Name:	Aize Cao
Affiliation:	Tennessee Valley Healthcare System
Email:	Aize.Cao@vanderbilt.edu
Presentation type (s):	Poster

Quality Assurance of Demographics Consistency between Veterans Affairs and Medicare Data

Aize Cao, PhD[1,2]; Margaret Gonsoulin, PhD[5] ; Kristin de Groot, MPH[5]; Elizabeth Hanchrow, RN, MSN[1,2;] Daniel Parker, BS[1,2]; Kristine Lynch, PhD[3,4]; Scott L. DuVall, PhD[3,4]; Michael E. Matheny, MD, MPH[1,2]; Stephen A. Deppen, PhD[1,2]

[1]Tennessee Valley Healthcare System, Veterans Affairs Medical Center, Nashville, TN; [2]Vanderbilt University, Nashville, TN; [3]VA Salt Lake City Health Care System, Salt Lake City, UT; [4]University of Utah, Salt Lake City, UT; [5]VA Information Resource Center, Edward Hines Jr. VA Hospital, Hines, IL

Abstract: *We developed two OMOP CDMs: Veteran Health Administration (VHA) Corporate Data Warehouse (OMOP VA) and Medicare enrollment and claims data (OMOP CMS). We examined the consistency between the two datasets on date of birth (DOB), gender and race. As one quality assurance approach, this cross validation provided data consistency and an estimate of possible reduction of missing data in each dataset. Using scrambled social security number, we linked the two datasets. Out of 23 million of unique patients in OMOP-VA and 12 million in OMOP-CMS, about 7 million patients were common between the two OMOP CDMs. The consistency between these two CDMs was 93.82% for DOB, 98.87% for gender, and 48.1% for race. Low race matching rates occurred due to extensive missing race value in the VA dataset. Missing DOB and gender in OMOP VA could be reduced by 0.54% and 0.07% respectively. Missing race could be reduced by 48.8% and 2.4% for the OMOP VA and OMOP CMS datasets, respectively. By comparing demographic information from two longitudinal datasets we were able to significantly reduce missing data in both.*

Introduction: VINCI (VA Informatics and Computing Infrastructure) began transforming VA CDW into OMOP in 2015. Starting in 2017, VINCI began collaborating with VIReC (VA Information Resource Center) to transform CMS data into the OMOP common data model. Demographic information in electronic medical records and claims data were collected from multiple sources and multiple times. Longitudinal data commonly have conflicting states of what should be relatively stable demographic variables, namely DOB, gender and race. Thus, demographic data often require a great deal of cleaning, costing researchers' significant time and effort.

Between these two large datasets we had an opportunity to compare each demographic variable and estimate the potential to reduce the level of missing or conflicting data. To validate the demographic information in both OMOP CDM person table, gender, race and DOB were compared.

Method and Results: The VA cohort included over 23 million veterans who accessed the VA healthcare system between January 2000 and December 2016. Before converting the VHA CDW demographic information into the OMOP person table, a multi-faceted effort was taken to evaluate conflicting data and determine the unique demographic values for each person based on the best practice logic developed by content experts (VIReC) and source data managers (VINCI).[1,2,3,4]

The Medicare data comprised 12 million unique veterans who have ever been enrolled in Medicare. Medicare date of birth and gender were sourced from the Medicare Vital Status file. Medicare race was sourced from the one of two sources. The Medicare Vital Status file contains a race value; however, these data underreport minority groups, especially Hispanic or Asian or Pacific Islanders[5]. Therefore, an imputed race value (RTI Race) which was based on an algorithm utilizing surname, state of residence, and language preference was used when available. Otherwise, the race value from the Medicare Vital Status file was the race value of record.

The two OMOP CDMs were linked through scramble SSN resulting in about 7 million common veterans. We found 6,523,073 same DOB, a 93.82% rate of concordance. No missing DOB was observed in the OMOP-CMS cohort, but 37,537

missing DOB occurred in the OMOP-VA cohort (0.5%).

Table 1. Consistency of gender in OMOP-VA and OMOP-CMS person table

		Gender in OMOP-CMS person table		
		Male	Female	Unknown
Gender in OMOP-VA person table	Male	6246548	33366	2
	Female	40132	627399	1
	Unknown	3906	1273	0

Gender matched 98.87% across the two datasets (**Table 1**). Three patients in OMOP-CMS had missing gender and a non-missing gender value in OMOP-VA. Over 5,179 patients with missing gender in OMOP-VA had non-missing value in OMOP-CMS.

Unknown race was the most common result in the VA dataset (51.5%). Unknown race may arise if the individual refuses to identify their race, if race is missing or if conflicting race values could not be resolved. This high rate of unknown race caused the lack of consistency (48.14%) observed for race between the two datasets (**Table 2**). About 3.4 million patients with missing or unknown race in OMOP-VA were not missing race in OMOP-CMS. Specifically, almost 3 million patients race were identified as white in OMOP-CMS data, doubling those identified as white in the VA dataset. On the other side, 163,640 patients with unknown race in OMOP-CMS could be found in OMOP-VA. These two sources put together could be used to assign a race to all but 184,608 members of the cohort, assuming resolution of conflicting race values between the two datasets.

Table 2. Consistency of race in OMOP-VA and OMOP-CMS person table

		Race in OMOP-CMS person table				
		White	Black or African American	Asian	American Indian or Alaska Native	Unknown
Race in OMOP-VA person table	White	2,788,832	12,249	4,649	6,150	146,095
	Black or African American	9,116	350,637	314	318	7832
	Asian	1,367	142	13,680	29	6,964
	American Indian or Alaska Native	13,197	1,368	106	9,250	2,749
	Unknown	2,958,531	378,929	34,416	21,099	184,608

Conclusion: This study found high consistency for gender and DOB between two OMOP CDMs. Low consistency for race occurred due to the high portion of unknown race in VA data. Race data from CMS could reduce the missing VA race data by over 50%. We showed a potential improvement in demographic data accuracy above existing cleaning algorithms by comparing and combining the OMOP-VA and OMOP-CMS datasets.

Financial Support: This study's work supported with resources and the use of facilities at the TVHS and Salt Lake VA, and is funded by VA HSR&D VINCI. Support for VA/CMS Data provided by VA HSR&D, VA Information Resource Center (Project Numbers SDR 02-237 and 98-004).

References

1. Kevin T. Stroupe, Elizabeth Tarlov, Qiuying Zhang, Thomas Haywood, Arika Owens, Denise M. Hynes, Use of Medicare and DOD data for improving VA race data quality, Journal of Rehabilitation Research & Development, Vol 47 (8), Page 781-796, 2010
2. Data quality analysis team, CDW race data and multiple races, http://vaww.vhadataportal.med.va.gov/Portals/0/DataQualityProgram/Reports/CDW_Race_Data_and_Multiple_Races.pdf
3. Gonsoulin, Margaret. Using SQL to "Sort Out" Race in CDW: A method for cleaning multiple values of race. The Researcher's Notebook; no. 6. Hines, IL: VA Information Resource Center; 2016. http://vaww.virec.research.va.gov/Notebook/RNB/RNB6-CDW-SQL-to-Sort-Out-Race-CY16.pdf
4. Best practices guide race data, http://vaww.vhadataportal.med.va.gov/Portals/0/DataQualityProgram/Reports/Best_Practices_Guide_Race_Data.pdf
5. Eicheldinger, C. Bonito, A. More Accurate Racial and Ethnic Codes for Medicare Administrative Data. Health Care Financing Review, Spring 2008, Volume 29, Number 3. https://www.cms.gov/Research-Statistics-Data-and-Systems/Research/HealthCareFinancingReview/downloads/08Springpg27.pdf