

Demographics

Contains basic person-descriptives for the people found in VDW tables. Also serves as a lookup dataset for MRNs.

1. Overview

The demographics table contains basic person-descriptives for the people found in VDW tables. It should also serve as a lookup dataset for MRNs—**every MRN appearing in any other VDW file should appear in Demographics (even if nothing substantive is known about the person).**

2. Data Dictionary

Variable name	Type (Length)	Preferred Format	Description	Valid Values	Comments
MRN	char (varies)	n/a	An arbitrary identifier unique to an individual within a site.	Any, so long as they uniquely identify individual people.	Used to link people across files within a site. May or may not contain the official local indigenous person identifier (e.g., "medical record number"). Regardless of whether it contains the official local identifier, this variable should never leave the site. Projects needing to move individual-level data should create a study-specific person identifier and substitute it for MRN on any data that is to move (See, e.g., the %DeIDDset() standard macro.)
birth_date	numeric (4)	mmddyy10	Date of Birth	SAS Date	
gender	char(1)	n/a	Gender or sex of the person.	M Male F Female U Unknown O Other: Transsexual, Transgendered, or anything else that does not fit into one of the prior categories.	This variable is whatever the health plan collects as GENDER (role) or SEX (biology), in that order of preference. Information on transgendered and transsexual individuals should be coded as "Other". Sites able to provide reliable data on transgendered and transsexual persons should document the availability of this information on their site dataset implementation page on the portal.
Race1 Race2 Race3	char(2)	n/a	Race(s) of the person.	HP Native Hawaiian or Other Pacific Islander	See Note 1 below.

Race4 Race5				IN American Indian/Alaska Native AS Asian BA Black or African American WH White MU More than one race, particular races unknown or not reported UN Unknown or Not Reported	
hispanic	char(1)	n/a	Hispanic origin (ethnicity)	Y Yes N No U Unknown	Nulls are no longer allowed.
primary_language	char(3)	n/a	Primary language spoken at last contact	As defined in the ISO -639-2 specification, or 'unk', for unknown. Partial list: eng English spa Spanish	<ul style="list-style-type: none"> ▪ This is capable of changing over time—we are only concerned with the most recent information. ▪ People whose languages are not known should have the value 'und', (undetermined) or 'unk' (unknown). (The former is part of the official ISO-639-2 spec, and the latter is our own devising. The two codes should be interpreted as equivalent.) ▪ People whose languages are listed as 'other' in source data should be mapped to 'und'. ▪ Clarity sites may find this information in patient.language_c
needs_interpreter	char(1)	n/a	Needs interpreter to communicate w/an English -only speaker?	Y Yes N No U Unknown	<ul style="list-style-type: none"> ▪ This is capable of changing over time—we are only concerned with the most recent information. ▪ Clarity sites may find this information in pat_enc.interpreter_need_yn

3. Notes

Note 1: Race

Our goal is to have the most complete, reliable, and detailed race and ethnicity information in the demographics file. SDMs should gather race information from all sources permitted at your site—for example, tumor registry data; state birth & death data; and regular large-scale social surveys. If there is more than one source of race information for a set of individuals, SDMs should give preference to self-reported sources.

Where multiple sources (or multiple measures from a single source) conflict as to the race of a given person, and the SDM does not have a reason to prefer one source to another (e.g., no one source is known to be most trustworthy, etc.) code all races indicated by any source.

For data sources that treat Hispanic ethnicity as a value of race (so e.g., you know the person is Hispanic, but you don't know their race) code the person's race as "Unknown".

Format

Race is now coded in accordance with the NIH Policy And Guidelines On The Inclusion Of Women And Minorities As Subjects In Clinical Research.

- Code up to 5 races in race variables 1 through 5.
- Where more than one category applies to a given person, assign values to the race<x> variables in the order in which they are listed above. So, the highest-listed value should go in Race1, the second-highest-listed in Race2, and so forth (see example table below).
- Fill in any unused variables with the value for "Unknown or not reported". None of these variables should be null.

Example Race Coding	Race1	Race2	Race3	Race4	Race5
White only	WH	UN	UN	UN	UN
White and Pacific-Islander	HP	WH	UN	UN	UN
No Race Known	UN	UN	UN	UN	UN
African-American and Native-American	IN	BA	UN	UN	UN
Multi-racial, particular races not reported.	MU	UN	UN	UN	UN

Recoding From The Prior Scheme

The below format can be used to translate the values of the prior race coding scheme into those of the new scheme.

```
value $Race
'01' = 'WH'
'02' = 'BA'
'03' = 'IN'
'04', '05', '06', '08', '09',
'10', '11', '12', '13', '14',
'96' = 'AS'
'07', '20', '21', '22', '25',
'26', '27', '28', '30', '31',
'32', '97' = 'HP'
Other = 'UN'
;
```

4. Future Plans

We plan to revisit coding of sex (biology), gender (role), race, and ethnicity as our health plans increase their sophistication of collecting this information from their members.

5. Contact Information

Please contact the VDW Enrollment and Demographics working group for questions and suggestions.

Related content

Enrollment/Demographics QA Workplan

Change Proposal: Move language information out of Demographics & into a separate table

Comments (3)

Liu, Julie Feb 15, 2011 03:57 PM

we did not implement

"For data sources that treat Hispanic ethnicity as a value of race (so e.g., you know the person is Hispanic, but you don't know their race)

code the person's race as "White.""

We have several data sources that don't use Hispanic as race value except HealthConnect. If all the data sources have shown unknown in the race but Hispanic in the data source of HealthConnect, we set the race unknown and hispanic flag to 'Y'.

Hart, Gene Nov 14, 2012 06:15 PM

There is a possibility of a conflict in Note 1.

"If there is more than one source of race information for a set of individuals, SDMs should give preference to self-reported sources."

"Where more than one category applies to a given person, assign values to the race<x> variables in the order in which they are listed above".

Is it the expectation that race reported in Epic should always be in race1 ("give preference to self report") or would a state death report of black go into race1 if Clarity said white ("assign values in the order listed above")?

This is not of major importance, but I just got asked this question on a project.

Pardee, Roy Nov 15, 2012 01:06 PM

Aye, that's a legit complaint. The way I interpret these sentences is that SDMs should use self-report-or-not *in deciding whether to use data from a given local source for a particular person*. So for instance, if you have race data for Patty Patient both from EPIC and from your tumor registry data, you might want to leave out the tumor race info because (at least at Group Health) tumor race is not self-reported, and EPIC is.

Once you've dropped any less-preferred race info, then you should code all races indicated. So if I have 2 self-reported data sources (say, EPIC and data from an annual Health Risk Assessment survey) & I include data from both for Patty Patient (because as far as I know they're both equally valid) then I would code all races indicated for Patty in either source, and place them in the Race1-5 vars strictly according to the hierarchy & irrespective of the source of the data. So if EPIC says Patty is AS and HRA says Patty is WH, then the EPIC data goes in Race1, but if it's reversed, then the HRA data goes in Race1.