Uninsurance Estimates from the Redesigned Current Population Survey. An Examination of the Imputation Process

By Heide Jackson

SEHSD Working Paper Number: 2019-20

DRB Number: CBDRB-FY19-ROSS-B0085

Abstract:

Objective: To evaluate the quality of imputation procedures in the Current Population Survey's (CPS) Annual Social and Economic Supplement (ASEC).

Data Source: The 2017 CPS ASEC production and 2017 CPS ASEC research files.

Study Design: Comparison of probability of health insurance coverage by imputation status,

Data Collection/Extraction Methods: Secondary analysis of household based data.

Principal Findings: After adjusting for factors related to health insurance, the production file showed significant differences between fully imputed and reported cases; no such differences were found in the research file.

Conclusions: Changes to imputation introduced in the research file have eliminated previous biases and have improved the distribution of multiple coverage types.

Key Words: Health insurance coverage, imputation, missing data

.

The Current Population Survey Annual Social and Economic Supplement (CPS ASEC) is an important and widely used source of information on the nation's income, labor force participation, and health insurance status. Data from the survey has been used widely by the research and policy community for a variety of purposes including the allocation of children's health insurance funds and to score congressional legislation [1,2]. The CPS ASEC has been noted for its many strengths including its large sample size, rich set of demographic, social, and economic measures, and timely release of its data.

Despite these strengths, the CPS ASEC also has some notable limitations. Namely, in past years, the survey had uninsured estimates higher than other federal surveys [3,4,5,6]. Research has identified several factors contributing to the higher uninsured rate in the CPS ASEC. Among these are that the CPS ASEC questions captured people's health insurance at the time of the interview or in recent months rather than over the course of year and that the Census Bureau's imputation methodology over-estimated the number of uninsured persons [5].

To respond to critiques and improve data quality, the CPS ASEC was redesigned in two stages[7,8]. First, the questionnaire was redesigned for calendar year 2013 to introduce new health insurance questions and an adaptive questionnaire design. Second, beginning with the 2017 CPS ASEC research file, improvements to the health insurance processing and imputation were rolled out to maximize information released from the redesigned questionnaire and address previously noted limitations of the imputation procedure.

In this paper, we examine whether and how new imputation procedures affect estimates of the uninsured population in 2016 using data from the 2017 CPS ASEC research file and production files. Specifically, we investigate whether the prevalence of major insurance types differs by imputation status in the 2017 CPS ASEC production file which used legacy imputation

procedures and in the 2017 CPS ASEC research file which introduces innovations in the imputation processing.

The CPS ASEC and Missing Data

The Current Population Survey (CPS) is a monthly survey conducted by the Census Bureau and sponsored by the Bureau of Labor Statistics. The Annual Social and Economic Supplement (ASEC) is added to the CPS survey from February through April. The ASEC collects information on a variety of topics including income from the prior (calendar) year as well as information on health insurance and health status [9]. Like other surveys, the CPS ASEC has missing data arising from unit and item non-response. For cases where a full or partial interview was conducted, the Census Bureau imputes information missing for these cases. In the 2017 CPS ASEC, 22% of cases had all health insurance information imputed.

For calendar year 2016, the Census Bureau released two files based on the same underlying data and which have the same missing data rate. The first file, the 2017 CPS ASEC production file, was based on legacy imputation and processing procedures. The second file, the 2017 CPS ASEC research file, features updated imputation and processing. Recent research suggests that although these files are based on the same underlying data, health insurance estimates are appreciably different [8,10]. Specifically, estimates of overall health insurance coverage increased in the research file while estimates of multiple coverage types declined [8,10]. This paper will be the first to examine whether changes to imputation procedures affected health insurance estimates across files.

Imputation Procedures in the CPS ASEC

In the CPS ASEC, the Census Bureau imputes health insurance data using a hotdeck imputation procedure [11]. Hotdeck imputation is a type of model-based imputation which takes information from cases with reported data, "donors", and assigns it to cases with missing data "recipients" [12]. Donors and recipients are matched so that they are comparable on observed demographic and social characteristics. Built into hotdeck imputation is the assumption that data are missing at random [12]. That is, controlling for relevant observed characteristics, the uninsured rate for imputed cases should be comparable to cases with reported values [5, 12].

Prior research revealed some limitations to legacy CPS ASEC hotdeck imputation procedures [5]. Prior to 2011, the health insurance imputation consisted of two parts. First, the imputation selected whether or not a person was the policyholder of a private insurance plan. If a person was identified as a policyholder, they were imputed to have either an individual or family plan. For policyholders of family plans, coverage was extended to other family members. This procedure was found to downwardly bias private insurance estimates in two ways. First, family size was not included when imputing whether or not a person was a policyholder of an insurance plan which proved problematic because the likelihood of being a policyholder varies by household size. Additionally, each insurance type was imputed separately and independently, increasing the prevalence of multiple insurance coverage among imputed cases.

Beginning in 2011, a few modifications to the CPS ASEC imputation procedures sought to address these shortcomings[13]. First, for private insurance, all non-policyholders were allowed to be imputed as dependents on health insurance plans. Next, the variables used in the hotdecks were updated to be more comprehensive and, importantly for private health insurance, include a measure of family size. Finally, to improve imputation of public health insurance, public insurance was imputed prior to private insurance and was factored into the hotdecks when

imputing private coverage. While these changes improved estimates of health insurance [13], they did not fully account for the imputation bias, nor did they address the higher than expected prevalence of dual health insurance coverage for fully imputed cases.

With the 2017 CPS ASEC research file, the health insurance imputation procedure was fully overhauled to further address these limitations [8]. In the research file, the Census Bureau constructed health insurance units (HIU) based on individuals who were eligible to share insurance coverage[15]. If no one in the HIU reported health insurance information, data for the entire HIU was imputed simultaneously by matching this recipient HIU with a donor HIU with completed health insurance information. This procedure was developed to address limitations of prior imputation strategies in three key ways. First, the new imputation procedure imputes all health insurance information simultaneously for all individuals within a health insurance unit; allowing the distribution of coverage to better match across donor and recipient cases. Second, the number of characteristics included in the hotdeck has been greatly expanded. Third and finally, this procedure allows better matching of multiple insurance types across imputed and reported cases.

Analytic Approach

To examine how changes to the health insurance imputation have affected estimates, we use the 2017 CPS ASEC production and research files. The 2017 CPS ASEC production file contains health insurance information processed using the legacy imputation procedure updated in 2011, and the 2017 CPS ASEC research file contains health insurance information processed using the new imputation procedure. In this analysis we make two adjustments; first, we only include people in-sample for both research and production files [14]. Second, we treat all military coverage as public coverage for the purposes of this analysis [8].

Our analysis consists of two parts. In the first, we perform a descriptive analysis of health insurance status by imputation status across the CPS ASEC production and research files. We examine health insurance at the person level by looking at the prevalence of any, private, public, and combined (private and public health insurance) by imputation status across files. As the health insurance imputation procedure is now done simultaneously for all people within a health insurance unit, we also examine coverage patterns at the household level looking at the percent of households where everyone in the household is: uninsured, covered by private insurance, covered by public insurance, covered by a combination of private and public insurance, and has some other coverage configuration (e.g. one person is uninsured and another has private insurance).

Next, we move to a regression-based analysis. On both the production and research files, we perform a multinomial logistic regression analysis where three outcomes are examined: (1) no insurance, (2) public insurance alone or in combination with private insurance, (3) private insurance. Given health insurance dynamics across age --health insurance is almost universal for adults over the age of 65 and persons under the age of 26 are eligible to be covered as dependents -- we restrict our sample to people under the age of 65 and run separate models for people under age 26 and working age adults 26 to 64. The key covariate in these models is a dichotomous variable indicating whether or not a person has health insurance coverage fully imputed. In the models run for people under the age of 26, age, race, sex, self-rated health, household poverty status, citizenship status, and country of birth are included as controls. In the models run for adults 26 to 64, we account for marital status, education, disability status, and labor force participation in addition to the controls included for persons under the age of 26.

The use of multinomial logistic regression with these outcomes is consistent with past approaches evaluating the quality of imputation in the CPS ASEC [5, 13]. If the relative risk ratio (RRR) for the imputation coefficient is not significantly different from 1, the imputation is considered unbiased. That is, to be consistent with the missing at random assumptions of hotdeck imputation, being fully imputed should not be associated with insurance status after controlling for relevant respondent characteristics.

Results

Consistent with findings from past studies [5,13], the production file shows large differences in the prevalence of any, public, private, and combined public and private coverage by imputation status at both the person and household level. Table 1 shows the person-level distribution of insurance coverage by imputation status across the research and production files. In the production file, fully imputed cases were 5.76 percentage points less likely to have coverage when compared with cases with some information reported. Significant differences by imputation status were found for all insurance types examined and were especially large for combined public and private coverage with imputed cases being 8.83 percentage points more likely to have this coverage pattern.

In the research file, differences by imputation status were greatly attenuated. Fully imputed cases were slightly less likely (0.68 percentage points) to have any insurance coverage, but there were no significant differences in the prevalence of private only and public only insurance. In contrast to the production file, fully imputed cases were less likely to have a combination of private and public coverage: 9.54 percent of imputed cases had both private and public coverage compared with 10.32 percent of cases with some reported data.

Table 2 shows distributions of insurance coverage by imputation status at the household level. In the production file, households with fully imputed health insurance information appeared different from households that had some information reported for most insurance types examined. Most notably, in households with two or more people, fully imputed households were much more likely to have people with different health insurance statuses than households with some reported information. In two person households, fully imputed cases are 6.48 percentage points more likely to have different coverage types; in households with three or more people, fully imputed cases are 13.14 percentage points more likely to have this profile.

In the research file, there were fewer significant differences between fully imputed and reported households. Again looking at households with different coverage types, we see differences between households with fully imputed versus some reported data, but they are of a smaller magnitude than those found in the production file. In a two-person household, fully imputed cases were 2.05 percentage points more likely to have different insurance types, and three person households were 3.28 percentage points more likely to have this profile. These small differences between fully imputed and reported cases are not unexpected as the new procedure imputes information simultaneously among people eligible to share health insurance coverage whereas during the course of the survey, people can report covering anyone within the household even if that person is not eligible to be covered under the same insurance plan. Overall, the descriptive analyses are suggestive; differences between fully imputed and reported cases appear to be attenuated in the research file. However, to detect whether either imputation procedure is biased, a regression-based analysis is needed.

Table 3 shows findings from the multinomial logistic regression analysis predicting coverage type. In the production file, cases with fully imputed data were less likely to have

public coverage, alone or in combination with private, or private coverage relative to cases that had some reported data, even after adjusting for other factors that should predict insurance status. For people under the age of 26 with imputed data, the relative risk ratios (RRR) were 0.78 for any public insurance and 0.54 for any private insurance, relative to being uninsured. For adults between the ages of 26 and 64, the RRR for fully imputed cases was 0.62 for public insurance alone or in combination and 0.41 for having exclusively private insurance. In contrast, the research file shows no such relationship between imputation status and insurance status. The RRRs for imputation status do not significantly differ from 1.00 for any of the insurance outcomes examined or for any age group.

Discussion

The CPS ASEC is a leading source of information on the nation's health insurance [1,2]. However, prior research noted that the uninsured rate was consistently higher than other federal surveys [3,4,5,6] in part due to limitations in imputation procedures[5,13]. Specifically, by imputing each insurance type separately and at the person-level, the CPS ASEC overestimated the number of people who were uninsured while inflating the prevalence of dual coverage and having multiple insurance types within a household [5,13].

The current paper makes two important empirical contributions. First, we find that limitations of the imputation procedure continue to bias estimates of health insurance in the 2017 CPS ASEC production file; people with fully imputed data were more likely to be uninsured even after adjusting for relevant characteristics. Fully imputed households were also much more likely to have heterogeneity in insurance coverage. Second, we find the bias in the health insurance imputation procedure appears largely addressed in the 2017 CPS ASEC research file. In the research file, differences between fully imputed and reported cases were greatly

attenuated. Moreover, after accounting for factors predictive of insurance status, there were no significant differences between cases that had all information imputed versus those that had some data reported. Not only does the revised imputation improve distributions at the person level, but imputation of health insurance units improves the distribution of coverage for entire households.

Given that 22% of cases in the CPS ASEC were fully imputed, the improvements to the imputation procedure appreciably affect health insurance estimates in the 2017 CPS ASEC research file. As the Census Bureau plans to incorporate these changes to the health insurance imputation into future data releases, the improvements to the imputation procedure will markedly improve data quality in coming years of the CPS ASEC. Future work should continue to consider the quality of imputation and its implications. One area not explored in this paper is the predictive validity of the new imputation procedures. While this paper has demonstrated that new procedures for imputing data reduce bias, it has not considered whether the new imputation improves the accuracy of the imputed survey data. Additional research is needed to quantify how often imputed outcomes match what outcomes respondents would have reported had their data been non-missing and to compare imputed answers against administrative data sources as available.

References

1. Davern M, Blewett LA, Bershadsky B, Call KT, Rockwood T. State Variation in SCHIP Allocations: How Much is There, What are Its Sources, and Can It be Reduced. Inquiry. 2003; 40 (2): 184–97.

2. Glied S, Remler DK, Zivin JG. Inside the Sausage Factory: Improving Estimates of the Effects of Health Insurance Expansion Proposals. Milbank Quarterly. 2002; 80 (4): 603–36.

3. Hess J, Moore J, Pascale J, Rothgeb J, Keeley C. The effects of person-level versus household-level questionnaire design on survey estimates and data quality. Public Opinion Quarterly. 2001; 65(4): 574-584.

4. Pascale J. Methodological Issues in Measuring the Uninsured. Proceedings of the Seventh Health Survey Research Methods Conference. 1999; pp. 167–173.

5. Davern M, Rodin H, Blewett LA, Call KT. Are the Current Population Survey uninsurance estimates too high? An examination of the imputation process. Health Services Research. 2007; 42(5): 2038-2055.

6. Kenney G, Lynch V. Monitoring children's health insurance coverage under CHIPRA using federal surveys. Databases for Estimating Health Insurance Coverage for Children: A Workshop Summary. Washington, DC: The National Academies Press; 2010.

7. CPS ASEC Health Insurance Redesign – 2014. U.S. Census Bureau Website. <https://www.census.gov/topics/health/health-insurance/guidance/cpsasec-redesign.html> Updated 2015. Accessed January 23, 2019.

8. Berchick ER, Jackson HM. Health Insurance Coverage in the 2017 CPS ASEC Research File. SEHSD Working Paper. <https://www.census.gov/data/datasets/time-series/demo/income-poverty/data-extracts.html> Updated January 23, 2019. Accessed January 23, 2019.

9. Berchick ER, Jackson HM. Health Insurance Coverage in the 2017 CPS ASEC Research File:Estimates from the 2017 Research File. SEHSD Working Paper. <https://www.census.gov/library/working-papers/2019/demo/SEHSD-WP2019-02.html> Updated January 23, 2019. Accessed January 23, 2019.

10. Current Population Survey. U.S. Census Bureau Website. <https://www.census.gov/programs-surveys/cps/about.html> Accessed January 31, 2019.

11. Design and Methodology. Technical Paper 66. <https://www.census.gov/prod/2006pubs/tp-66.pdf>. Accessed January 31, 2019.

12. Andridge RR, Little RJA. A review of hot deck imputation for survey non-response. International Statistical Review. 2010; 78(1):40-64.

13. Boudreaux M, Turner J. Modifications to the Imputation Routine for Health Insurance in the CPS ASEC: Description and Evaluation. State Health Access Data Assistance Center Working Paper. <https://www.shadac.org/sites/default/files/publications/CPS_Imputation_Dec2011.pdf>. Accessed January 31, 2019.

14. 2017 Research File Documentation. U.S. Census Bureau Website. < https://www2.census.gov/programs-surveys/demo/datasets/income-poverty/time-series/data-extracts/2017/cps-asec-research-file/2017-asec-research-file-documentation.pdf>. Accessed January 31, 2019.

15. Using SHADAC Health Insurance Unit (HIU) and Federal Poverty Guideline (FPG) Microdata Variables. Technical Brief. , https://www.shadac.org/publications/using-shadac-

health-insurance-unit-hiu-and-federal-poverty-guideline-fpg-microdata>. Accessed January 31, 2019.

Table 1: Person Health Insurance by Imputation Status in the 2017 CPS ASEC Production and Research Files.

	Production File						Research File					
	Fully Imputed		Not Fully Imputed		Fully Imputed-Not Fully Imputed		Fully Imputed		Not Fully Imputed		Fully Imputed-Not Fully Imputed	
	%	SE[1]	%	SE[1]	%	Sig.	%	SE[1]	%	SE[1]	%	Sig.
Any Coverage	86.82	0.29	92.58	0.11	-5.76	***	91.61	0.24	92.29	0.11	-0.68	**
Private Coverage Only	48.65	0.45	55.66	0.23	-7.01	***	55.98	0.41	55.73	0.24	0.25	
Public Coverage Only	17.82	0.32	25.40	0.23	-7.58	***	26.08	0.36	26.24	0.24	-0.16	
Combination of Private and Public Coverage	20.35	0.34	11.52	0.12	8.83	***	9.54	0.21	10.32	0.12	-0.77	***

* p<.05 **p<.01 ***p<.001

[1]SE stands for standard error.

Table 2: Household Health Insurance by Imputation Status in the 2017 CPS ASEC Production and Research Files.

	Production File						Research File					
	All Household Members Fully Imputed		Not All Household Members Fully Imputed		Fully Imputed-Non-Fully Imputed		All Household Members Fully Imputed		Not All Household Members Fully Imputed		Fully Imputed-Non-Fully Imputed	
	%	SE[1]	%	SE[1]			%	SE[1]	%	SE[1]		
1 person household												
Uninsured	12.44	0.60	6.23	0.23	6.21	***	9.62	0.58	6.23	0.23	3.39	***
Member Covered by Public Insurance Only	30.92	1.00	32.35	0.52	-1.43		33.90	1.08	33.07	0.52	0.82	
Member Covered by Private Insurance Only	35.22	1.00	42.28	0.48	-7.06	***	40.44	0.95	42.07	0.47	-1.63	
Member Covered by Private and Public Insurance	21.41	0.82	19.14	0.39	2.27	*	16.04	0.73	18.63	0.39	-2.59	**
Other Insurance Coverage Pattern	--	--	--	--	--	--	--	--	--	--	--	--
2 person household												
Fully Uninsured	7.45	0.46	2.78	0.13	4.67	***	3.25	0.29	2.95	0.13	0.29	
All Members Covered by Public Insurance Only	16.06	0.64	17.03	0.37	-0.97		17.60	0.62	17.66	0.37	-0.06	
All Members Covered by Private Insurance Only	31.52	0.87	42.06	0.40	-10.54	***	41.18	0.82	42.17	0.39	-0.98	
All Members Covered by Private and Public Insurance	14.45	0.64	14.08	0.30	0.37		12.56	0.52	13.86	0.28	-1.31	*
Other Insurance Coverage Pattern	30.52	0.83	24.05	0.35	6.48	***	25.41	0.69	23.36	0.37	2.05	*
3+ person household												
Fully Uninsured	3.44	0.27	2.24	0.12	1.20	***	2.33	0.22	2.63	0.12	-0.30	
All Members Covered by Public Insurance Only	5.62	0.33	11.26	0.25	-5.64	***	11.11	0.42	12.08	0.27	-0.97	*
All Members Covered by Private Insurance Only	34.27	0.67	48.90	0.36	-14.64	***	47.58	0.75	49.41	0.35	-1.84	**
All Members Covered by Private and Public Insurance	7.55	0.38	1.62	0.10	5.93	***	0.79	0.12	0.96	0.07	-0.18	
Other Insurance Coverage Pattern	49.11	0.67	35.97	0.36	13.14	***	38.20	0.70	34.91	0.34	3.28	***

* p<.05 **p<.01 ***p<.001

[1]SE stands for standard error.

Table 3: Association of Imputation Status with Health Insurance Status in the 2017 CPS ASEC Production and Research File: Results from Multinomial Logistic Regression Models	Production File		Research File	
	Ages 0 to 25	Ages 26 to 64	Ages 0 to 25	Ages 26 to 64
Public Insurance Alone or in Combination (Relative to Uninsured)				
RRR: Fully Imputed	0.78***	0.62***	0.99	0.97
	(0.05)	(0.02)	(0.06)	(0.04)
Private Insurance (Relative to Uninsured)				
RRR: Fully Imputed	0.54***	0.41***	0.89	0.95
	(0.03)	(0.02)	(0.06)	(0.04)
Controls				
Age	X	X	X	X
Race	X	X	X	X
Sex	X	X	X	X
Household Income to Poverty Ratio	X	X	X	X
Self Rated Health	X	X	X	X
Citizenship Status	X	X	X	X
Foreign Born	X	X	X	X
Disability Status		X		X
Marital Status		X		X
Labor Force Participation		X		X

* p<.05 **p<.01 ***p<.001
standard errors in parentheses