Variance Estimation for trend of nurse diversity using ACS PUMS data 2000-2015

Tiandong Li

Health Resources and Services Administration (HRSA)

5600 Fishers Lane, Rockville, MD 20857

Introduction

Recent research indicates that racial and ethnic diversity in the nursing workforce is linked to improved health outcomes, such as access to care for racial and ethnic minority patients, greater patient choice and satisfaction, and better patient–provider communication (Grumbach & Mendoza, 2008; Smedley, Butler & Bristow, 2004; IOM, 2004; U.S. Dept of Health and Human Services, 2006).

To provide a comprehensive review of the trends in racial and ethnic demographics for the nursing workforce, we used single-year American Community Survey (ACS) Public Use Microdata Sample (PUMS) data from years 2000 to 2015 to examine the trend of the race/ethnicity diversity of Registered Nurses (RNs). Specifically, we conducted a logistic regression model at the person level to evaluate the impact of race/ethnicity, time in years, and their interaction on the status of working as an RN among the general U.S. population.

For estimating the standard errors of estimates based on PUMS data, the U.S. Census Bureau provides two methods: 1) the replication method and (2) the design factor method. Generally, replication method will produce a more accurate estimate of a standard error. (United States Census Bureau, 2015).

However, replicate weights are only available for year 2005 and after. The current study analyzed the data from a longer period of time and can't use the replication method across all the years.

Design factor method (a type of generalized variance function) has been used to analyze ACS PUMS data since 2000. Although the design factors created for ACS PUMS data provided a way to estimate simple statistics without using replicate weights, they can't be applied to our study for a few reasons: 1) design factors were developed for simple descriptive statistics, but not for complex statistics, such as parameters in logistic regressions; 2) no design factors were developed for statistics based on combined data across 16 years; and 3) design factors were developed for the general US population and may not apply to subpopulations, such as Registered Nurses (RN).

As a practical solution, we developed a two-step method to estimate variance for the data across 16 years. First, we estimated adjustment factors for parameters in logistic regressions, using the years of data with replicate weights available (2005-2015). The adjustment factor is defined as the ratio of the variance estimate based on replicate weights and that based on the full sample weight only. To compute the variance estimates based on the data across all years (2000-2015), the adjustment factor is then applied to the variance estimated based the full sample weight from 2000-2015 data.

Data

Single-year ACS PUMS data for years 2000 to 2015 were concatenated for the trend analysis. ACS across these years included two stages. The demonstration stage of the ACS in 2000-2004 selected a sample of counties (1,240 counties) and 0.8 million addresses. During the full implementation stage, addresses were sampled from all 3,141 US counties and the number of addresses increased from 2.9 million in 2005-2011 to 3.54 million in 2011-2015.

The RNs in this study included Advanced Practice Registered Nurse (APRN), which consists of nurse practitioners, clinical nurse specialists, certified nurse midwives, and certified registered nurse anesthetists. These APRN categories were not separated from RNs in the PUMS file until year 2010, so they are included into this study as a whole.

For the logistic regression model, the outcome variable was the dummy variable for the status of working as an RN among the general U.S. population. The predictor variables included the time in years with values 0-15 and race/ethnicity. The race/ethnicity was categorized as Hispanic, NonHispanic White only, NonHispanic Black only, NonHispanic Asian only, Multiple races and NonHispanic Others, where NonHispanic White only is the reference group.

Method

As shown below, data analysis used a logistic regression to examine the trend of RN diversity by race/ethnicity

$$logit(\pi) = \alpha + \beta_1 race + \beta_2 time + \beta_3 (time * race)$$

where π is the probability of being an RN in the U.S. population; $exp(\beta_1)$ is the odds ratio of π comparing the racial/ethic groups to the reference group, averaged across time; $exp(\beta_2)$ shows the yearly change in the odds ratio of π for the reference group; and $exp(\beta_3)$ shows the differential yearly change in the odds ratio comparing the race/ethnicity groups of interest to the reference group.

To estimate the sampling variance of odds ratios, this study takes a two-step approach, as shown below.

Step 1: Estimate an Adjustment Factor (AF) for each parameter in the logistic regression model using 2005-2015 data

$$AF_{0515} = SE_{rep,\,0515} / SE_{norep,\,0515}$$

where $SE_{rep,\,0515}$ is the standard error from Proc Surveylogistic using replicate weights; $SE_{norep,\,0515}$ is the standard error from Proc Surveylogistic using full sample weight only.

Step 2: Apply Adjustment Factors to estimates (with full sample weight) based on 2000-2015 data

$$SE_{rep,\,0015} = AF_{0515} \times SE_{norep,\,0015}$$

Result

Table 1 demonstrates standard errors and t-values for parameter estimates in the logistic regression. The original estimates were based on the full sample weight using ACS PUMS data from year 2000 to 2015 and the adjustment factor was computed from the 2005-2015 data. The adjustment factors show a narrow range around 1, specifically between 0.92 and 1.15. The adjustment didn’t change the signification status at 0.05 level of the estimates.

Table 1. Parameter Estimates from Logistic Regression using ACS PUMS data in 2000-2015 (variance adjusted for complex sample design)

		Standard Error			t-value	
Parameter	**Category**	**Original**	**Design factor**	**Adjusted**	**Original**	**Adjusted**
Intercept		0.00633	0.97	0.0061	-720.30	-744.63
time		0.00063	0.95	0.0006	27.19	28.77
RACETH5	**1-Hispanic**	0.03350	1.15	0.0385	-46.12	-40.10
	2-NonHispanic White alone	NA	NA	NA	NA	NA
	3-NonHispanic Black alone	0.02350	0.94	0.0221	-20.61	-21.87
	4-NonHispanic Asian alone	0.02350	1.09	0.0255	16.42	15.10
	5-NonHispanic Other or Multiple	0.05180	0.92	0.0478	-13.17	-14.28
time*RACETH5	**1-Hispanic**	0.00313	1.12	0.0035	5.91	5.27
	2-NonHispanic White alone	NA	NA	NA	NA	NA
	3-NonHispanic Black alone	0.00226	0.96	0.0022	4.28	4.46
	4-NonHispanic Asian alone	0.00224	1.09	0.0024	-1.93	-1.77
	5-NonHispanic Other or Multiple	0.00493	0.93	0.0046	0.09	0.09

Discussion

The two-step adjustment factor approach follows a similar idea as the design factor approach. While the design factor reflects all components of complex sample design - unequal sample weights, stratification and clustering, the adjustment factor only reflects stratification and clustering. The adjustment factor approach intends to take advantage of all the information available from the data. First, as the full sample weight is available for all the years of PUMS data, sampling variance related to unequal sample weights was accurately calculated in the weighted analysis using full sample weight. Second, all the data years with replicate weights available were utilized to calculate the adjustment factor. For the trend analysis, most years in the model, 11 out of 16 years, were used to calculate the adjustment factor. Last, the adjustment factor was calculated for the same statistics as the ones in the final model.

However, a caveat exits when applying the adjustment factor calculated using 2005-2015 data to the trend analysis based on 2000-2015 data. The sample design for the demonstration stage of ACS in 2000-2004 is different to the full implementation stage in 2005-2015 – a sample of 1,240 counties were selected in the demonstration stage rather than taking all 3,141 counties in the full implementation stage. The clustering at the county level may cause underestimation of sampling variance using the adjustment factor approach. The scale of the underestimation should be very small, if not ignorable, given the large number of selected counties and the number of years with replicate weight available.

Given that the adjustment factors are in a narrow range around one, variance estimated from analysis using full sample weight may provide a reasonable estimation.

For future study, a simulation study will help demonstrate the scale of the underestimation related to the two-step approach and the accuracy of the variance estimated from the full sample weight. Also, Taylor Series Expansion method could be examined by treating counties as strata or clusters in the corresponding years, while the design information is not provided by Census.

Reference

Grumbach, K. & Mendoza, R. (2008). Disparities in human resources: Addressing the lack of diversity in the health professions. Health Affairs, 27(2): 413-422.

Institute of Medicine. (2004). In the Nation's compelling interest (1st ed.). Washington, DC: National Academics Press.

Smedley, B.D., Butler, A.S., & Bristow, L.R. (2004). In the nation's compelling interest: Ensuring diversity in the health care workforce. Washington, DC: National Academic Press.

U.S. Department of Health and Human Services, Health Resources and Services Administration, Bureau of Health Professions, (2006). The rationale for diversity in the health professions: A review of the evidence. Retrieved from http://bhpr.hrsa.gov/healthworkforce/reports/diversityreviewevidence.pdf.

U.S. Census Bureau, Design and Methodology, American Community Survey, U.S. Government Printing Office, Washington, DC, 2014.