Rate of Missing Socioeconomic Factors in the 4th KNHANES

Brief Communication

Hyun Ah Park*

Department of Family Medicine, Inje University Seoul Paik Hospital, Inje University College of Medicine, Seoul, Korea

This study is to assess how missing values in socioeconomic status (SES) variables were handled in the Korean Journal of Family Medicine (KJFM) article using the Korea National Health and Nutrition Examination Survey (KNHANES) data and to estimate the rate of missing SES variables from the 4th KNHANES. We searched all original articles published in the KJFM from 2007 to 2011 and identified those that used KNHANES as their primary source of data. None of the 11 articles which presented KNHANES SES variables took into account of omitions in the analysis. The estimated rate of missing data on education, household income, marital status, and occupation data of the 4th KNHANES was 0.3 (0.05)%, 2.7 (0.2)%, 0.5 (0.1)%, and 9.4 (0.9)%, respectively. When all four variables were used simultaneously, the rates increased to 11.8 (0.9)%. Respondents with missing household income tended to be older ($P < 0.001$), less educated ($P < 0.001$), and more likely to be unemployed ($P < 0.001$), and widowed ($P < 0.001$). A similar relationship was shown for missing occupation data. Omissions in SES variables in KNHANES were related to certain characteristics of study participants. Researchers using KNHANES data should keep in mind the possible bias which can be introduced by missing SES values.

Keywords: Missing Values; Socioeconomic State; Korea National Health and Nutrition Examination Survey

INTRODUCTION

Missing values are endemic across health and social studies.[1] Missing data reduce statistical power and representativeness of the sample and might cause misinterpretation of the results by introducing bias.[2] Socioeconomic status (SES) variables such as education, income, marital status, and occupation are often unanswered in the large public data sets. In most cases, the missingness in SES variables is related to certain characteristics of the individuals surveyed.[3] However, this issue rarely receives specific focus as a shortcoming of studies, and has rarely been the focus of specific discussion within academic print.

The Korea National Health and Nutrition Examination Survey (KNHANES) provides a rich source of data for studying the relationships between health and SES for primary care physicians. In this report, SES variables (i.e., education, household income, marital status, and occupation) within the Korean Journal of Family Medicine (KJFM) original articles using the KNHANES data were reviewed. Rates of missing SES variables from the 4th KNHANES were estimated, when used independently or used in combination with other SES variables. Finally, other SES characteristics related to the omissions of household income and occupation were assessed.

METHODS

This study was composed of two main parts. The first part

Received: July 12, 2012, Accepted: October 19, 2012

*Corresponding Author: Hyun Ah Park
Tel: 82-2-2270-0952, Fax: 82-2-2267-2030
E-mail: drparkhyunah@gmail.com

included a detailed hand search to select KJFM articles from 2007 to 2011 which used KNHANES as their primary source of data. The methods and results sections of each relevant article were carefully reviewed, and the SES variables used in the univariate and multivariate analysis were checked.

In the second part of the study, rates of missing SES variables (i.e., education, household income, marital status, and occupation) from the Health Interview Survey (HIS) of the 4th KNHANES were estimated including that for all men and women aged more than 19 years old. The HIS consisted of four components: the household core, the sample adult core, the sample adolescent core, and the sample child core. The household core component included the household income and marital status of individuals, and the sample adult core component included their educational and occupational classification. Detailed descriptions of the plan and operation of the survey have been described on the KNHANES website (http://knhanes.cdc.go.kr/).

Educational levels were categorized according to less than elementary school graduate, middle school graduate, high school graduate, and college graduate. In order to calculate the household income level, the mean monthly household income was divided by the root of the number of household members, and was classified into quartiles. Marital status of the individuals (married and living with a partner, divorced or separated, widowed, or unmarried) was also included. Occupational classification used the KNHANES system of classification, which is a modified version of the Korean Standard Classification of Occupation, 6th revision (2007) supplemented by an indicator reflecting unemployment status.

Analyses were performed with Stata ver. 10 (Stata Co., College Station, TX, USA) to incorporate sampling weight. Indicator variables for missing values of SES variables were created. Chi-square tests were used to assess the relationship between the missingness of household income data and occupation classification and other SES variables.

RESULTS

Of the reviewed literature, one article in 2007, three articles in 2008, four articles in 2009, three articles in 2010, and five articles in 2011 used the KNHANES data as their primary data source, totaling 16 articles (5.4%) among 296 original articles during the same period. Eleven articles presented SES data to describe the participants' characteristics with univariate analysis, and 9 articles used them as covariates in multivariate analysis. The most frequently used SES variables in the multivariate analysis were education (9 articles), household income (8 articles), marital status (3 articles), and occupation (4 articles). None of these 11 articles took into account the omissions within the analysis (data not shown).

The estimated rates of missing data on education, household income, marital status, and occupation were 0.3 (standard error, SE, 0.05), 2.7 (0.2), 0.5 (0.1), and 9.4 (0.9), respectively. The variable of occupation had the highest rate of omission. When all four variables were used simultaneously, the rate increased to 11.8 (0.9) (Table 1).

Table 1. Estimated missing rates when each variable is used simultaneously in the analysis of the 4th Korea Health and Nutrition Examination Survey.

	Unweighted n (%)	Weighted % (SE)	Estimated n
Education	64 (0.4)	0.3 (0.05)	115,577
Household income	438 (2.5)	2.7 (0.2)	1,013,253
Marital status	67 (0.4)	0.5 (0.1)	178,261
Occupation	1,577 (9.1)	9.4 (0.9)	3,530,556
Education, household income	494 (2.9)	3.0 (0.2)	1,117,988
Education, household income, occupation	1,916 (11.1)	11.6 (0.9)	4,341,659
Education, household income, marital status	534 (3.1)	3.3 (0.3)	1,234,059
Education, household income, marital status, occupation	1,947 (11.2)	11.8 (0.9)	4,424,218

Table 2. Socioeconomic characteristics comparing respondents and non respondents to the categories of household income and occupation.

	Household income		P-value*	Occupation		P-value*	Total
	Reported	Missing		Reported	Missing		
Unweighted n	16,873	438		15,734	1,577		17,311
% (SE)	97.3 (0.2)	2.7 (0.2)		90.6 (0.9)	9.4 (0.9)		
Age	44.5 (0.2)	48.9 (1.2)	<0.001	44.3 (0.2)	47.2 (0.8)	0.001	44.6 (0.2)
Sex							
Male	49.6 (0.4)	46.2 (2.3)	0.158	51.1 (0.4)	33.8 (1.9)	<0.001	49.5 (0.4)
Female	50.4 (0.4)	53.8 (2.3)		48.9 (0.4)	66.2 (1.9)		50.5 (0.4)
Highest education achieved							
Less than elementary school	19.9 (0.5)	31.8 (2.7)	<0.001	19.5 (0.6)	27.8 (2.0)	<0.001	20.3 (0.5)
Middle school	10.3 (0.3)	11.7 (1.8)		10.4 (0.3)	9.6 (1.1)		10.3 (0.3)
High school	40.1 (0.6)	36.4 (3.2)		40.2 (0.7)	37.8 (1.9)		40.0 (0.6)
College	29.4 (0.7)	19.1 (2.2)		29.9 (0.7)	21.9 (2.3)		29.2 (0.7)
Missing	0.3 (0.05)	1.1 (0.4)		ND	2.9 (0.5)		0.3 (0.05)
Income							
Low	ND	ND		15.4 (0.6)	20.8 (1.9)	<0.001	15.9 (0.5)
Mid-low	ND	ND		24.4 (0.6)	24.5 (1.8)		24.4 (0.6)
Mid-high	ND	ND		28.4 (0.7)	4.9 (1.7)		28.0 (0.6)
High	ND	ND		29.5 (0.9)	23.8 (2.6)		29.0 (0.9)
Missing	ND	ND		2.5 (0.2)	6.0 (0.9)		2.7 (0.2)
Occupational classification†							
Managers, professional	13.2 (0.4)	9.2 (1.9)	<0.001	ND	ND		13.1 (0.4)
Office worker, clerical support workers	8.5 (0.3)	5.2 (1.6)		ND	ND		8.4 (0.3)
Service workers, sales workers	13.9 (0.4)	7.7 (1.5)		ND	ND		13.7 (0.4)
Skilled agricultural, forestry, and fishery workers	5.2 (0.5)	2.9 (0.8)		ND	ND		5.1 (0.5)
Craft, plant and machine operators, and assemblers	11.5 (0.4)	8.3 (1.7)		ND	ND		11.4 (0.4)
Elementary occupations	8.2 (0.3)	7.2 (1.9)		ND	ND		8.2 (0.3)
Unemployed	30.4 (0.9)	38.5 (3.3)		ND	ND		30.6 (0.9)
Missing	9.1 (0.9)	20.9 (3.1)		ND	ND		9.4 (0.9)
Marital status							
Married & living with a partner	68.7 (0.6)	48.1 (3.1)	<0.001	68.4 (0.7)	64.8 (2.0)	<0.001	68.0 (0.6)
Divorced or separated	4.4 (0.2)	5.0 (1.2)		4.4 (0.2)	3.7 (0.5)		4.4 (0.2)
Widowed	6.8 (0.2)	15.0 (2.0)		6.5 (0.3)	12.0 (1.0)		7.0 (0.2)
Unmarried	20.0 (0.6)	28.4 (3.0)		20.4 (0.6)	18.1 (1.8)		20.2 (0.6)
Missing	0.3 (0.1)	3.5 (1.1)		0.3 (0.1)	1.4 (0.4)		0.4 (0.1)

*By chi-square test. †By the classification of the Korea Health and Nutrition Examination Survey, which is a modified version of the Korean Standard Classification of Occupation, 6th revision (2007) supplemented by an indicator reflecting unemployment status.

Table 2 presented SES characteristics according to the missingness of household income and occupation data. Respondents with missing household income tended to be older ($P < 0.001$), less educated ($P < 0.001$), and more likely to be unemployed ($P < 0.001$), and widowed ($P < 0.001$). A similar relationship was shown by the missingness of occupation classification.

DISCUSSION

The rates of missing data for the categories of household income and occupation within KNHANES were not low, i.e., 2.7% and 9.4%, respectively, and the missingness was not randomly dispersed throughout the data. However, no articles in KJFM clarified the process by which missing values or attrition reduce the sample size, nor did they explicate how these problems introduce potential bias to their findings.

Traditional approaches to working with missing values are case deletion, pair wise deletion, mean substitution, or the inclusion of indicator variables. KJFM Articles using SES variable from KNHANES used the case deletion method, presumably as it is the default in standard statistical packages. Use of the case deletion method using KNHANES SES data could result in the loss of up to 12% of the data, and this figure will increase when researchers use multiple components of KNHANES together. Therefore, these approaches can result in serious biases in a positive or a negative direction, increasing type II errors.[4,5]

This study showed that low educational achievement, unemployment state, and lower household income were all associated with omission in SES data. Similar results have been reported in oversea studies. A postpartum survey in California[3] and in the National Health Interview Survey[6] showed that respondents with missing income information were, in general, more likely to be socioeconomically disadvantaged.

Missing values cannot be avoided, and naturally, the best solution is to minimize missing values at the point of collection. However, this may not be possible in most of cases. Hence, researchers should, first and foremost, carefully examine the profiles of respondents with missing information prior to analysis.[3] Secondly, researchers should keep in mind the possible bias which can be introduced by missing values. Finally, modern alternative techniques for working with missing values, such as single or multiple imputation, or full information maximum likelihood approaches should be introduced to the analysis.[7]

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Juster FT, Smith JP. Improving the quality of economic data: lessons from the HRS and AHEAD. J Am Stat Assoc 1997; 92:1268-78.
2. Roth PL. Missing data: a conceptual review for applied psychologists. Pers Psychol 1994;47:537-60.
3. Kim S, Egerter S, Cubbin C, Takahashi ER, Braveman P. Potential implications of missing income data in population-based surveys: an example from a postpartum survey in California. Public Health Rep 2007;122:753-63.
4. Vach W, Blettner M. Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. Am J Epidemiol 1991;134:895-907.
5. Graham JW, Donaldson SI. Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of follow-up data. J Appl Psychol 1993; 78:119-28.
6. Winkleby MA, Cubbin C. Influence of individual and neighbourhood socioeconomic status on mortality among black, Mexican-American, and white women and men in the United States. J Epidemiol Community Health 2003;57:444-52.
7. He Y. Missing data analysis using multiple imputation: getting to the heart of the matter. Circ Cardiovasc Qual Outcomes 2010;3:98-105.