Health Services Research and Development Service
Data Quality Alerts
Information for Applicants and Reviewers

VA Information Resource Center
June 2007

This informational memorandum alerts researchers and reviewers to some data quality issues that should be considered in the design of HSR&D research applications. This alert addresses VA race/ethnicity and mortality data.

Background

Experienced investigators know that selection and use of secondary data for research must take into account the quality and validity of the data sources. For example, most datasets used in research are susceptible to out-of-range or missing values. Careful examination of the data to determine the degree and nature of the problem should be routine practice when creating analytic datasets for research.

Race/Ethnicity

Summary: Issues regarding VHA race and ethnicity data quality include missing values and inconsistency across time due to changes in data collection method and in allowable response categories and format.

Inconsistency Across Time

- Patient race in the Department of Veterans Affairs (VA) information system was previously recorded based on an administrative or clinical employee's observation. Since 2003, the VA has collected self-reported race in compliance with a new federal guideline from the Office of Management and Budget (OMB).
  - In the VHA Medical SAS Inpatient and Outpatient Datasets there was previously a single variable (RACE) containing race and ethnicity information.
  - In FY 2003 the variables RACE1 - RACE6 were added to the SAS Inpatient Datasets and in FY 2004 RACE1 - RACE7 were added to the SAS Outpatient Datasets to allow for multiple race reporting. These new variables are coded to incorporate both race and method of collection (e.g., 9O = Black/African American observer reported, 9S = Black/African American self-reported).
  - Hispanic ethnicity is recorded as a separate variable under the new coding. The variable ETHNIC was added to the SAS Inpatient Datasets in FY 2003 and to the Outpatient Datasets in FY 2004.
  - The RACE variable is only partially populated in the FY 2003 Medical SAS Inpatient Main Dataset and is entirely empty in the FY 2004 Inpatient Main Dataset.

- VIReC conducted a study examining issues related to the transition to the new race/ethnicity data collection standards in the VA (Sohn et al., 2006a).
  - Using Medical SAS Inpatient and Outpatient Datasets for FY 1997 through FY 2002 and FY 2004, the researchers found that the agreement between observer-recorded race before the transition and self-reported race/ethnicity in FY 2004 was high overall and among White and African American (AA) VHA users. This

result suggests that observer-recorded and self-reported data for these groups can be used across years without creating serious bias.

  - Observer-recorded race was less reliable for non-AA minorities. The investigators demonstrated that combining those groups to create a higher-level more inclusive group improved accuracy of observer-reported race/ethnicity and that the AA and non-AA distinction provided the greatest agreement between observer- and self-reported race/ethnicity.

Missing Values

- Missing values on race are not uncommon. For the period from FY 1997 to FY 2005 the overall proportion of VHA users with missing or unknown race could be as high as 45%. Following the implementation of the new standard, the frequency of missing values has increased (Sohn et al., 2006a).
- To address incompleteness issues, researchers can attempt to fill in missing race/ethnicity using
  - race/ethnicity data from previous years (as mentioned above, agreement between self-reported values and previous observer values has been shown to be good) or
  - alternative data sources (e.g., for a substantial proportion of elderly veterans whose race is missing or unknown in VA data, this information can be found in Medicare data).

Mortality

Summary: The VA Vital Status Master File provides the most complete information on veterans' vital status and date of death. Investigators whose study outcome measures include mortality are advised to use this file in their research.

- Historically, four sources of information on veterans' vital status and date of death have been available to VA researchers: the Beneficiary Identification Records Locator System (BIRLS), the Medical SAS Inpatient Datasets (for veterans who died during an inpatient stay), the Medicare Vital Status File, and the Social Security Administration's Death Master File.
- The VA-NDI Mortality Data Merge Project[1] compared the dates of death available in the VA with those obtained from the National Death Index (NDI) which uses data obtained from death certificates and is considered the gold standard for mortality data. Results of this project can be found in VIReC's Technical Report #2 (available on the VIReC website) and have been published (Sohn, et al, 2006b).
  - The study found that agreement between any one of the four sources and the NDI was suboptimal or worse but that combining information from these sources greatly improved completeness and accuracy of vital status ascertainment. The Project recommended that the VA develop a registry containing death dates from all available sources
  - The project also developed and tested an algorithm to determine the "best" date of death when the sources contain conflicting data.

[1] The VA-NDI Mortality Data Merge Project was funded by VA HSR&D Service (SDR 03-157) and the VA Information Resource Center (SDR98-004).

- The VA Vital Status Master File, as recommended by the VA-NDI Mortality Data Merge Project, is a registry containing death dates from all available sources.
- The VA Vital Status Mini File, constructed by VA National Data Systems, uses an algorithm to select a "best" date of death for each social security number. The existence of the two datasets allows researchers to determine which dataset best fits the needs of a particular study.
- More information about these datasets and how to access them is available on the VIReC website.
- The mortality ascertainment algorithm used to produce the VA Vital Status Mini File has not yet undergone rigorous evaluation. Early feedback from researchers using this dataset raises concern regarding the sensitivity and specificity of the methodology.
    - A small proportion of veterans whose vital status, as determined using the methodology, was "presumed dead" when the Mini File was first constructed were found to have utilization records in databases from later fiscal years.
    - Conversely, some very elderly veterans (over 90 years old) whose vital status, as determined using the methodology, is "presumed living" have had no utilization for a period of 5 years or more.
    - Researchers should be aware of these issues and use appropriate caution. VIReC will continue to report data quality information relating to this new dataset.

More information on these and other data quality topics can be found on the VA Information Resource Center (VIReC) website.

References

Arnold N, Sohn M, Maynard C, & Hynes DM. *VIReC Technical Report 2: VA-NDI Mortality Data Merge Project*. Edward Hines, Jr. VA Hospital, Hines, IL: VA Information Resource Center, April 9, 2006. Available at: http://vaww.virec.research.va.gov/References/TechnicalReports/VIReCTechnicalReports.htm.

Sohn, M.W., Zhang, H., Arnold, N., Stroupe, K., Taylor, B.C., Wilt, T.J., & Hynes, D.M. (2006). Transition to the new race/ethnicity data collection standards in the Department of Veterans Affairs. *Population Health Metrics, 4*(7). Available at: http://www.pophealthmetrics.com/content/4/1/7.

Sohn, M.W., Arnold, N., Maynard, C., & Hynes, D.M. (2006). Accuracy and completeness of mortality data in the Department of Veterans Affairs. *Population Health Metrics, 4*(2). Available at: http://www.pophealthmetrics.com/content/4/1/2.