The Effects of Measurement Error on Health Estimates in Web-Based Versus Face-to-Face National Health Surveys

Leanna P. Moron[1]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850

Abstract

Objectives: This report explores the effects of measurement error on seven national health estimates among the general U.S. adult population based on two data sources: a web-based survey, the third round of the Research and Development Survey (RANDS 3), and an in-person household survey, the 2019 National Health Interview Survey (2019 NHIS). These seven health estimates are based on five physical health variables and two mental health variables. The five physical health variables include *ever diagnosed by a physician or other medical professional with asthma, diabetes, high blood pressure or hypertension, high cholesterol, or chronic obstructive pulmonary disease (COPD).* The two mental health variables are *mild, moderate, and severe levels of major depressive disorder (depression)* and *generalized anxiety disorder (GAD)*.

Methods: Data from the 2019 NHIS (n = 31,997) and RANDS 3 (n = 2,646) were used to produce the weighted national estimates. The statistical analysis included two main components: (1) comparing weighted estimates by data source and conducting Rao-Scott significance testing to detect initial evidence of significant differences by data source and (2) building logistic regression models for each of the seven health outcomes and conducting Wald tests to determine the statistical significance of interaction terms. The interaction terms are between the RANDS indicator and age groups (45 to 64 years, and 65 years and older) as well as interactions between the RANDS indicator and the education groups (some college, bachelor's degree or above).

The initial logistic regression models contain a 16 predictor variables, including a RANDS indicator, sociodemographic characteristics (gender, age group, race and Hispanic origin, education, census region), and four interaction terms. The four interaction terms include the interaction between RANDS and two age groups (45 to 64 years, and 65 years and older) as well as the interaction between RANDS and two education groups (some college, and bachelor's degree or above). These interaction terms help examine the potential effects on the health outcome variables based on data source and age group, or based on data source and education group.

Results: Based on the results of the Rao-Scott significance testing, we find that the national health estimates based on RANDS 3 are significantly higher compared to the estimates based on the 2019 NHIS for six out of the seven health outcomes (asthma, high blood pressure, high cholesterol, COPD, depression, and GAD). The RANDS 3 and 2019 NHIS estimates were comparable for only one health outcome: diabetes. Among the seven final logistic regression models, three final models included no interaction terms (asthma, high blood pressure, and GAD), three final models included interaction terms between the RANDS indicator and both age groups (diabetes, high blood pressure, and COPD), and one final model included all four interaction terms (depression).

Conclusions: These findings suggest that the health information collected through RANDS differs in a systematic way from the information collected through NHIS, such that respondents more frequently reported having health conditions in the web-based survey compared to the in-person survey. This study makes the argument that these systematic differences are primarily due to measurement error, as opposed to other sources of survey error (i.e., coverage error, nonresponse

error). Although the reason for the observed differences in the estimates between the two data sources is unknown, a potential explanation could be social desirability bias.

Key Words: web survey, face-to-face survey, total survey error, secondary data analysis

1. Introduction

In-person household surveys have been traditionally regarded as a gold-standard approach for conducting data collection due to their relatively higher response rates in comparison to alternative survey modes (Biemer et al., 2021). However, due to declining response rates in recent decades (Brick & Williams, 2018), researchers have become increasingly interested in alternative survey modes. The major advantages of web-based surveys include the ability to collect more timely and cost-effective data in comparison to traditional, face-to-face surveys (Braekman et al., 2022; Heiervang & Goodman, 2011; Tourangeau et al., 2013). However, there is limited research on understanding the potential mode effects of web-based versus face-to-face surveys in terms of survey error and data quality, particularly in the context of national health surveys.

One meta-analysis has found that web-based surveys have an average response rate of 12 percentage points lower than other survey modes (Daikeler, 2020). Therefore, web-based surveys are at risk of achieving lower response rates in comparison to traditional survey modes. Research also suggests that measurement error may be of concern when it comes to the use of web-based surveys in place of face-to-face surveys (Couper et al., 2018).

1.1 Overview of Study

This paper explores the effects of measurement error on national health estimates introduced by a web-based panel survey, the third round of the Research and Development Survey (RANDS 3), and a primarily face-to-face survey, the 2019 National Health Interview Survey (2019 NHIS). These two data sources are from the Centers for Disease Control and Prevention (CDC) National Center for Health Statistics (NCHS). NCHS is the nation's principal health statistics agency, monitoring the health of the U.S. through various data collection systems. The RANDS 3 data was collected by NORC at the University of Chicago's AmeriSpeak Panel (https://amerispeak.norc.org/Pages/default.aspx) in April 2019. This study focuses on the following research question: What are the effects of measurement error introduced by web-based and face-to-face national health surveys on select physical and mental health estimates?

Through a secondary data analysis, the study examines the effects of measurement error on seven health outcomes, including five physical health variables and two mental health variables. The five physical health variables are whether or not the respondent has ever been diagnosed with asthma, diabetes, high blood pressure or hypertension, high cholesterol, or chronic obstructive pulmonary disease (COPD). The two mental health variables are having mild, moderate, or severe levels of major depressive disorder (depression) and generalized anxiety disorder (GAD) in 2019.

The statistical analysis includes two major components. First, design-adjusted Rao-Scott chi-squared tests are conducted to determine initial evidence of statistically significant differences between the health estimates by data source. Second, logistic regression models are fitted for each health outcome. The initial models include a predictor variable to indicate the data source (referred to as the "RANDS indicator"), select sociodemographic characteristics (sex, age, race and Hispanic origin, education, and census region), as well as interaction terms between the RANDS indicator and age, and interactions between RANDS and education. These interaction terms were included to examine the possibility that mode effects vary across these subgroups based on age and education.

Note that, although prior research compared comparing health estimates between RANDS and NHIS (He et al., 2020), those studies did not specifically focus on the effects of measurement error.

1.2 Total Survey Error Framework

Based on the total survey error framework, survey error may be attributed to multiple sources, including coverage, nonresponse, measurement, and processing error. In this study, we attribute measurement error as the primary source of error associated with the differences in health estimates between the two data sources, RANDS 3 and the 2019 NHIS. We make this argument because the coverage, nonresponse, and processing errors are anticipated to have little to no effect on the health estimates, due to the survey design and final survey weights. Both data sources have probability-based sample designs that control for selection biases associated with nonprobability-based surveys. In addition, nonresponse error is expected to be minimal because the postsurvey adjustment weighting approaches used for each data source account for differential nonresponse across subgroups. Third, processing error is expected to be minimal due to the quality control procedures used by both surveys. Therefore, based on the total survey error framework, this study attributes differences in the health estimates between RANDS 3 and 2019 NHIS to be primarily based on measurement error.

2. Data Sources

The comparisons were based on U.S. adults ages 18 years and older, using two public-use data sources from NCHS: RANDS 3 (n = 2,646), which was fielded in 2019, and the 2019 NHIS Sample Adult data file (n = 31,997). Both data sources are probability-based, complex surveys that include variables representing the final survey weights, primary sampling units (PSUs), and strata.

2.1 Research and Development Survey Round III

Established in 2015 by NCHS, RANDS is a primarily web-based series of cross-sectional surveys from probability-based commercial panel surveys. The target population of RANDS is the general U.S. adult population, ages 18 years and older. The third round of RANDS (referred to as RANDS 3), was administered as a web-based survey by NORC at the University of Chicago's AmeriSpeak panel between April 11 and April 24, 2019. Out of the 4,255 invitations sent to web-based panelists to participate in the survey, there were 2,646 completed interviews for an overall completion rate or conditional response rate of 62.2 percent. The specific American Association for Public Opinion Research (AAPOR) response rate formula used was not specified in the technical documentation. The sampling design stratifies the target population by the following characteristics: age, race and Hispanic origin, gender, and education. Of note, information about the response rate from earlier stages of the recruitment was not publicly available.

2.2 2019 National Health Interview Survey

Since 1957, the NHIS has been one of the major surveys and data collection systems of NCHS. The NHIS is a nationally representative, cross-sectional, population health survey. The inferential target population of the NHIS is the non-institutionalized, civilian, U.S. adult population, ages 18 years and older. The 2019 NHIS (n = 31,997) was conducted primarily as an in-person household survey, specifically through computer-assisted personal interviewing. The final reported response rate of the 2019 NHIS sample adult data file was 59.1 percent. The 2019 NHIS technical documentation does not specify which, if any, AAPOR response rate formula was used to calculate the final reported response rate.

3. Methods

3.1 Key Study Variables

This study focuses on seven health outcomes, including five physical health variables (asthma, diabetes, high blood pressure or hypertension, high cholesterol, and COPD) and two mental health variables (depression and GAD). The physical health variables indicate whether the respondent has ever been diagnosed with the condition by a physician or other health professional. The mental health conditions are any level of major depressive disorder symptoms and any level of GAD symptoms. All seven key study variables are binary variables, where 1 represents "yes," or the presence of the health condition, and 0 represents "no," or the absence of the health condition. The binary mental health outcome variables are based on guidance published by the Washington Group

on Disability Statistics (2021). For depression and GAD, the variables are recoded such that 1 indicates that the respondent has mild, moderate, or severe levels, and 0 indicates that the respondent has none to minimal levels.

3.2 Auxiliary Variables

The data includes the following demographic characteristics of the respondent: sex, age, race and Hispanic origin, and census region. For both surveys, sex is reported as male or female. The age variable was recoded into the following age groups: 18- to 44-year-olds, 45- to 64-year-olds, and 65 years and older. The race and Hispanic origin variable was recoded into the following four categories: Hispanic, non-Hispanic White, non-Hispanic Black, and non-Hispanic other or multiple races. The four census-defined regions are the Northeast, West, South, and Midwest.

3.3 Statistical Analysis: Weighted Estimates and Rao-Scott Significance Testing

The complex sample design features i.e., final survey weights, PSU, strata) of RANDS 3 and the 2019 NHIS are accounted for in the statistical analyses by using the R "survey" package created by Thomas Lumley. To examine the effects of measurement error on the selected health estimates, weighted percentages and standard errors are calculated for the seven key health variables by data source. The survey design objects are specified by using the svydesign() function, and the weighted estimates are calculated by using the svymean() function. In addition, design-adjusted Rao-Scott chi-squared tests are conducted to examine initial evidence of significant differences between the RANDS 3 and 2019 NHIS estimates. The svychisq() function is used to conduct the Rao-Scott significance testing.

3.4 Statistical Analysis: Logistic Regression Modeling

Separate logistic regression models are fitted for each key outcome variable. The logistic regression models include a predictor for the data source (where 0 = data source is the 2019 NHIS, 1 = data source is RANDS 3) (referred to as the "RANDS indicator"), as well as the following sociodemographic characteristics: sex, age, race and Hispanic origin, and census region. In the model output, the beta coefficients of the RANDS indicator are examined to determine whether the 95 percent confidence interval contains 0 or not. The models account for non-independent observations introduced by cluster sampling by accounting for the complex design features when computing standard errors, specifically by using Taylor series linearization for variance estimation.

The initial logistic regression model contains a binary health outcome variable denoted by y_{ij} where 1 indicates the presence of the health condition and 0 indicates otherwise. The initial logistic regression models contain 16 predictor variables, including 12 main predictor variables and 4 interaction terms. The 12 main predictor variables include the following: the RANDS indicator, male, 45 to 64 years, 65 years and older, non-Hispanic White, non-Hispanic Black, non-Hispanic other or multiple races, some college, bachelor's degree or greater, Midwest, West, and South. The four interaction terms include the interaction between RANDS and two age groups (45 to 64 years, and 65 years and older) as well as the interaction between RANDS and two education groups (some college, and bachelor's degree or above). In addition, Wald test statistics were produced by using the regTermTest() function in the survey package to help determine the statistical significance of the interaction terms (see equation 1).

In terms of the reference levels, the reference level for age is the age group of 18 to 44 years. For race and Hispanic origin, the reference level is Hispanic. For education, the reference level is having a high school level of education or less. For the census region, the reference level is the Northeast region of the United States. For more information about the predictor variables included in the initial logistic regression models, see Appendix Table A1.

$$log\left(\frac{Pr(y_{ij}=1)}{Pr(y_{ij}=0)}\right) = \beta_0 + \beta_1 RANDS + \beta_2 MALE + \beta_3 AGE45to64 + \beta_4 AGE65UP + \beta_5 WHITE$$

$$+ \beta_6 BLACK + \beta_7 OTHER + \beta_8 SOMECOL + \beta_9 BA + \beta_{10} WEST + \beta_{11} SOUTH + \beta_{12} MIDWEST + \beta_{13} RANDS{:}\, AGE45to64 + \beta_{14} RANDS{:}\, AGE65UP + \beta_{15} RANDS{:}\, SOMECOL + \beta_{16} RANDS{:}\, BA$$

Equation 1. Initial logistic regression model

4. RESULTS

4.1 Results: Weighted Estimates and Rao-Scott Significance Testing

Presented in **Table 1** are the weighted national health estimates of the U.S. population based on RANDS 3 and 2019 NHIS for the following seven health conditions: asthma, diabetes, high blood pressure, high cholesterol, COPD, depression, and GAD.

Based on the results of the Rao-Scott significance testing, the RANDS 3 estimates are significantly higher compared to the 2019 NHIS estimates for six out of the seven health conditions (asthma, high blood pressure, high cholesterol, COPD, depression, and GAD). For instance, the 2019 NHIS estimates that about 24.85 percent of U.S. adults have been diagnosed with high cholesterol at some point in their lifetime. However, based on RANDS 3, the prevalence of ever being diagnosed with high cholesterol among the general U.S. adult population is estimated to be closer to 36.44 percent.

Results suggest that the health estimates were comparable between the two data sources for only one out of the seven health outcomes: ever diagnosed with diabetes. Based on RANDS 3 and the 2019 NHIS, nearly 1 in 10 adults in the United States have ever been diagnosed with diabetes by a physician or other medical professional. More specifically, this estimate is 9.35 percent based on the 2019 NHIS, and 10.55 percent based on RANDS 3.

Table 1: Unweighted sample sizes and weighted percentages of physical and mental health conditions among adults aged 18 years and older: United States, 2019

	RANDS 3 (n = 2,646)			*2019 NHIS (n = 31,997)*			
	n	*Percent*	*SE*	*n*	*Percent*	*SE*	*Rao-Scott*[+]
Physical health variables[1]							
Asthma	2,628	16.86	0.98	31,947	13.46	0.25	13.30***
Diabetes	2,631	10.55	0.8	31,949	9.35	0.20	2.36
High blood pressure	2,637	37.00	1.21	31,938	31.65	0.38	18.89***
High cholesterol	2,633	36.44	1.19	31,876	24.85	0.32	106.67***
COPD[2]	2,633	8.40	0.72	31,945	4.62	0.15	42.12***
Mental health variables[3]							
Depression	2,525	28.05	1.19	31,253	15.32	0.28	50.94***
GAD	2,536	46.33	1.32	31,240	35.68	0.41	37.69***

Notes:
[1] SE stands for standard error.
[2] The physical health variables (asthma, diabetes, hypertension, high cholesterol, and COPD) refer to respondents who responded "yes" to the question "Have you ever been diagnosed [with this condition] by a physician or other health professional?"
[3] COPD refers to chronic obstructive pulmonary disease.
[4] The mental health variables include any (mild, moderate, or severe) levels of depressive symptomatology and any (mild, moderate, or severe) levels of generalized anxiety disorder (GAD) symptomatology.
[+] Rao-Scott chi-squared tests are based on a significance level of 0.05.
[*] p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

4.2 Results: Wald Test Statistics and Logistic Regression Modeling

Presented in **Table 2** are the beta coefficients and standard errors of the final logistic regression models for the seven health outcomes. These health outcomes include asthma, diabetes, high blood pressure or hypertension, high cholesterol, COPD, depression, and GAD. As previously described in the Methods section, the initial logistic regression models contained a total of 16 predictor variables. These variables include the RANDS indicator, sociodemographic characteristics (sex, age group, race and Hispanic origin, education, census region), and interaction terms between the RANDS indicator and age groups (45 to 64 years, and 65 years and older), as well as interaction terms between the RANDS indicator and the education groups (some college, bachelor's degree or above).

4.3 Model 1: Ever Diagnosed with Asthma

Based on the initial logistic regression model for asthma, none of the interaction terms were statistically significant based on the Wald test statistics. The final model for asthma contains only the 12 main predictor variables (RANDS indicator, male, 45 to 64 years, 65 years and older, non-Hispanic White, non-Hispanic Black, non-Hispanic other or multiple races, West, South, Midwest, high school graduate or less, and some college-level education) and no interaction terms.

Based on the final model for asthma, 8 out of the 12 predictor variables were statistically significant predictors: the RANDS indicator, male, age group 45 to 64 years, non-Hispanic White, non-Hispanic Black, bachelor's degree or above, and the West and South census regions. The individuals in the population represented by RANDS 3 have significantly greater odds (1.26 times) of having ever been diagnosed with asthma in comparison to respondents in the 2019 NHIS, after adjusting for the other covariates in the model.

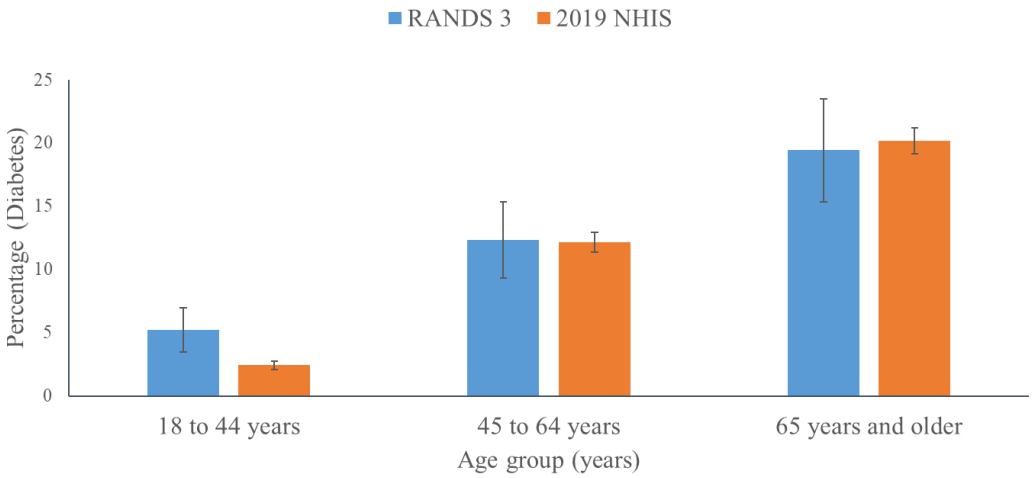
4.4 Model 2: Ever Diagnosed with Diabetes

Based on the initial model for diabetes, two out of the four interaction terms were statistically significant predictors: the two-way interactions between the RANDS indicator and each age group (45 to 64 years, and 65 years and older). The interaction terms between the RANDS indicator and both education groups (some college, bachelor's degree or above) were not statistically significant predictors of having ever been diagnosed with diabetes.

Based on the final model for diabetes, 11 out of 14 predictor variables were found to be statistically significant. The significant predictors were the following: the RANDS indicator, male, 45 to 64 years, 65 years and older, non-Hispanic White, non-Hispanic other or multiple races, some college, bachelor's degree or higher, West, the interaction term between the RANDS indicator and 45 to 64 years, and the interaction term between the RANDS indicator and 65 years and older.

Based on the interaction terms in the final model, for respondents in RANDS 3, the odds of having ever been diagnosed with diabetes are 2.27 times higher compared to NHIS respondents aged 18 to 44 years, while holding the other covariates in the model fixed. In addition, the odds of having ever been diagnosed with diabetes are 1.09 times higher for RANDS individuals between the ages of 45 to 64 years old, and 1.08 times higher for individuals who are 65 years and older, in comparison to the NHIS respondents. These results suggest that the effects of RANDS 3 are significantly weaker for respondents in the older age groups in comparison to respondents in the youngest age group.

Depicted in **Figure 2** are the weighted percentages of U.S. adults who have ever been diagnosed with diabetes by data source and age group. We observe that the prevalence of ever being diagnosed with diabetes is more common among the older age groups (45 to 65 years, and 65 years and older) in comparison to the youngest age group, 18 to 44 years. The effect of the data source for RANDS 3 respondents is most pronounced for the youngest age group, specifically for *ever been diagnosed with diabetes*.



Note: 95 percent confidence intervals, shown with error bars.

Figure 2. Percentage of adults aged ⩾ 18 years who have ever been diagnosed with diabetes, by data source and age group

4.5 Model 3: Ever Diagnosed with High Blood Pressure or Hypertension

Based on the initial model for high blood pressure, none of the interaction terms were statistically significant predictors. Based on the final model for high blood pressure, 8 out of 12 predictors were found to be statistically significant. The significant predictors were the RANDS indicator, male, 45 to 64 years, 65 years and older, non-Hispanic White, some college, bachelor's degree or above, and West. When adjusting for the other covariates in the model, the individuals in the population represented by RANDS 3 have significantly greater odds (1.34 times) of having ever been diagnosed with high blood pressure or hypertension in comparison to the respondents in the 2019 NHIS.

4.6 Model 4: Ever Diagnosed with High Cholesterol

Based on the initial model for high cholesterol, the Wald test statistics indicate that two out of the four interaction terms were statistically significant: the interactions between the RANDS indicator and both age groups (45 to 64 years, and 65 years and older).

Based on the final model for high cholesterol, 9 out of the 14 predictor variables were statistically significant (RANDS indicator, male, 45 to 64 years, 65 years and older, bachelor's degree or above, West, South, interaction term between RANDS indicator and 45 to 64 years, and interaction term between RANDS indicator and 65 years and older).

Among individuals between the ages of 18 to 44 years, the odds of having ever been diagnosed with high cholesterol are 2.5 times greater for individuals in the population represented by RANDS 3 compared to the respondents in the 2019 NHIS, while holding other covariates in the model fixed. Based on the interaction terms in the final model, the odds of having ever been diagnosed with high cholesterol are higher for RANDS individuals between the ages of 45 to 64 years, as well as those that are 65 years and older (1.80 and 1.54 times, respectively), in comparison to the 2019 NHIS respondents, when adjusting for the other covariates in the model.

4.7 Model 5: Ever Diagnosed with Chronic Obstructive Pulmonary Disease

Based on the initial model for COPD, two out of the four interaction terms were found to be statistically significant: the interaction terms between the RANDS indicator and both age groups (45 to 64 years, and 65 years and older). Based on the final model for COPD, all 14 out of the 14 predictor variables are statistically significant predictors.

Among respondents between the ages of 18 to 44 years, the odds of having ever been diagnosed with COPD are 3.4 times higher for those in RANDS 3 in comparison to those in the 2019 NHIS, while adjusting for the other covariates in the model. Based on the interaction terms, we find that older respondents have significantly greater odds of having ever been diagnosed with COPD in RANDS 3 versus the 2019 NHIS (1.89 times higher for individuals between 45 to 64 years old, and 1.49 times higher for individuals that are 65 years old or older), while adjusting for the other covariates in the model.

4.8 Model 6: Depression – Mild, Moderate, or Severe Levels

Based on the initial model for depression, all four interaction terms are statistically significant predictors. The significant interaction terms include the interactions between the RANDS indicator and both age groups (45 to 64 years, and 65 years and older), as well as the interactions between the RANDS indicator and both education groups (some college, and bachelor's degree or above).

Based on the final model for depression, 13 out of the 16 predictor variables were statistically significant (RANDS indicator, male, 45 to 64 years, 65 years and older, non-Hispanic White, non-Hispanic other or multiple races, some college, bachelor's degree or above, South, and all four interaction terms).

The odds of having mild, moderate, or severe levels of depression symptoms are 2.2 times higher for respondents in RANDS 3 compared to the 2019 NHIS, specifically among individuals who are 18 to 44 years old and who have no more than a high school level of education. There is little to no effect of the RANDS indicator for individuals between the ages of 45 to 64. RANDS 3 respondents who are 65 years and older have 0.68 times lower odds of having mild, moderate, or severe symptoms of depression, in comparison to respondents in the 2019 NHIS. In terms of education, RANDS 3 respondents with some college-level education or at least a bachelor's degree have greater odds of having depression in comparison to respondents in the 2019 NHIS (4.22 and 3.90 times, respectively).

4.9 Model 7: Generalized Anxiety Disorder – Mild, Moderate, or Severe Levels

Based on the initial model for GAD, we found none of the interaction terms to be statistically significant. However, based on the final model for GAD, we found 8 out of the 12 predictors to be statistically significant. These significant predictors include the following: the RANDS indicator, male, 45 to 64 years, 65 years and older, non-Hispanic White, non-Hispanic other or multiple races, some college, and South. The predictor variables that were not found to be significant include non-Hispanic Black, bachelor's degree or above, Midwest, and West.

The individuals in the population represented by RANDS 3 have 1.93 times the odds of having mild, moderate, or severe levels of GAD in 2019 compared to those from the NHIS data source, adjusting for other covariates in the model.

Table 2: Beta coefficients and standard errors (in parentheses) of final logistic regression models for health outcomes

Predictor Variable	***Asthma***	***Diabetes***	***High Blood Pressure***	***High Cholesterol***	***COPD***	***Depression***	***GAD***
Intercept	-2.39 (0.10)***	-3.84 (0.10)***	-1.88 (0.06)***	-2.36 (0.06)***	-4.09 (0.13)	-1.58 (0.07)***	-0.37 (0.05)***
RANDS	0.23 (0.07)**	0.82 (0.20)***	0.30 (0.07)***	0.92 (0.11)***	1.23 (0.23)***	0.80 (0.25)**	0.66 (0.11)***
Male	0.21 (0.04)***	-0.10 (0.05)*	-0.16 (0.03)***	-0.18 (0.03)***	0.21 (0.06)***	0.51 (0.04)***	0.58 (0.03)***
45 to 64 years	0.43 (0.05)***	1.79 (0.08)***	1.59 (0.04)***	1.72 (0.05)***	1.43 (0.11)***	-0.17 (0.05)***	-0.48 (0.04)***
65 years+	0.08 (0.06)	2.41 (0.08)***	2.55 (0.05)***	2.54 (0.05)***	2.00 (0.11)***	-0.45 (0.06)***	-1.02 (0.04)***
White, non-Hispanic	0.25 (0.07)***	0.50 (0.10)***	0.52 (0.05)***	-0.09 (0.05)	-0.41 (0.10)***	-0.34 (0.07)***	-0.55 (0.06)***
Black, non-Hispanic	0.31 (0.09)***	0.17 (0.29)	0.29 (0.21)	0.15 (0.20)	0.34 (0.33)***	-0.15 (0.30)	0.13 (0.37)
Other, non-Hispanic	0.43 (0.29)	0.45 (0.08)***	-0.10 (0.05)	0.10 (0.05)	-0.71 (0.13)***	-0.54 (0.07)***	-0.65 (0.05)***
Some college	0.08 (0.05)	-0.18 (0.06)**	-0.14 (0.04)***	-0.06 (0.04)	-0.31 (0.08)***	-0.12 (0.05)*	0.08 (0.04)*
Bachelor's+	0.23 (0.05)***	-0.67 (0.06)***	-0.47 (0.04)***	-0.12 (0.04)**	-1.42 (0.08)***	-0.57 (0.05)***	0.04 (0.04)
Midwest	-0.07 (0.07)	0.16 (0.08)	-0.01 (0.06)	0.02 (0.06)	0.26 (0.10)**	0.09 (0.07)	-0.01 (0.05)
West	-0.20 (0.07)**	0.31 (0.07)***	0.22 (0.05)***	0.12 (0.05)*	0.41 (0.09)***	0.03 (0.07)	-0.04 (0.05)
South	-0.15 (0.06)*	0.03 (0.09)	-0.03 (0.06)	-0.11 (0.05)*	0.25 (0.12)*	0.16 (0.07)*	0.14 (0.05)*
RANDS:45 to 64 years		-0.73 (0.23)**		-0.33 (0.14)*	-0.60 (0.28)*	-0.80 (0.22)***	
RANDS:65 years+		-0.74 (0.24)**		-0.49 (0.16)**	-0.83 (0.28)**	-1.18 (0.26)***	
RANDS:some college						0.64 (0.25)**	
RANDS:bachelor's+						0.56 (0.26)*	

Note: * p-value < 0.05; ** p-value < 0.01; *** p-value < 0.001

5. Discussion

5.1 Summary of Findings: Weighted Estimates and Rao-Scott Significance Testing

The main objective of this study was to investigate the effects of measurement error on select physical and mental health estimates in a web-based survey (RANDS 3) and a traditional primarily face-to-face survey (the 2019 NHIS). Based on the results of the weighted estimates and Rao-Scott significance testing, we found that the RANDS estimates were significantly higher compared to NHIS estimates for six out of the seven health outcomes: asthma, high blood pressure or hypertension, high cholesterol, COPD, depression, and GAD. Based on the results of the Rao-Scott significance testing, the RANDS 3 and 2019 NHIS estimates were comparable for only one health outcome: diabetes. Based on both data sources, it is estimated that nearly 1 in 10 adults in the United States have ever been diagnosed with diabetes by a physician or other medical professional.

5.2 Summary of Findings: Logistic Regression Modeling

The initial logistic regression models contain 16 predictor variables: the RANDS indicator, gender (male), age group (45 to 64 years, 65 years and older), race and Hispanic origin (non-Hispanic White, non-Hispanic Black, non-Hispanic other or multiple races), education (some college, bachelor's degree or greater), census region (Midwest, West, and South) and four interaction terms. The four interaction terms include the interaction between RANDS and two age groups (45 to 64 years, and 65 years and older) as well as the interaction between RANDS and two education groups (some college, and bachelor's degree or above).

Among the seven final logistic regression models, three final models included no interaction terms (asthma, high blood pressure, and GAD), three final models included interaction terms between the RANDS indicator and both age groups (diabetes, high blood pressure, and COPD), and one final model included all four interaction terms (depression).

The three final models that did not include any significant interaction terms (asthma, high blood pressure, and GAD) allow us to better understand the main effects of the RANDS indicator. In summary, the persons in the population represented by RANDS 3 have significantly greater odds of having ever being diagnosed with asthma, of having ever being diagnosed with high blood pressure, and having mild, moderate, or severe levels of GAD symptomatology, in comparison to the respondents in the 2019 NHIS (while adjusting for the other covariates in the model).

Based on the final logistic regression models for diabetes, high cholesterol, and COPD, we conclude that the web-based survey effects of RANDS 3 are significantly weaker for respondents in the older age groups (45 to 64 years, and 65 years and older) in comparison to the youngest age group (18 to 44 years). For diabetes, while the findings based on the Rao-Scott significance testing suggest that there are no significant differences in the overall population of RANDS 3 versus the 2019 NHIS, the logistic regression results suggest that there are significant differences for particular subpopulations based on their age group.

In terms of having mild, moderate, or severe levels of depression, the effects of RANDS 3 are significantly weaker for older age groups (45 to 64 years, and 65 years and older) in comparison to the youngest age group (18 to 44 years). In addition, we conclude that the RANDS effect is also significantly weaker for those with higher levels of education (some college, or bachelor's degree or more) in comparison with those with only a high school degree or less in face-to-face interviews. Social desirability bias could be a potential reason behind the underreporting of chronic health conditions such as depression.

5.3 Limitations

While this study has several strengths, such as the comparison of two probability-based surveys, we also acknowledge the limitations of this research. The first major limitation of this study is that more detailed model diagnostics are necessary. In particular, conducting design-adjusted goodness-of-fit tests on the logistic regression models would be helpful information to assess model fit. This would

allow us to continue the model-building process and ultimately work toward improving the final logistic regression models. The second limitation is that there are slight differences in the target population of the two data sources. As described in the technical documentation, the target population for RANDS 3 is the general U.S. adult population, ages 18 years and older, and the target population for the 2019 NHIS is the non-institutionalized, U.S. adult population, ages 18 years and older. The term "non-institutionalized" is referenced in the documentation for the 2019 NHIS but not for RANDS 3. Another limitation is that RANDS 3 used a raking approach to create calibration weights based on select demographic and health variables from the 2019 NHIS as the reference dataset. These select health variables used for calibration include some of the key outcome variables examined in this paper, including *ever diagnosed with asthma, high cholesterol, high blood pressure or hypertension, and diabetes*.

Another weakness is the assumption that only measurement error contributes to the total survey error for each of the surveys. The response rate published for the RANDS survey does not provide the response rate for recruiting these respondents into the sample. Once doing this, it would clearly push the response rate significantly below the NHIS. While both surveys make rigorous adjustments for nonresponse, this may not have controlled for non-random nonresponse.

5.4 Social Desirability Bias: A Potential Reason for Observed Differences

While we do not know for certain why most of the health estimates were significantly higher based on a web-based survey (RANDS 3) in comparison to the face-to-face survey (the 2019 NHIS), a potential reason could be social desirability bias. Social desirability bias can be described as survey respondents' tendency to present themselves in a favorable light by underreporting socially undesirable behaviors or attributes and overreporting socially desirable ones (Groves et al., 2009). In other words, respondents, or select subpopulations of respondents, might underreport diagnoses of physical or mental health conditions.

5.5 Research Implications and Future Research

These findings suggest that the health information collected through RANDS differs in a systematic way from the information collected through NHIS in some instances and for particular subpopulations. In addition, this paper makes the argument that these systematic differences can be attributed primarily to measurement error. As previously discussed, the survey design and final survey weights for RANDS 3 and the 2019 NHIS minimize potential coverage and nonresponse errors. Therefore, the differences in health estimates by data source in this study are arguably introduced specifically by measurement error.

Based solely on the results of the Rao-Scott significance testing, we determined that the RANDS 3 estimates are significantly higher compared to the 2019 NHIS estimates for all of the health outcomes except for diabetes—that is, for asthma, high blood pressure or hypertension, high cholesterol, COPD, depression, and GAD. However, based on the results of the logistic regression modeling, we found significant interaction terms between the respondent's age group and the RANDS indicator, allowing us to better understand that the effects of RANDS 3 are more salient for the youngest age group (18 to 44 years). This leads us to recommend that future research regarding measurement error consider examining interactions between survey mode and respondent sociodemographic characteristics, such as age group and education.

Future research is also necessary to assess how the effects of social desirability may be affecting the NHIS results. For example, a record check study could be conducted to compare respondents' medical records and self-reports. This type of study would help determine whether respondents are accurately reporting their health diagnoses, under the assumption that the medical records are the absolute truth.

The NCHS continues to research the impact of using calibration weights to improve web-based estimates from RANDS. Given that the NHIS is typically regarded as the gold-standard data source, one idea is to have the RANDS estimates become more similar to the NHIS estimates. Alternatively, researchers might consider calibrating health estimates based on the NHIS to a web-based data

source, such as RANDS, if it is believed that the web-based responses are more accurate due to its self-administrative mode. Further investigation in the areas of measurement error and survey mode effects is necessary for us to have a better understanding as to why some of these systematic differences are occurring for certain populations between RANDS and NHIS.

References

Biemer, P. P., Harris, K. M., Burke, B. J., Liao, D., & Halpern, C. T. (2021). Transitioning a panel survey from in-person to predominantly web data collection: Results and lessons learned. *Journal of the Royal Statistical Society, Series A, 185*(3), 798-821. doi: 10.1111/rssa.12750.

Biemer, P. P., Murphy, J., Zimmer, S., Berry, C., Deng, G., & Lewis, K. (2018). Using bonus monetary incentives to encourage web response in mixed-mode household surveys. *Journal of Survey Statistics and Methodology*, *6*(2), 240–261. doi: 10.1093/jssam/smx015.

Braekman, E., Demarest, S., Charafeddine, R., Drieskens, S., Berete, F., Gisle, L., & Van Hal, G. (2022). Unit response and costs in web versus face-to-face data collection: Comparison of two cross-sectional health surveys. *Journal of Medical Internet Research, 24*(1), e26299. doi: 10.2196/26299.

Brick, J. M. & Williams, D. (2018). Trends in U.S. face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, *6*(2), 186-211. doi: 10.1093/jssam/smx019.

Couper, M. P., Gremel, G., Axinn, W. G., Guyer, H., Wagner, J., &West, B. T. (2018). New options for national population surveys: The implications of internet and smartphone coverage. *Social Science Research, 73*(1), 221–235.

Daikeler, J., Bošnjak, M., & Lozar Manfreda, K. (2020). Web versus other survey modes: An updated and extended meta-analysis comparing response rates. *Journal of Survey Statistics and Methodology*, *8*(3), 513–539. doi: 10.1093/jssam/smz008.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology*. John Wiley & Sons.

He, Y., Cai, B., Shin, H.-C., Beresovsky, V., Parsons, V., & Irimata, K. (2020). The National Center for Health Statistics' 2015 and 2016 Research and Development Surveys. National Center for Health Statistics. *Vital Health Statistics, 1*(64). https://www.cdc.gov/nchs/data/series/sr_01/sr01-64-508.pdf

Heiervang, E., & Goodman, R. (2011). Advantages and limitations of web-based surveys: evidence from a child mental health survey. *Social Psychiatry and Psychiatric Epidemiology, 46*(1), 69–76.

Lumley, T., Gao, P., Schneider, B. (2024). R "survey" package documentation: Analysis of complex survey samples. (2024). https://cran.r-project.org/web/packages/survey/survey.pdf

National Center for Health Statistics (NCHS). National Health Interview Survey, 2019. Public-use data file and documentation. https://www.cdc.gov/nchs/nhis/2019nhis.htm

National Center for Health Statistics (NCHS). Research and Development Survey (RANDS) 3 public-use data file and technical documentation. Hyattsville, MD: NCHS, 2020. https://www.cdc.gov/nchs/rands/data.htm

Scott, A. J., & Rao, J. N. K. (1981). Chi-squared tests for contingency tables with proportions estimated from survey data. In *Current Topics in Survey Sampling* (pp. 247–265), Academic Press.

Tourangeau, R., Frederick, C., & Couper, M. (2013). *The science of web surveys.* New York: Oxford University Press.

Washington Group on Disability Statistics. (2021). *Analytical guidelines: Creating disability identifier using the Washington Group Short Set on Functioning – Enhanced (WG-SS Enhanced) SPSS Synthax.* https://www.washingtongroup-disability.com/fileadmin/uploads/wg/WG_Document__7A_-_Analytic_Guidelines_for_the_WG-SS_Enhanced__SPSS_.pdf

Appendix

Table A1: Predictor variables for initial logistic regression models

	Predictor Variable	*Description*
RANDS	RANDS	1 = RANDS 3, 0 = 2019 NHIS
Sex	MALE	1 = male, 0 = female
Age Group	AGE45to64	1 = 45 to 64 years old, 0 = otherwise
	AGE65UP	1 = 65 years old or older, 0 = otherwise
Race and Hispanic Origin	WHITE	1 = non-Hispanic White, 0 = otherwise
	BLACK	1 = non-Hispanic Black, 0 = otherwise
	OTHER	1 = non-Hispanic other or multiple races, 0 = otherwise
Census Region	WEST	1 = North census region, 0 = otherwise
	SOUTH	1 = South census region, 0 = otherwise
	MIDWEST	1 = Midwest census region, 0 = otherwise
Education	SOMECOL	1 = some college, 0 = otherwise
	BA	1 = bachelor's degree or above, 0 = otherwise
Interaction Terms	RANDS: AGE45to64	Interaction term between RANDS indicator and age group 45 to 64 years
	RANDS: AGE65UP	Interaction term between RANDS indicator and age group 65 years and older
	RANDS: SOMECOL	Interaction term between RANDS indicator and having some college-level education
	RANDS:BA	Interaction term between RANDS indicator and having at least a bachelor's degree

The reference levels are the following: 18 to 44 years for age group, Hispanic for race and Hispanic origin, Northeast for Census region, and high school graduate or less for education.