race data

Cohort

10% representative sample of VA patients obtaining services during FY2004-2005 (N=570,018)

## Reduction in Missing Data

52% were missing usable race from VA data sources

<table>
<thead>
<tr>
<th>Age ≥ 65</th>
<th>Age &lt; 65</th>
</tr>
</thead>
<tbody>
<tr>
<td>53% missing usable VA race data</td>
<td>51% missing usable VA race data</td>
</tr>
<tr>
<td><em>Of those…</em></td>
<td><em>Of those…</em></td>
</tr>
<tr>
<td>95% had usable Medicare data</td>
<td>18% had usable Medicare data</td>
</tr>
<tr>
<td></td>
<td>37% had usable DoD data</td>
</tr>
<tr>
<td></td>
<td>52% had usable data from Medicare and/or DoD data</td>
</tr>
</tbody>
</table>
## Concordance with Non-VA Data Sources

*Table compares non-VA data sources to self-reported VA race/ethnicity data*

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>--</th>
</tr>
</thead>
<tbody>
<tr>
<td>White and African Americans</td>
<td>Agreement was good (93-99%) for both non-VA data Sources</td>
</tr>
<tr>
<td>Non-African American Minorities</td>
<td>Agreement was poor (27-55%) for both Medicare and DoD</td>
</tr>
<tr>
<td>Hispanics</td>
<td>Classified as White (64%) rather than Hispanic (25%) in the Medicare data</td>
</tr>
<tr>
<td>Asian, Pacific Islanders, and Other Minorities</td>
<td>Had to be collapsed into one category for comparisons</td>
</tr>
</tbody>
</table>
Session Outline

- Introduction
- Locating race and ethnicity in VA data
- Locating race and ethnicity in Medicare/Medicaid
- Quality of VA race/ethnicity data
- **Examples**
  - Recommendations to address data quality issues
  - Where to go for more help
SQL Examples in CDW

Getting Started with Using CDW
Includes several seminars on using SQL to join and manipulate CDW data

http://vaww.virec.research.va.gov/CDW/Documentation.htm (VA Intranet only)

Race Data Best Practices Guide
Several SQL examples for multiple tasks utilizing race and ethnicity data


Researcher’s Notebook: Using SQL to “Sort Out” Race in CDW


Connected to server vhacdwa01.vha.med.va.gov
Please note that the location of race data is now different from what is in these guides. 04/2019
Example: Patsub.PatientRace

```
SELECT Race, FORMAT(COUNT(Race), 'N0') as Freq
FROM CDWWork.Patsub.PatientRace
GROUP BY Race
ORDER BY COUNT(Race)
```

Format to show commas

<table>
<thead>
<tr>
<th>Race</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Unknown at this time</em></td>
<td>2</td>
</tr>
<tr>
<td><em>Missing</em></td>
<td>4</td>
</tr>
<tr>
<td>WHITE NOT OF HISP ORIG</td>
<td>74,231</td>
</tr>
<tr>
<td>ASIAN</td>
<td>269,125</td>
</tr>
<tr>
<td>NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER</td>
<td>273,076</td>
</tr>
<tr>
<td>AMERICAN INDIAN OR ALASKA NATIVE</td>
<td>281,759</td>
</tr>
<tr>
<td>UNKNOWN BY PATIENT</td>
<td>695,252</td>
</tr>
<tr>
<td>DECLINED TO ANSWER</td>
<td>871,746</td>
</tr>
<tr>
<td>BLACK OR AFRICAN AMERICAN</td>
<td>3,938,842</td>
</tr>
<tr>
<td>WHITE</td>
<td>18,054,251</td>
</tr>
</tbody>
</table>
Example: Mapping to Standard Race Values

• Create a table that maps between non-standard and standard values
  → Code is on p.10 of “Race Data Best Practices Guide”

• Map these additional entries to “Unable to Map:”
  “*Unknown at this time*”  “*Missing*”  “Asian/Pacific Islander”

• Change mapped categories to match project needs

See Researcher’s Notebook: Using SQL to “Sort Out” Race in CDW for alternate method for programming standard race values
Example: Race Translation Table

```sql
drop table if exists #RaceTranslationTable

create table #RaceTranslationTable
    (InboundRace varchar(50),
    StandardRace varchar(50));

insert into #RaceTranslationTable
    values('NULL','Unable to Map')

insert into #RaceTranslationTable
    values('AMER INDIAN OR ALASKAN NATIVE','AMERICAN INDIAN OR ALASKA NATIVE')

insert into #RaceTranslationTable
    values('AMERICAN INDIAN','AMERICAN INDIAN OR ALASKA NATIVE')

insert into #RaceTranslationTable
    values('AMERICAN INDIAN / ALASKAN NATIVE','AMERICAN INDIAN OR ALASKA NATIVE')

insert into #RaceTranslationTable
    values('AMERICAN INDIAN OR ALASKA NATIVE','AMERICAN INDIAN OR ALASKA NATIVE')

insert into #RaceTranslationTable
    values('AMERICAN INDIAN OR ALASKAN NATIVE','AMERICAN INDIAN OR ALASKA NATIVE')

insert into #RaceTranslationTable
    values('AMERICAN INDIAN/ALASKAN NATIVE','AMERICAN INDIAN OR ALASKA NATIVE')

insert into #RaceTranslationTable
    values('AMERICAN INDIAN/ALASKA NATIVE','AMERICAN INDIAN OR ALASKA NATIVE')
```

Text ‘NULL’ ≠ null value

Delete table if it already exists

Use # to create temporary tables

*See page 10 of Race Data Best Practices Guide for the remaining code*
Example: Convert to Standard Values

```sql
SELECT b.StandardRace, FORMAT(COUNT(b.StandardRace), 'N0') as Freq
FROM CDWWork.PatSub.PatientRace as a
LEFT JOIN #RaceTranslationTable as b
ON a.Race = b.InboundRace
GROUP BY b.StandardRace
ORDER BY COUNT(b.StandardRace)
```

<table>
<thead>
<tr>
<th>StandardRace</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unable to Map</td>
<td>6</td>
</tr>
<tr>
<td>ASIAN</td>
<td>269,125</td>
</tr>
<tr>
<td>NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER</td>
<td>273,076</td>
</tr>
<tr>
<td>AMERICAN INDIAN OR ALASKA NATIVE</td>
<td>281,759</td>
</tr>
<tr>
<td>UNKNOWN BY PATIENT</td>
<td>695,252</td>
</tr>
<tr>
<td>DECLINED TO ANSWER</td>
<td>871,746</td>
</tr>
<tr>
<td>BLACK OR AFRICAN AMERICAN</td>
<td>3,938,842</td>
</tr>
<tr>
<td>WHITE</td>
<td>18,128,482</td>
</tr>
</tbody>
</table>
Example: Patsub.PatientEthnicity

```sql
SELECT Ethnicity, FORMAT(COUNT(Ethnicity), 'N0') as Freq
FROM CDWWork.PatSub.PatientEthnicity as a
GROUP BY Ethnicity
ORDER BY COUNT(Ethnicity)
```

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Missing</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Unknown at this time</em></td>
<td>1</td>
</tr>
<tr>
<td>DECLINED TO ANSWER</td>
<td>473,961</td>
</tr>
<tr>
<td>UNKNOWN BY PATIENT</td>
<td>989,292</td>
</tr>
<tr>
<td>HISPANIC OR LATINO</td>
<td>1,408,195</td>
</tr>
<tr>
<td>NOT HISPANIC OR LATINO</td>
<td>21,794,870</td>
</tr>
</tbody>
</table>
Example: Collection Method

```sql
SELECT CollectionMethod, FORMAT(COUNT(CollectionMethod), 'N0') as Freq
FROM CDWWork.PatSub.PatientRace
GROUP BY CollectionMethod
ORDER BY COUNT(CollectionMethod)
```

<table>
<thead>
<tr>
<th>CollectionMethod</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Unknown at this time</em></td>
<td>1</td>
</tr>
<tr>
<td><em>Missing</em></td>
<td>3</td>
</tr>
<tr>
<td>PROXY</td>
<td>444</td>
</tr>
<tr>
<td>OBSERVER</td>
<td>1,832</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>218,142</td>
</tr>
<tr>
<td>SELF IDENTIFICATION</td>
<td>24,237,866</td>
</tr>
</tbody>
</table>

Default Value, rarely changed
Example: LegacyRace

```
SELECT a.LegacyRace, FORMAT(COUNT(1), 'N0') as Freq
FROM (SELECT DISTINCT PatientSID, LegacyRace from PatSub.PatientRace) as a
GROUP BY a.LegacyRace
ORDER BY COUNT(1) DESC;
```

Need to remove duplicates
Example: LegacyRace (Standard Values)

```
SELECT b.StandardRace, FORMAT(COUNT(b.StandardRace), 'N0') as Freq
FROM (SELECT DISTINCT PatientSID, LegacyRace from PatSub.PatientRace) as a
LEFT JOIN #RaceTranslationTable as b
    ON a.LegacyRace=b.InboundRace
GROUP BY b.StandardRace
ORDER BY COUNT(b.StandardRace) DESC;
```

<table>
<thead>
<tr>
<th>StandardRace</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unable to Map</td>
<td>20,837,492</td>
</tr>
<tr>
<td>WHITE</td>
<td>2,481,836</td>
</tr>
<tr>
<td>BLACK OR AFRICAN AMERICAN</td>
<td>622,239</td>
</tr>
<tr>
<td>AMERICAN INDIAN OR ALASKA NATIVE</td>
<td>15,106</td>
</tr>
<tr>
<td>UNKNOWN BY PATIENT</td>
<td>329</td>
</tr>
<tr>
<td>DECLINED TO ANSWER</td>
<td>112</td>
</tr>
<tr>
<td>ASIAN</td>
<td>10</td>
</tr>
<tr>
<td>NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER</td>
<td>6</td>
</tr>
</tbody>
</table>
Example: Multiple Sources (Long Format)

```
DROP TABLE IF EXISTS #SampleCohort
SELECT TOP 100 PatientSID, Sta3n
    INTO #SampleCohort
    FROM CDWWork.Patient.Patient;

SELECT a.PatientSID, a.Sta3N, b.Race, b.CollectionMethod
FROM #SampleCohort as a LEFT JOIN cdwork.PatSub.PatientRace AS b
ON a.PatientSID = b.PatientSID

UNION ALL

SELECT c.PatientSID, c.Sta3N, d.LegacyRace, NULL as CollectionMethod
FROM #SampleCohort AS c
LEFT JOIN (SELECT DISTINCT PatientSID, LegacyRace from PatSub.PatientRace) as d
ON c.PatientSID = d.PatientSID

ORDER BY 1;
```

- Names don’t need to match as long as data type and column order are the same
- Can select different value for CollectionMethod but must have the same # of columns for each table
- Sorts by the 1st column
Session Outline

• Introduction
• Locating race and ethnicity in VA data
• Locating race and ethnicity in Medicare/Medicaid
• Quality of VA race/ethnicity data
• Examples
• **Recommendations to address data quality issues**
• Where to go for more help
Recommendations: VA Data

When multiple sources of race and ethnicity exist...
- Use data from newer collection methods, if available
- Only consider data from old collection methods (< FY 2003), if newer data are not available*
  - LegacyRace contains race and ethnicity in CDW
  - RACE contains race and ethnicity in MedSAS

If conflicting values are still present...
- Prioritize values from specific sites, if relevant for the project
- Consider using all recorded values

When using MedSAS...
- Obtain race and ethnicity from both inpatient and outpatient files

* Unless the cohort has less recent utilization, LegacyRace may not be helpful
Recommendations: Non-VA Data

- Use of non-VA race data can reduce missing data.
- Carefully consider any potential bias (e.g., age or disability) in the outside data source.
- Classifying non-Black minorities as “Other” results in better agreement with other data sources.
- Potential supplementary data sources:
  - Medicare
  - Department of Defense
  - Medicaid
  - Special Surveys
Recommendations: Medicare

When using VA VSF…

→ Match on date of birth and gender, in addition to (scrambled) SSN

→ Researchers most likely to identify the right individuals if they use all 3 elements when conducting their VSF-study cohort record match

Note that…

→ Medicare data cannot be used to identify Hispanics with any degree of accuracy or completeness, but

→ RTI_RACE in the Medicare Denominator file can increase the identification of Hispanics and Asians
Session Outline

• Introduction
• Locating race and ethnicity in VA data
• Locating race and ethnicity in Medicare/Medicaid
• Quality of VA race/ethnicity data
• Examples
• Recommendations to address data quality issues
• Where to go for more help
VIReC resources on Race and Ethnicity

Race and Ethnicity overview:
http://vaww.virec.research.va.gov/RaceAndEthnicity/Overview.htm
(Intranet only)
Quick links for VA data resources

**Quick Guide: Resources for Using VA Data**

**VIReC:** http://vaww.virec.research.va.gov/Index.htm (VA Intranet)

**VIReC Cyberseminars:** http://www.virec.research.va.gov/Resources/Cyberseminars.asp

**VHA Data Portal:** http://vaww.vhadataportal.med.va.gov/Home.aspx (VA Intranet)

**VINCI:** http://vaww.vinci.med.va.gov/vincicentral/ (VA Intranet)

**CDW:** https://vaww.cdw.va.gov/Pages/CDWHome.aspx (VA Intranet)
## VIReC Options for Specific Questions

<table>
<thead>
<tr>
<th><strong>HSRData Listserv</strong></th>
<th><strong>HelpDesk</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Community knowledge sharing</td>
<td>Individualized support</td>
</tr>
<tr>
<td>• ~1,400 VA data users</td>
<td><a href="mailto:virec@va.gov">virec@va.gov</a></td>
</tr>
<tr>
<td>• Researchers, operations, data stewards, managers</td>
<td>(708) 202-2413</td>
</tr>
</tbody>
</table>
Contact information

VA Information Resource Center
Hines VA Hospital
virec@va.gov
708-202-2413

Maria Mor
Maria.Mor@va.gov
Next session:
May 6, 2019 at 1 pm Eastern

Database & Methods Cyberseminar Series

Session #8: VetPop: A New Data Source with Socioeconomic
Variables

R. George Hauser, MD & Ankur Bhargava
Veterans Affairs Connecticut Healthcare System
Selected Recent References on Race/Ethnicity Data


Selected Recent References on Race/Ethnicity Data


Selected Recent References on Race/Ethnicity Data


Selected Recent References on Race/Ethnicity Data


Selected Recent References on Race/Ethnicity Data


