Imputation of Non-Response in Height and Weight in the **"Mexican Health and Aging Study"**

Imputación de no-respuesta en peso y talla **en el Estudio Nacional de Salud y Envejecimiento**

Matthew Miller,* Alejandra Michaels-Obregón, Karina Orozco Rocha,*** Rebeca Wong******

The way missing data in population surveys are treated can influence research results. Therefore, the aim of this paper is to explain the reasons and procedure for imputing anthropometric data such as height and weight self-reported by individuals in the first four waves of the Mexican Health & Aging Study (MHAS). We highlight the effect of the imputation versus the exclusion of the cases with missing data, by comparing the distribution of these values and their associated effects on the Body Mass Index using a regression model. We conclude that the incorporation of imputed data offers more solid results as opposed to eliminating the cases with missing data. Hence the importance of applying these statistical procedures, with appropriate treatment of the data, making the methodology and the imputed data available to the users by the same source of information, as offered in the MHAS.

Key words: MHAS; imputation; height; weight; BMI.

El manejo de los datos faltantes en entrevistas por encuestas puede influenciar los resultados de una investigación. Por ello, el objetivo de este trabajo es explicar las razones y el procedimiento de imputación de datos antropométricos, como la altura y el peso, autorreportados en las primeras cuatro rondas del Estudio Nacional sobre Salud y Envejecimiento en México (ENASEM). Destacamos el efecto de la imputación *versus* la eliminación de los casos con datos faltantes, comparando la distribución de dichos valores y sus efectos asociados con el Índice de Masa Corporal mediante un modelo de regresión. Se concluye que la incorporación de datos imputados ofrece resultados más sólidos en comparación con la eliminación de los casos con datos faltantes. De ahí la importancia de aplicar estos procedimientos estadísticos con un manejo adecuado de los datos y difundir la metodología aplicada para obtener los datos imputados desde la misma fuente de información, tal como se ofrece en el ENASEM.

Palabras claves: ENASEM; imputación; altura; peso; IMC.

Recibido: 28 de junio de 2021.
Aceptado: 5 de noviembre de 2021.

* mrmiller@utmb.edu
** almichae@utmb.edu
*** korozco9@ucol.mx
**** rewong@utmb.edu

Body Mass Index Abstract / Hailshadow/ iStock

Introduction

Missing data are a common problem in statistical information collected through population surveys, and an inadequate treatment in the processing and analysis of the information can generate biases and inaccuracies in the results obtained (Abellana & Farran, 2015; Kontopantelis et al., 2017). Missing data in the Mexican Health and Aging Study (MHAS) are no exception, since they are present in a variety of variables including social, economic, and health dimensions. The source of missing data tends to be that the respondent has no knowledge or refuses to disclose the information to the interviewer. In the variables on income and assets, the fraction of missing data is around 10% (Wong et al., 2017a), while in anthropometric variables, such as self-reported height and weight, it is close to 20% (Montevarde & Novak, 2008). In MHAS, the advantage in the economic variables is that the study includes bracket questions as follow-up after a non-response, in order to recover some of the missing data. However, the self-report of anthropometric variables such as height and weight do not use this strategy.

Regarding these two types of variables, there has been more documentation on the mechanisms or techniques to impute missing data in economic variables, such as earned-income variables in the National Survey of Occupation and Employment, ENOE (Durán, 2019), household-income variables in the National Survey of Household Income and Expenditure, ENIGH (Vargas & Valdés, 2018) or economic indicators in National Economic Surveys, EEN (Corona, et al. 2019). These data are collected by the Mexican National Institute of Statistics and Geography (Instituto Nacional de Estadística y Geografía, INEGI). We know less about the mechanisms to impute missing data in the anthropometric var-

iables, hence the importance of documenting the procedure performed for the MHAS.

The anthropometric variables of weight and height are used to calculate quite an important indicator for health and aging research: body mass index (BMI), providing an assessment for level of underweight, normal weight, overweight or the obesity of a person. This indicator is critical and used by multiple studies related to a variety of health dimensions of older adults. Palloni et al. (2015) research the effects of overweight and obesity on the incidence of type 2 diabetes and older adult mortality; or research such as Kumar et al. (2015) that analyze longitudinally the effects of BMI on disability and mortality over an 11-year follow up among Mexicans aged 50 years and older who are non-disabled at baseline in 2001. Now we know that obesity is also a risk factor for severe Covid-19 infection (Satter et al., 2020; González et al., 2021). Indeed, it is estimated that the prevalence of obesity has been rising over the last decade, with 45% of adults 50 years of age and older being overweight and 23% obese in Mexico in 2015 (Rodriguez & Wong, 2019).

This paper aims to provide the rationale and explain the procedure of imputation of the missing data in height and weight self-reported by the individuals in the MHAS. To highlight the effect of imputation versus deletion of observations with missing data, we compared the distributions of these variables among three groups: cases where the data were observed (non-imputed cases), cases where the data were imputed (imputed cases), and all cases (non-imputed plus imputed). Finally, we constructed a database containing the means and standard deviations of height, weight, and BMI of each individual in each wave, along with dummy variables indicating whether height and weight were imputed. These variables are shared with users in an MHAS data file along with the proper documentation.

This work has five sections. First, we present conceptual aspects about missing data and imputation. In the second section, we describe the anthropometric data for weight and height in the Mexican Health and Aging Study for the four waves. Next, we present how we prepared the data for imputation, the procedure for imputation, and the creation of final datasets for end-users. In the fourth section, we present results highlighting the differences between imputed and non-imputed weights and heights, and their effect on the calculated BMI. Finally, we present the conclusions about the importance of imputation in anthropometric data.

1. Conceptual aspects of imputation

There is a variety of ways to handle missing data, such as case deletion or imputation. The selection of the proper mechanism depends on how the missing data are considered: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (Kontopantelis et al., 2017).

There are two types of case deletion. The first excludes from the analysis all the cases with missing data in the variables of interest (listwise)—that is, working only with the cases with complete information for all variables. This implies a reduction of the analytical sample size and, depending on the proportion of missing data, the statistical power of hypothesis tests and standard errors may be affected. This method assumes that missing data are MCAR, meaning that the likelihood that data are missing is totally independent of all observed or missing data. The other alternative is pairwise deletion (or available case analysis), which eliminates those cases with missing data in a specific variable in each analysis. But they are included in other analyses using variables with complete information. This means working with different sample sizes in different parts of the analysis. Like the previous method, this one assumes that missing data are MCAR (Abellana & Farran, 2015).

There are different alternatives for the imputation procedure, such as simple or multiple im-

putation. Imputation seeks to replace missing data with plausible values of each incomplete variable. Plausible values are simulated by estimating relationships between imputed variables and those with no missing values. Imputation adds a layer of uncertainty to results derived from imputed data, as it is not definitively known that the missing values would equal the imputed values if they were observed. Therefore, We recommend creating multiple sets of imputed data using a process that involves a degree of randomness. Such a procedure is called "multiple imputation," and we use it here to compute plausible heights and weights for subjects in each wave of the MHAS for whom such data are missing.

Multiple imputation typically assumes that missing data are MAR, meaning that the likelihood that data are missing is independent of the missing values themselves, given the observed values. Although it is difficult to tell whether our data are MAR from the observed data alone, we believe that assuming as much is reasonable, considering how many variables contributed to our imputations (Rässler, Rubin, & Zell, 2012; van Buuren, Boshuizen, & Knook, 1999).

We identified other associated variables, which can contribute to imputation of height and weight in our data, and we needed to impute any missing values in those other variables too. Height, weight, and other variables that contributed to imputation of height and weight had a non-monotonic pattern of missingness, so we employed the multivariate imputation using chained equations (MICE) or a fully conditional specification (FCS) algorithm because it provides the flexibility that we need (van Buuren, 2007).

2. Data

The Mexican Health and Aging Study (MHAS) is longitudinal and representative of adults aged 50 and over living in rural and urban areas of Mexico. The study is also known by its name in Spanish (Estudio Nacional de Salud y Envejecimiento en México, ENASEM**)**. The goal is to study aging with a broad health, economic and sociodemographic perspective. Furthermore, this study is highly comparable to the U. S. Health and Retirement Study (HRS). The baseline sample was surveyed in 2001. It included households with at least one resident aged 50 years or older (born no later than 1951) and his/her spouse or partner, regardless of age (Wong et al, 2017b). The follow-up surveys were successfully fielded in 2003, 2012, 2015, and 2018. In 2012 and 2018, the MHAS cohort was supplemented with representative samples of adults born between 1952 and 1961 and of those born between 1962 and 1967, respectively.

For this research, the first four waves are used. The MHAS questionnaire is made up of various sections such as: demographic, non-resident children, health, health care services, cognition, help and children, employment, housing, pension, income, and assets. Within the health section, various aspects of self-reporting are asked, such as the diagnosis of chronic diseases as well as weight and height. The latter information is captured with the following questions: "How much do you weigh now?", the answers to which are coded in kilos; and "How tall are you without shoes?", the answers to which are coded in meters and centimeters.

3. Methods

a) Preparing data for imputation

The variables that we seek to impute are self-reported height and weight. The first step is to prepare the data so that the values to be imputed in each variable are identified.

In the raw dataset, numeric variables contain values that although they appear as real numbers are intended to denote observations where those variables were unobserved for some known reason (usually "refused to answer" or "don't know"). These are values such as 888, or 999 in a 3-field variable. Stata, the software used to perform all imputations and analyses described in this document, regards such values as observed and valid, so these values

need to be replaced with explicitly missing values. The MHAS codebooks for each wave list such values for each variable (MHAS, 2001–2015). Stata has 27 different missing values: ".," ".*a*," ".*b*," …, and ".*z*." Because only the soft missing value "." can be imputed in STATA, we assign a soft missing value (.) to the values in every variable that will be imputed.

MHAS selected a subsample in each wave to obtain objective anthropometric measurements, including height and weight, which contributed to the imputation of self-reported heights and weights for those observations selected for the subsample in each wave. Some recorded values of self-reported heights and weights differed so greatly from measured values that the accuracy of the recorded self-reported value is suspect. Therefore, for the imputation exercise, self-reported heights and weights that differed from observed measured values in the same survey participant by more than 10% of the measured value were replaced with soft missing values. Table 1 shows the numbers of self-reported heights and weights in each wave that are missing for this reason.

Furthermore, if a height reported in 2003, 2012, or 2015 differed from the height reported by the same respondent in at least one prior wave by more than 10% of the height in the prior wave, the height in the later wave was also assumed to be inaccurate and replaced with a soft missing value ($N_{2003} = 313, N_{2012} = 477, N_{2015} = 696$). This is because heights in the target population of the MHAS should not change significantly over time.

The process of preparing or "cleaning" the data for imputation in this way is outlined in Figure 1, and the proportions of observations in each wave with missing height and missing weight after the data were cleaned are shown in Figure 2.

b) Imputation Procedure

As previously mentioned, height, weight, and other variables that contributed to the imputation thereof were imputed with the MICE technique. MICE involves random draws from posterior predictive distributions. Thus, for the sake of reproducibility, the seed for pseudorandom-number generation was set to 101 each time that the command "*mi impute chained*" was called in Stata. The covariates for imputation of self-reported height and weight included sex, age, locality size, and years of education. MICE requires that any variable X involved in imputation of another variable Y also be imputed if X has missing values. Table 2 shows the numbers of observations in which each of those variables was imputed.

In addition to these covariates, measured heights and weights contributed to imputation of self-reported heights and weights within the subsam-

Table 1

Numbers of Self-Reported heights and Weights that Differed from Measured Values by more than 10%

Wave	2001	2003	2012	2015
Total MHAS sample size	15,186	13,704	15,723	14,779
Anthropometric subsample size	2,944	2,641	2,086	2,054
Cases with different height	43	53	67	70
Cases with different weight	317	263	252	270

Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

Figure 1

Flowchart of Process of Preparing Data for Imputation in Each Wave

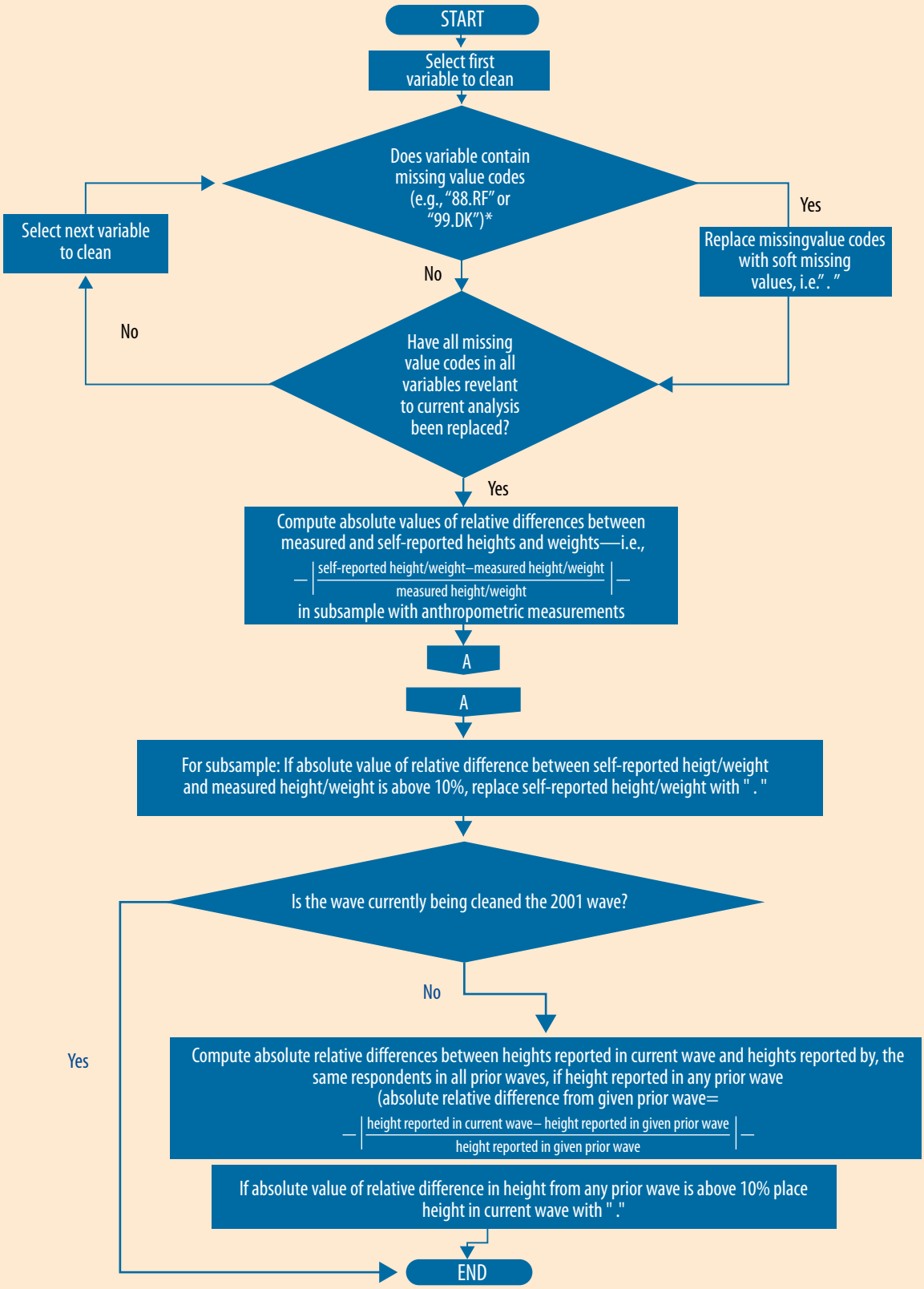
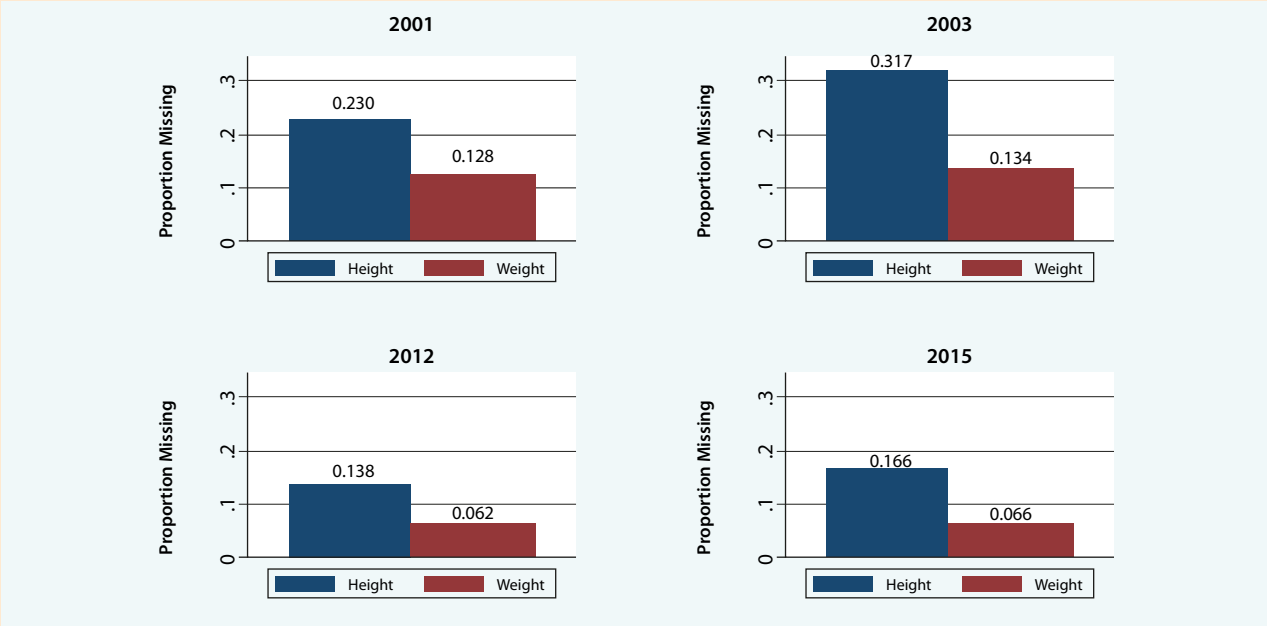


Figure 2

Proportion of Individuals Missing Self-Reported Height and Self-Reported Weight in each Wave



Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

ples selected for anthropometric measurements. We substituted zeroes for measured heights and weights outside of the subsample to allow Stata to perform the MICE algorithm. MICE sequentially performs a univariate imputation on each variable with missing values, in our case predictive mean matching (PMM) for all such variables. More detailed justifications for these choices can be found in the document "Imputation of Height and Weight in the Mexican Health and Aging Study" found on the MHAS webpage.

The following table shows which variables were imputed in each wave and the univariate imputation method used to impute them (Table 3).

All variables with missing values were imputed using the univariate method predictive mean matching (PMM)—called "regression switching" by van Buuren, Boshuizen, and Knook (1999). For each observation with a missing value of the imputed variable, the PMM algorithm finds a predetermined number of observations that are "closest" to the observation with a missing value, according to a certain measure of distance, among all observations with non-missing values of the imputed variable. One of those observations is selected at random, and the observed value from the selected observation is assigned for the missing value. In each wave, each imputed value was selected from one of the five closest observations with non-missing values.

In 2001, missing values of education, self-reported height and weight, and measured height and weight were imputed; sex, age, and locality size had no missing values in 2001. The length of the burn-in period —the number of times PMM was performed before settling on an imputed value— in 2001 was set at 450 iterations.

In 2003, missing values of self-reported height and weight, age, education, and measured height and weight were imputed, and the length of the burn-in period was set at 350 iterations.

In 2012, missing values of age, education, and self-reported height and weight were imputed for the entire sample, and the averages of two meas-

Table 2

Number of missing values among covariates used in the imputations of height and weight

Wave (total sample size)	2001 (15,186)	2003 (13,704)	2012 (15,723)	2015 (14,779)
Sex	0	0	0	0
Age	0	11	25	4
Locality size	0	0	0	0
Education	19	90	68	175

Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

urements each of height and weight were imputed for the subsample selected for anthropometric measurements. The length of the burn-in period was set at 300 iterations for this wave.

In 2015, missing values of age, education, and self-reported height and weight were imputed for the entire sample, and averages of two measurements each of height and weight were imputed for the subsample selected for anthropometric measurements. The length of the burn-in period was set at 300 iterations for this wave.

c) Post-Imputation

After imputation of height and weight, BMI was generated as a "passive variable," a function of one or more imputed variables, in each wave. To examine how imputing missing values can affect results versus entirely excluding observations with missing values from analysis, three linear regression models of the natural logarithm of BMI were estimated in each wave using both imputed and non-imputed data. Each model had one independent variable at a time: diabetic status, years of education, or locality size; and similar

Table 3

Variables imputed and imputation method in each MHAS wave

Variable	Wave			
	2001	2003	2012	2015
Self-reported height	PMM			
Self-reported weight	PMM			
Self-reported body mass index (BMI)	Calculated from self-reported height and weight after imputation			
Measured height (subsample only) [1/]	PMM			
Measured weight (subsample only) [1/]	PMM			
Years of education	PMM			
Age	Complete	PMM		
Locality size[2/]	Complete			
Gender	Complete			

Notes:
Complete indicate that the variable had no missing values and, thus, was not imputed in that wave.
[1/] Average of two measurements each— for waves 2012 and 2015.
[2/] 2003 data from 2001.w
Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

models were constructed using only non-imputed data. The models that included imputed data were pooled across 10 imputations, and the standard errors of estimated coefficients were adjusted to account for the added variability introduced by such pooling.

Finally, for each wave we calculated the means and standard deviations of height, weight, and BMI across 10 imputations for each subject. These are the imputed variables that are provided in the MHAS website (http://www.mhasweb.org/). In cases where such values are observed, the imputed values are the same as the observed values. For each case in each wave, two separate dummy variables are included which indicates if the values for height and for weight were imputed. The goal is to provide as much information as possible to the MHAS data user, who can decide whether or not to use the imputed variables.

4. Results

Tables 4 and 5 show great similarity between the distributions of self-reported height and weight

Table 4 Continue

Self-Reported Heights (cm) and Weights (kg) by Percentile in Non-Imputed Cases and by Imputation Status, 2001–2015

	Self-Reported Heights			Self-Reported Weights		
2001	Non-Imputed	Imputed	All*	Non-Imputed	Imputed*	All*
Percentile	*n*= 11,677	3,509	15,186	13,225	1,961	15,186
1st	135.0	144.7	139.0	40.0	46.7	41.1
5th	146.0	148.5	147.0	50.0	51.9	50.0
10th	150.0	150.1	150.0	53.0	55.0	53.0
25th	155.0	153.2	154.0	60.0	60.5	60.0
50th	160.0	156.4	160.0	70.0	66.3	68.8
75th	168.0	161.7	166.0	78.0	72.1	78.0
90th	173.0	166.1	172.0	87.0	78.0	86.0
95th	177.0	168.2	175.0	94.0	82.4	93.0
99th	185.0	171.2	183.0	109.0	94.1	108.0
2003	Non-Imputed	Imputed*	All*	Non-Imputed	Imputed*	All*
Percentile	*n*= 9,278	4,426	13,704	11,765	1,939	13,704
1st	140.0	145.1	140.0	40.0	51.0	40.0
5th	146.0	148.1	147.0	49.0	55.6	50.0
10th	150.0	150.2	150.0	53.0	57.7	54.0
25th	155.0	153.0	153.7	60.0	62.1	60.0
50th	160.0	156.4	160.0	69.0	67.1	68.2
75th	168.0	161.5	165.0	78.0	73.3	78.0

Table 4 Concludes

Self-Reported Heights (cm) and Weights (kg) by Percentile in Non-Imputed Cases and by Imputation Status, 2001–2015

Self-Reported Heights				Self-Reported Weights		
2003	Non-Imputed	Imputed*	All*	Non-Imputed	Imputed*	All*
Percentile	*n*= 9,278	4,426	13,704	11,765	1,939	13,704
90th	173.0	166.1	170.0	87.0	80.5	86.0
95th	176.0	168.1	175.0	94.0	85.7	93.0
99th	183.0	172.0	182.0	108.0	92.9	106.0
2012	Non-Imputed	Imputed*	All*	Non-Imputed	Imputed*	All*
Percentile	13,622	2,101	15,723	14 ,746	977	15,723
1st	140.0	144.7	140.0	42.0	44.6	42.0
5th	145.0	147.9	145.8	50.0	49.7	50.0
10th	149.0	149.7	149.4	53.0	53.9	53.0
25th	153.0	152.4	153.0	60.0	60.3	60.0
50th	160.0	155.5	160.0	69.0	66.3	69.0
75th	167.0	160.5	165.0	78.0	72.5	78.0
90th	172.0	165.4	171.0	88.0	79.0	87.0
95th	175.0	167.6	175.0	95.0	84.7	95.0
99th	182.0	171.3	180.0	108.0	103.5	108.0
2015	Non-Imputed	Imputed*	All*	Non-Imputed	Imputed*	All*
Percentile	12,386	2,393	14,779	13,807	972	14 ,779
1st	140.0	145.0	140.0	41.0	43.7	41.0
5th	145.0	147.7	146.0	49.0	49.1	49.0
10th	149.0	149.2	149.0	53.0	51.4	53.0
25th	153.0	151.8	152.4	60.0	58.8	60.0
50th	160.0	155.0	159.0	69.0	65.1	68.0
75th	166.0	160.4	165.0	78.0	71.3	78.0
90th	172.0	165.6	170.0	88.0	78.5	87.0
95th	175.0	168.0	175.0	95.0	83.3	95.0
99th	182.0	172.5	180.0	109.0	95.6	109.0

***** Imputed values in each observation averaged across 10 imputations.
Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

among the non-imputed cases and among all cases (combining imputed and non-imputed). For additional analysis, we include box plots of BMI in 2012 that control for locality size and diabetic status (Figures 3 and 4, respectively), showing similar results. The results showing similar distributions between all cases and non-imputed cases are expected, as imputation of missing values should not distort the distribution of the data used to perform imputations.

Histograms of self-reported height and weight in 2012 among imputed cases showed more centralized distributions than histograms among non-imputed cases (see Figure 5). The values for imputed cases were averaged across 10 imputations; this

Table 5

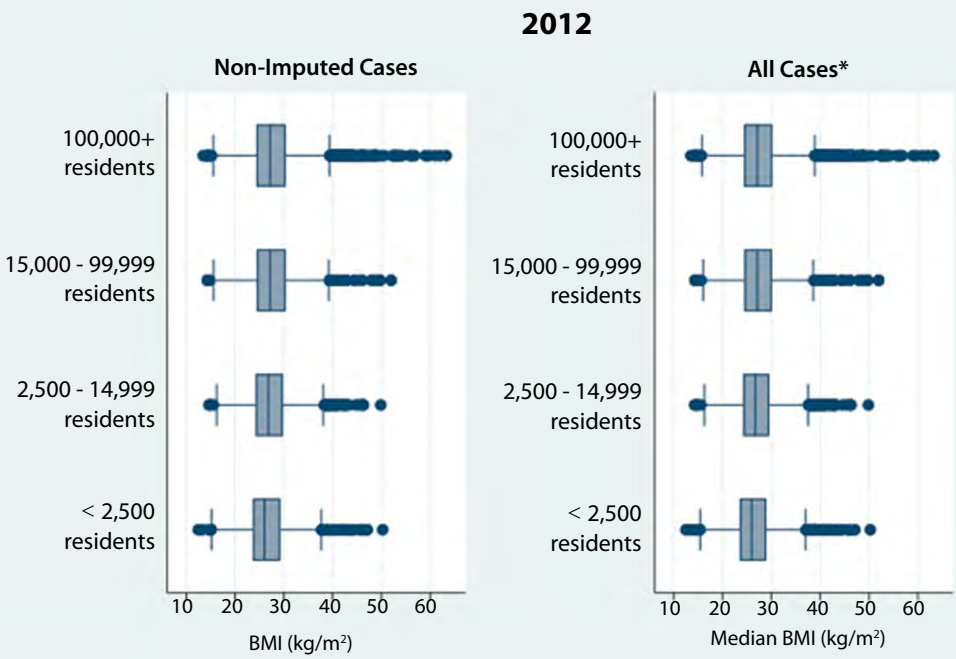
Summary Statistics of Self-Reported Heights and Weights by Imputation Status, 2001–2015

	Number of Missing Values (% of Sample)	Non-Imputed Cases		Imputed Cases*		All Cases*	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
2001 (*N* = 15,186)							
Self-Reported Height (cm)	3,493 (23.0)	160.94	9.99	157.20	6.07	160.08	9.37
Self-Reported Weight (kg)	1,937 (12.8)	69.93	13.98	66.52	9.62	69.49	13.55
2003 (*N* = 13,704)							
Self-Reported Height (cm)	4,347 (31.7)	161.05	9.62	157.33	6.18	159.87	8.85
Self-Reported Weight (kg)	1,833 (13.4)	69.80	14.11	66.49	9.98	69.36	13.68
2012 (*N* = 15,723)							
Self-Reported Height (cm)	2,170 (13.8)	159.99	9.43	156.63	6.06	159.53	9.12
Self-Reported Weight (kg)	977 (6.2)	70.00	13.99	66.81	10.90	69.80	13.84
2015 (*N* = 14,779)							
Self-Reported Height (cm)	2,452 (16.6)	159.85	9.33	156.39	6.34	159.28	9.00
Self-Reported Weight (kg)	971 (6.6)	69.77	14.19	65.72	10.79	69.50	14.03

* Imputed values in each observation averaged across 10 imputations.
Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

Figure 3

Box Plots of BMI by Locality Size Among Non-Imputed Cases and Among All Cases (2012)

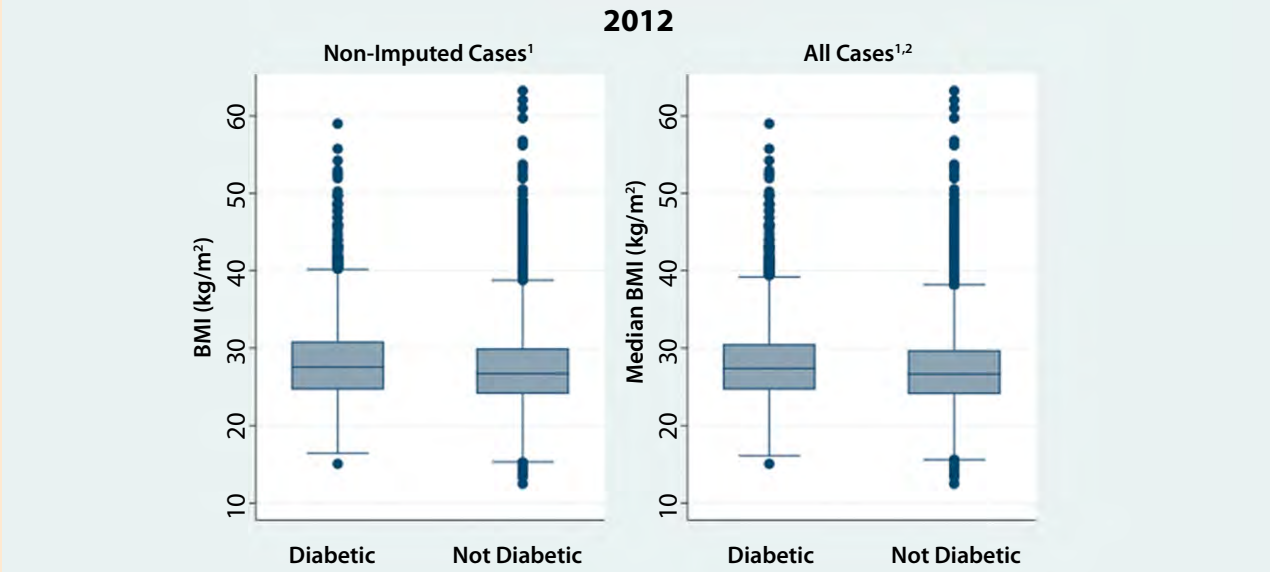


* BMI´s for imputed cases are medians among 10 imputations.

Source: Own calculation using data from the Mexican Health and Aging Study 2012.

Figure 4

Box Plots of BMI by Diabetic Status Among Non-Imputed Cases and Among All Cases (2012)

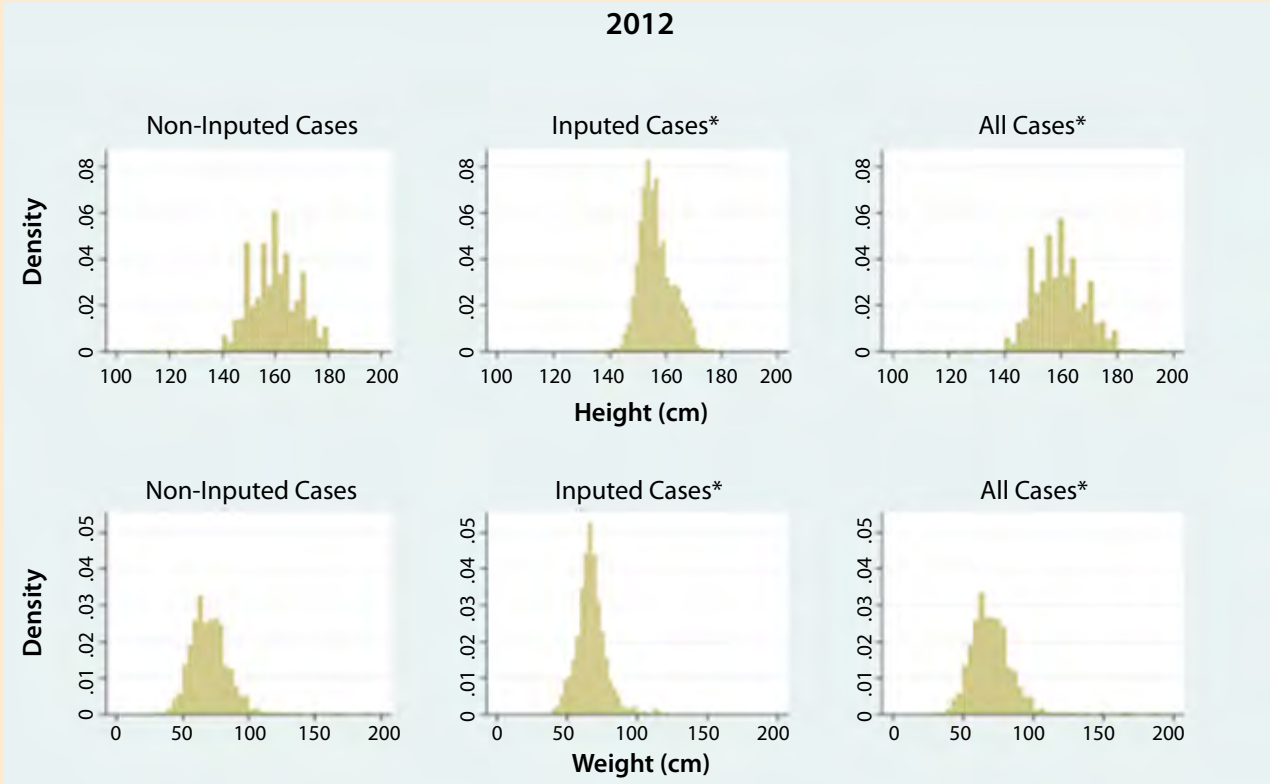


[1] Excludes observations where diabetic status is missing.
[2] BMI´s for imputed cases are medians among 10 imputations.

Source: Own calculation using data from the Mexican Health and Aging Study 2012.

Figure 5

Histograms of Self-Reported Heights and Weights (2012 Wave)



* Imputed values in each observation are averaged across 10 imputations.

could explain the differences between the two distributions because sample means are less variable than the data from which they are computed.

Table 6 presents the regression coefficients of the aforementioned models of log-BMI along with *p*-values, and shows how outright excluding observations with missing data can bias results. For example, in 2001 and 2003, using all (imputed and non-imputed) data showed a statistically significant positive association between log-BMI and education. On the other hand, in the models that excluded observations in which either BMI or education was missing, those associations were estimated to be smaller in magnitude and not statistically significant. Also, in every wave the models with education and locality size as independent variables had smaller coefficients when missing data were excluded than when their imputed values were included. Although the differences varied in magnitude, the fact that such differences were consistently evident across waves implies that the impact of deleting observations with missing data on analysis of these data may be meaningful.

Table 6 Continue

Regression Parameter Estimates of Log-Transformed BMI on Years of Education, Locality Size[1], or Diabetic Status

	Non-Imputed Cases Only			All Cases[2]		
	N[3]	β	*p*-value	*N*[3]	β	*p*-value
2001						
Years of Education	11,107	3.269×10^{-4}	NS	15,186	1.3979×10^{-3}	***
Locality Size (2,500–14,999)		0.026	**	15,186	0.038	***
Locality Size (15,000–99,999)	11,117	0.034	***		0.043	***
Locality Size (100,000+)		0.044	***		0.052	***
Diabetes (Yes)	10 ,830	0.016	***	14,721	0.016	**
2003						
Years of Education	8,875	6.377×10^{-4}	NS	13,704	2.8981×10^{-3}	***
Locality Size in 2001 (2,500–14,999)		0.031	***		0.041	***
Locality Size in 2001 (15,000–99,999)	8,929	0.040	***	13,704	0.052	***
Locality Size in 2001 (100,000+)		0.045	***		0.062	***
Diabetes (Yes)	8,914	0.024	***	13,650	0.025	***
2012						
Years of Education	13, 042	2.07×10^{-3}	***	15,723	2.5986×10^{-3}	***
Locality Size (2,500–14,999)		0.025	***		0.028	***
Locality Size (15,000–99,999)	13,104	0.037	***	15,723	0.039	***
Locality Size (100,000+)		0.037	***		0.041	***
Diabetes (Yes)	13,081	0.034	***	15,689	0.031	***
2015						
Years of Education	11,746	2.4337×10^{-3}	***	14,779	3.5923×10^{-3}	***
Locality Size (2,500–14,999)		0.025	***		0.033	***

Table 6 Concludes

Regression Parameter Estimates of Log-Transformed BMI on Years of Education, Locality Size[1], or Diabetic Status

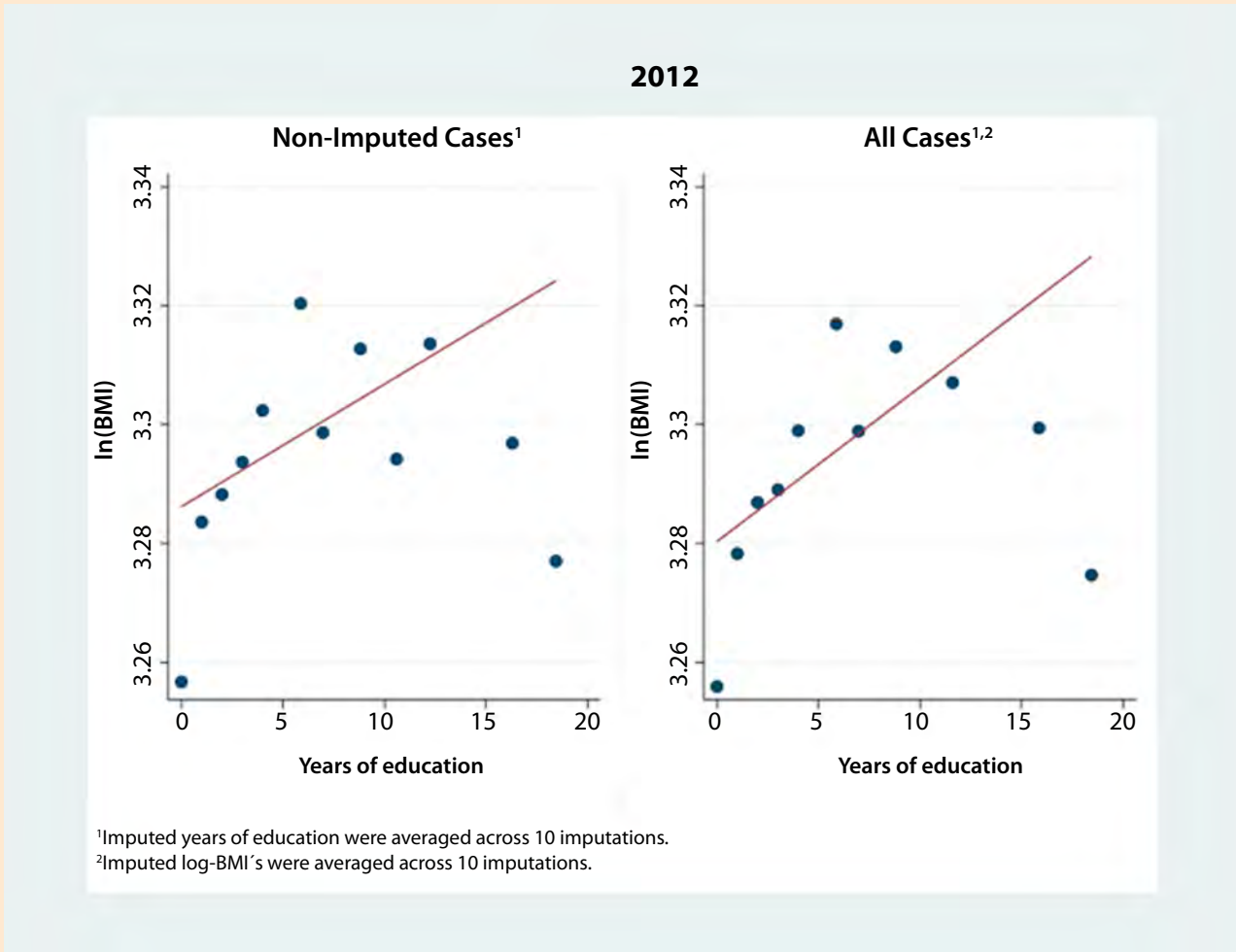
	Non-Imputed Cases Only			All Cases[2]		
	N[3]	*β*	*p*-value	*N*[3]	*β*	*p*-value
Locality Size (15,000–99,999)	11,909	0.023	***	14,779	0.028	***
Locality Size (100,000+)		0.032	***		0.043	***
Diabetes (Yes)	11,898	0.034	***	14,759	0.032	***

Notes: [1] Versuslocalitysize<2,500.
[2] Models pooled across 10 imputations.
[3] Number of observations in which neither log-transformed BMI nor independent variable is missing.
*: $p<0.05$, **: $p<0.01$, ***: $p<0.001$. NS ($p>.05$).
Each model had one independent variable at a time: years of education, locality size, or diabetes. Models were constructed using only non-imputed cases and all cases.

Source: Own calculation using data from the Mexican Health and Aging Study 2001, 2003, 2012 and 2015.

Figure 6

Binned Scatter Plots of Mean Log-BMI by Years of Education among Non-Imputed Cases and All Cases (2012 Wave)



[1]Imputed years of education were averaged across 10 imputations.
[2]Imputed log-BMI´s were averaged across 10 imputations.

Source: Own calculation using data from the Mexican Health and Aging Study 2012.

Comparison between case deletion and multiple imputation with respect to the estimated association between log-BMI and diabetes is more complicated, however. In 2003 multiple imputation showed a stronger association than pairwise deletion showed, as with education and locality size, but in 2012 and 2015 the opposite was true.

Table 6 and Figure 6 show that –although exclusion of cases with missing values biases the slope of the linear association between log-BMI and each of education, locality size, and diabetic status towards zero– this effect is less pronounced in the 2012 and 2015 waves than in the 2001 and 2003 ones. This result may be explained because the last two waves had smaller fractions of missing height and weight than the earlier two waves.

5. Conclusion

We provided a rationale and explained the procedure for imputation of non-response across MHAS waves. Multiple imputation produced more powerful results than case deletion did, without significantly distorting the distributions of height, weight, and body mass index (BMI) computed from these heights and weights. Therefore, we recommend imputing missing data and/or using the imputed values that we have generated here when analyzing data that includes self-reported height and weight from MHAS 2001, 2003, 2012, and/or 2015. More generally, when working with data with missing values, we recommend that users consider multiply imputing missing data whenever possible.

Our results justify the strategy of providing imputed values for the MHAS users, in particular because BMI is a critical variable for many studies of health of mid- and old-age Mexican adults. Our strategy is to provide users with an alternative to excluding the cases with missing values in height or weight, which could bias their results in a meaningful manner. We believe that our imputed variables provide a robust alternative for most users, and that researchers should not need to perform their own imputations. Even though the extent of bias when excluding cases with missing values may vary depending on the specific research and analyses performed, the researchers may at least now be able to test the sensitivity of their results when the cases with missing values are excluded.

As previously stated, the imputations described in this document used data from the 2001, 2003, 2012, and 2015 MHAS waves. Raw data from another wave, fielded in 2018, is now publicly available. Next, we will use the process described above to impute self-reported heights, weights, and BMI's in 2018.

References

Abellana, R., & Farran, A. (2015). "The identification, impact and management of missing values and outlier data in nutritional epidemiology", in: *Nutrición Hospitalaria*. 31(3), 189–195 (DE) https://doi.org/10.3305/nh.2015.31.sup3.8766

Corona, F., López-Pérez, J., & Muriel, N. (2019). "Funcionamiento en muestras finitas de técnicas de imputación y retropolación: caso de las series de encuestas económicas nacionales del INEGI", in: *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía*. 10(3), 100–116.

Durán, B. (2019). "Comparación de metodologías de imputación aplicadas a ingresos laborales de la ENOE", in: *Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía*. 10(3), 4-27.

González-González, C., Orozco-Rocha, K., Samper-Ternent, R., & Wong, R. (2021). "Adultos mayores en riesgo de COVID-19 y sus vulnerabilidades socioeconómicas y familiares: un análisis con el ENASEM", in: *Papeles de Población*. 27(107), 141-165. Epub 06 de diciembre de 2021 (DE) https://doi.org/10.22185/24487147.2021.107.06

Kontopantelis, E., Parisi, R., Springate, D. A., & Reeves, D. (2017). "Longitudinal multiple imputation approaches for body mass index or other variables with very low individual-level variability: the mibmi command in Stata", in: *BMC Research Notes*. 10(1), 1–21 (DE) https://doi.org/10.1186/s13104-016-2365-z

Kumar, A., Karmarkar, A., Tan, A., Graham, J., Arceri, C., Ottenbacher, K., & Al Snih, S. (2015). "The effect of obesity on incidence of disability and mortality in Mexicans aged 50 years and older", in: *Salud Publica Mex*. 57(1), s31–s38.

MHAS Mexican Health and Aging Study, (2001–2015). Data Files and Documentation (public use): Mexican Health and Aging Study, (Data File Codebooks). Retrieved from www.MHASweb.org on September 9, 2020.

Monteverde, M., & Novak, B. (2008). "Obesidad y esperanza de vida en México", in: *Población y Salud Mesoamérica.* 6(1), 1–13 (DE) https://doi.org/10.1038/jid.2014.371

Palloni, A., Beltrán-Sánchez, H., Novak, B., Pinto, G., & Wong, R. (2015). "Adult obesity, disease and longevity in Mexico", in: *Salud Pública de México.* 57(1), s22–s30.

Rässler, S., Rubin, D. B., & Zell, E. R. (2012). "Imputation", in: *WIREs Computational Statistics.* 5(1), 20–29. doi: 10.1002/wics.1240

Rodriguez, M., & Wong, R. (2019). "Envejecimiento en México: Obesidad", in: *Boletín Informativo del ENASEM*. 19(1), 1–2 (DE) http://www.enasem.org/MHAS_AgingInMexico.pdf

Sattar, N., McInnes, I. B., & McMurray, J. J. V. (2020). "Obesity is a risk factor for severe COVID-19 infection: Multiple potential mechanisms", in: *Circulation,* 4–6 (DE) https://doi.org/10.1161/CIRCULATIONAHA.120.047659

Van Buuren, S. (2007). "Multiple imputation of discrete and continuous data by fully conditional specification", in : *Statistical Methods in Medical Research*. 16(3), 219–242. doi: 10.1177/0962280206074463

Van Buuren, S., Boshuizen, H.C., & Knook, D.L. (1999). "Multiple imputation of missing blood pressure covariates in survival analysis", in: *Statistics in Medicine*. 18(6), 681–694. doi: 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R

Vargas Chanes, D., & Valdés Cruz, S. (2018). "Ajuste estadístico a la distribución del ingreso en el Módulo de Condiciones Socioeconómicas 2015 mediante imputaciones múltiples", in: *Realidad, Datos y Espacio, Revista Internacional de Estadística y Geografía.* 9(Número especial), 155–175.

Wong, R., Michaels-Obregon, A., & Palloni, A. (2017b). "Cohort Profile: The Mexican Health and Aging Study (MHAS)", in: *International Journal of Epidemiology.* 46(2), 1–10 (DE) https://doi.org/10.1093/ije/dyu263

Wong, R., Orozco, K., Zhang, D., & Michaels, A. (2017a). "Imputation of non-response on economic variables in the Mexican Health and Aging Study (MHAS / ENASEM) 2015", in: *Aging. University of Texas Medical Branch*.