RESEARCH ARTICLE

Breast Cancer and Modifiable Lifestyle Factors in Argentinean Women: Addressing Missing Data in a Case-Control Study

Julia Becaria Coquet[1], Natalia Tumas[2], Alberto Ruben Osella[3], Matteo Tanzi[3], Isabella Franco[3], Maria Del Pilar Diaz[4]*

Abstract

A number of studies have evidenced the effect of modifiable lifestyle factors such as diet, breastfeeding and nutritional status on breast cancer risk. However, none have addressed the missing data problem in nutritional epidemiologic research in South America. Missing data is a frequent problem in breast cancer studies and epidemiological settings in general. Estimates of effect obtained from these studies may be biased, if no appropriate method for handling missing data is applied. We performed Multiple Imputation for missing values on covariates in a breast cancer case-control study of Córdoba (Argentina) to optimize risk estimates. Data was obtained from a breast cancer case control study from 2008 to 2015 (318 cases, 526 controls). Complete case analysis and multiple imputation using chained equations were the methods applied to estimate the effects of a *Traditional dietary pattern* and other recognized factors associated with breast cancer. Physical activity and socioeconomic status were imputed. Logistic regression models were performed. When complete case analysis was performed only 31% of women were considered. Although a positive association of Traditional dietary pattern and breast cancer was observed from both approaches (complete case analysis OR=1.3, 95%CI=1.0-1.7; multiple imputation OR=1.4, 95%CI=1.2-1.7), effects of other covariates, like BMI and breastfeeding, were only identified when multiple imputation was considered. A Traditional dietary pattern, BMI and breastfeeding are associated with the occurrence of breast cancer in this Argentinean population when multiple imputation is appropriately performed. Multiple Imputation is suggested in Latin America's epidemiologic studies to optimize effect estimates in the future.

Keywords: Body mass index- breastfeeding- cancer epidemiology- dietary pattern- multiple imputation

Asian Pac J Cancer Prev, **17 (10)**, 4567-4575

Introduction

Cancer is the second leading cause of death by disease in Argentina, only preceded by cardiovascular diseases (Dirección de Estadísticas e Información de Salud, 2013). A large body of literature has identified several socio-cultural and biological risk factors associated with most incident cancers in Argentina (Niclis et al., 2015; Pou et al., 2014; Román et al., 2014; Tumas et al., 2014). Most of these works has been conducted by the Group of Environmental Epidemiology of Cancer in Córdoba (GEECC). Diet is a recognized modifiable factor associated with most incident cancers in the country. Argentinean traditional diet is characterized by a high consumption of animal protein and fat (obtained mainly from red meat), and low intakes of fish, fruits and vegetables (Navarro et al., 2003; Pou et al., 2014). In Argentina, breast cancer is the most commonly diagnosed cancer in total population and, specifically, in women. Moreover, a strong diet-breast cancer relationship has been reported (Tumas et al., 2014). Other modifiable factors associated with the disease that have being well studied are breastfeeding and nutritional status (Chan and Norat, 2015; Islami et al., 2015).

In general, health sciences researchers do not inform how much data is missing in their studies or how they handle it appropriately (Klebanoff and Cole, 2008). Even though in the last years recommendations on how to address this problem have been published (Klebanoff and Cole, 2008; Sterne et al., 2009; Von Elm et al., 2014), epidemiologic works reporting numerically this weakness are scarce. This may be partly because health researchers avoid these analyses as they lack confidence in the practice of bias analysis and, in some cases, do not apply

*[1]Instituto de Investigaciones en Ciencias de la Salud (INICSA-UNC-CONICET), Universidad Nacional de Cordoba (UNC), [2]Centro de Investigaciones y Estudio sobre Cultura y Sociedad (CIECS), Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET), Ciudad Universitaria, Cordoba Capital, [4]Biostatistics Unit. School of Nutrition, Faculty of Medical Sciences, University of Cordoba, Avenida Enrique Barros s/n, Ciudad Universitaria, CP 5,000, Cordoba, Argentina [3]Laboratorio di Epidemiologia e Biostatistica, Istituto di Ricerca e Cura a Carattere Scientifico (IRCCS) Saverio de Bellis, Castellana Grotte, Bari, Italia. *For Correspondence: pdiaz@fcm.unc.edu.ar*

appropriate methods to tackle specific problems such as missing data. Moreover, in certain instances, researchers do not realize that missing data can bias results. A review about cohort studies in major medical journals showed that most papers excluded participants with missing data and performed a complete-case analysis (66%) and only 7% applied multiple imputation or Bayesian methods. The rest performed bias methods (Karahalios et al., 2013). Wood et al (Wood, White, and Thompson, 2004) reviewed several randomized controlled trials and stated that only less than a fourth of them included a sensitivity analysis about missing data.

A case-control is a design study frequently used in analytical research in cancer epidemiology. This type of study requires significant planning to avoid bias and the information obtain is important to identify risk factors associated with diseases with long exposure period as cancer. Case-control studies have to deal with missing data, especially missing multiple information in covariates. This lack of information in predictors can lead to biased and/or inefficient estimates of parameter and biased standards errors resulting in incorrect confidence intervals and significance tests. Consequently, effort must be put in obtaining valid and precise risk estimates to translate these into recommendations to the population (Lash et al., 2009). One way of improving validity is to address the missing data, a common source of bias in biomedical research.

In all statistical analysis, some assumptions are made about the missing data mechanism. The validity of results obtained after applying imputation methods depends on the compliance of the assumptions made about the missing data mechanism (White et al., 2011). Little and Rubin's framework is often used to classify the missing data as being missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR) (Acock, 2005; White et al., 2011).

Most statistical software applies by default a complete case analysis to address the problem of incomplete data, excluding from the analysis subjects with one or more variables with missing information. Several relatively simple methods have been developed in last decades, such as substitution by the mean, median or linear regression. These are simple imputation methods where the missing data is imputed with a single value and the complete dataset is utilized to perform the subsequent analysis (Allison, 2009). The uncertainty associated with the imputed value must be considered to obtain valid estimations because the imputed value is not the real value that we would have observed if all variables where complete in the dataset. Thus, to solve this problem multiple imputation method was developed (Rubin, 1976).

In Argentina, most epidemiologic studies probing the cancer causal pathway are case-control studies. However, none of these researches have addressed the missing data problem in the country. The main objective of this study was to optimize risk estimates associated with an identified dietary pattern while using multiple imputation for missing values on covariates in a breast cancer case-control study.

Materials and Methods

Study Design

Data come from an ongoing breast cancer case-control study conducted in female adult population of Córdoba province. Three hundred and eighteen cases under 85years old with a histopathologically confirmed incident primary diagnosis of breast cancer (ICD-10th Edition, ICIE10:C50) have been enrolled between 2008 and 2015 (identified by the Córdoba Tumor Registry). In the same time, 526 controls were randomly chosen. These subjects were healthy women matched by age (± 5 years) and place of residence with cases. All women gave their informed consent and ethics approval was obtained (RePIS 058/10/E).

Data

Data were collected by trained interviewers following a structured questionnaire including auto reported information about sociodemographic and anthropometric characteristics, physical activity, smoking habits, family and personal disease history and dietary habits. Data on diet 5 years (Ambrosini et al., 2008) before interview (for controls) or diagnosis (for cases) was obtained using a food frequency questionnaire and a photographical atlas, both validated (Navarro et al., 2001, 2007).

Imputation and Statistical Analysis

Method

Multivariate imputation using chained equations (MICE) is a practical approach for handle missing data (Acock, 2005; White et al., 2011). In this case, the MAR mechanism of missingnes was assumed. The imputation process has been described elsewhere (White et al., 2011). Briefly, MICE method imputes missing values in different steps; initially, all missing values are filled at random and multiple imputed data sets are generated. For each one of the imputed variables an imputation model is build considering all variables that are included in the subsequent analysis, as well as those that may be predictive of the missing values. Second, the imputed data sets are analyzed separately and, finally, all the independent estimations are combined into an overall estimate.

In this paper, 20 dataset were generated (Sterne et al., 2009) and the imputation method was performed when variables had more than 10% of missing values (Bennett, 2001). The final model selected was the most appropriate model based on a set of imputation models and the average relative variance increase (RVI) obtained (Acock, 2014). In addition, diagnostic plots were performed comparing the distribution of the imputed values with the observed values for the continuous variable imputed (Eddings and Marchenko, 2012).

Imputed Variables

Physical activity (PA) and socioeconomic status (SES) variables were imputed and then used as predictor variables in the final risk logistic regression model. These

variables had a significant amount of missing data mainly because they were included after the study began.

The International Physical Activity Questionnaire (IPAQ, short form) was used to obtain PA information (The International Physical Activity Questionnaire, n.d.) and was included as a continuous variable. The volume of PA was achieved by weighting each type of activity by its energy requirements defined in METs (multiples of resting metabolic rate).

SES variable is build with eight variables from the dataset(Asociación Argentina de Marketing, 2002). When one of these variables is missing, the SES will have missing values too. Therefore, these observed variables were imputed and then the SES was calculated. This strategy prevents the loss of information. Six out of 8 variables were imputed. Education level, number of economic providers in the house, occupation of main provider, having computer at home, having internet at home, having debit card, having health care and having cars, were the variables used to construct SES. All of these variables were imputed except education level and occupation of main provider (<10.0% missing).

In total, seven variables were imputed (PA and six variables used to build SES).

Models

The outcome was the presence/absence of breast cancer. The exposure covariate was *Traditional dietary Pattern*, which was previously identified in Cordoba's population through a principal component factor analysis. This pattern was characterized by positive high loadings of fat meats, bakery products, and vegetable oil and mayonnaise (Tumas et al., 2014). Other recognized risk factors for breast cancer were included: age (Benz, 2008), SES (Bigby and Holmes, 2005), body mass index (BMI) (Chan and Norat, 2015), PA (Amadou, Torres-Mejía, Hainaut, and Romieu, 2014), reproductive variables (having children, breastfeeding, year of menarche, gynecological status) (Sisti et al., 2015; Zhou et al., 2015). A Logistic multiple regression model was used.

Stata 13.0 software (StataCorp LP, USA) was used for analysis.

Results

Table 1 and 2 show the distribution of subjects by variable and distribution of missing data, illustrating shadowed variables with more than 10% of missing. There were 318 breast cancers cases and 526 controls. About half of the women were older than 60 years. More than a third of women presented a higher adherence to the *Traditional dietary Pattern* identified in the female population and more than half were overweight or obese (51.3%). Regarding PA, 21.6% of women were sedentary but it should be noted that the percentage of missing values in this variable was high (29.1%). In relation to education, half of the women did not finished high school education, and 29.5% of them had higher education (university or tertiary education, completed or not) (Table 1).

Relating to gynecologic variables, around 79% of the women had children, 61.7% had breastfed, 69.8% were

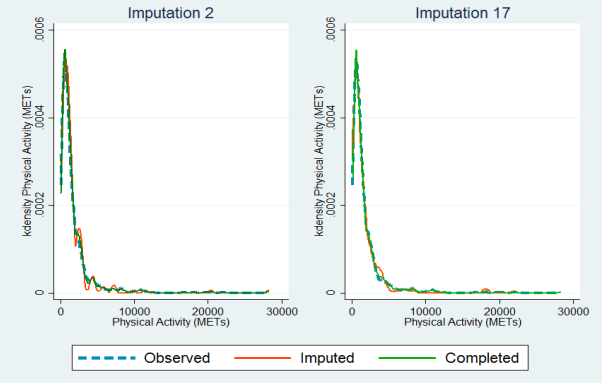

Figure 1. Diagnostic Plots for Physical Activity After Imputation for the Imputed Dataset 2 and 17, Breast Cancer Case-Control Study Córdoba, Argentina 2008-2015.

menopause at the time of the diagnosis (or at interview in controls) and a 17.4% stated were younger than 12 years old at the time of the menarche. Concerning SES variables, around 68% declared having health care and 24.3% were missing values; around a half of the women interviewed had computer, internet and debit card and a 43% of subject did not have a car. All of these SES variables had about 10% of missing values (Table 2, occupation of main provider shown in Supplement and Supporting Data SSD).

Generally, most of the distributions of the missing values differ among categories of the variables included in the analysis, yet not all of these are statistically significant. For example, the distribution of missing data in Health Care variable was different among categories of PA. Sedentary women had 52.7% of missing value in Health Care variable, compare with the 10.7% of missing in the higher category of PA. The differential distribution of missing values was also statistically significant regarding the education, having children and breastfeeding categories. Similar patters occur with categories of the SES variables as well. The percentages of missing values in the SES variables seem to decline as the participants' education level increases. When the *Traditional dietary Pattern* is considered, percentages of missing data elevate as people adhere more to this food pattern. Missing data patterns seems to be MAR mechanism.

Most frequent missing data patterns are shown in SSD, combining all variables included in the analysis, except *Traditional dietary Pattern* (without missing values). Only 31% of women have complete information in all variables. The most frequent pattern of missing data is observed in 21% of subject with missing value only on PA. 20% of women had no information on Health Care only (see Table in SSD).

Table 3 shows the estimated effects (OR, 95 % CI) of covariates from the logistic regression model, when Complete Case (CC) analysis and Multiple Imputation (MI) method are performed. CC is only applied in 31% of subjects. Although significant promoting effect of the Traditional dietary Pattern was observed from both approaches, effects of other covariates, like BMI and breastfeeding, were only identified when MI is

Table 1. Subjects and Missing Data: Absolute and Relative Distributions of Outcome, Exposure and Other Covariates, Breast Cancer Case-Control Study Córdoba, Argentina 2008-2015

Total	n	%	% of missing values						
	844	100							
Breast Cancer			Physical Activity	Health Care	N° of providers	Computer	Internet	Debit Card	Cars
No	526	62.3	29.7	26.0	11.2	11.2	11.2	11.2	11.2
Yes	318	37.7	28.3	21.4	8.8	8.8	8.8	8.8	9.4
Traditional dietary Pattern									
Tertil 1	245	29	31	24.1	6.9*	6.9*	6.9*	6.9*	7.3*
Tertil 2	277	32.8	28.5	24.2	10.5	10.1	10.1	10.1	10.1
Tertil 3	322	38.2	28.4	24.5	12.7	13.0	13.0	13.0	13.3
Age									
<45 years	137	16.2	26.3	21.9	9.5	9.5	9.5	9.5	10.2
45-60 years	307	36.4	24.4	26.4	11.4	11.4	11.4	11.4	11.4
>60 years	400	47.4	33.7	23.5	9.7	9.7	9.7	9.75	10
BMI									
<25kg/mt2	400	47.4	29	27.2	8.5*	8.7*	8.7*	8.75*	8.75*
25-30kg/mt2	268	31.7	27.9	22	9.3	9.3	9.3	9.33	9.33
>30kg/mt2	165	19.6	31.5	22.4	15.8	15.1	15.1	15.1	16.4
Unknown	11	1.3	27.3	0.0	18.2	18.2	18.2	18.2	18.2
Physical Activity									
Sedentary	182	21.6	_	52.7*	4.9*	4.4*	4.4*	4.4*	4.9*
Moderate	228	27	_	28.1	8.8	8.8	8.8	8.8	8.8
Vigorous	188	22.3	_	10.6	23.9	24.5	24.5	24.5	25
Unknown	246	29.1	_	10.2	5.3	5.3	5.3	5.3	5.3
Education									
No Studies	5	0.6	80.0*	0.0*	20.0*	20.0*	20*	20*	20*
Incomplete primary	63	7.5	33.3	0.0	17.6	15.9	15.9	15.9	17.9
Complete primary	281	33.3	22.1	46.6	9.2	9.2	9.2	9.2	9.2
Incomplete high school	79	9.4	34.2	0.0	8.9	10.1	10.1	10.1	10.1
Complete high school	145	17.2	21.4	19.3	10.3	10.3	10.3	10.3	10.3
Higher education	249	29.5	36.9	13.6	6.0	6.0	6.0	6.0	6.4
Unknown	22	2.6	40.9	54.5	54.5	54.5	54.5	54.5	54.5

*Statistically Significant Differences (p <0.05).

considered. Even though uncertainty associated with the imputation process is taking into account in estimations, more precise

95% confidence intervals are observed after MI method is applied. Finally, the imputation diagnostic measure (RVI) of the final risk model shows a value equal to 0.07 indicating that the estimated sampling variability for this set of covariates was just 7% larger than what would have been in the case of complete values of covariates. Figure 1 shows some of the distributions of the imputed and observed values for the physical activity covariate, as an example of the behavior of the imputation modeling. For all imputed dataset the imputed and observed distributions were similar.

Discussion

Estimates obtained applying CC analysis and MI methods differ from each other. *Traditional dietary Pattern*, BMI and Breastfeeding were the variables that showed significant changes in their effects on the occurrence of breast cancer, when the imputation method was considered. Furthermore, the imputation mechanism chosen in the modeling process had a successful performance, based on the value of the diagnostic measure RVI coupled to the combined distributions analysis.

Eating habits of women with high adherence to the *Traditional dietary Pattern* may be linked to breast cancer through different pathways. Fat and carbohydrates intake may influence circulating level of plasma sex hormones and/or growth factors (Amadou et al., 2014; Lajous et al., 2005; Renehan et al., 2015). Fat meat intake may be associated through its high lipid content and the production of heterocyclic amines as well (Ronco et al., 2010) among other mechanisms; and the low presence of dietary fiber and antioxidant vitamins in this dietary pattern may also

Table 2. Subjects and Missing Data: Absolute and Relative Distributions of Gynecologic and Socioeconomic Variables, Breast Cancer Case-Control Study Córdoba, Argentina 2008-2015

Total	n	%	% of Missing Values						
	844	100							
Having Children			Physical Activity	Health Care	N° of providers	Computer	Internet	Debit Card	Cars
No	120	14.2	32.5*	24.2*	10.8	10.8	10.83	10.8	11.7
Yes	666	78.9	29.9	26.3	10.4	10.4	10.36	10.4	10.5
Unknwon	58	6.9	13.8	1.7	8.6	8.6	8.62	8.6	8.6
Breastfeeding									
No	266	31.5	31.2	21.8*	7.5	7.9	7.89	7.9	8.3
Yes	521	61.7	29.4	28.2	11.7	11.5	11.52	11.5	11.7
Unknown	57	6.8	17.5	0.0	10.5	10.5	10.53	10.5	10.5
Menopause									
No	201	23.8	25.4*	24.9	9.5	9.5	9.45	9.4	9.5
Yes	589	69.8	31.9	26.3	11.4	11.4	11.38	11.4	11.5
Unknown	54	6.4	12.9	0.0	1.8	1.8	1.85	1.8	3.7
Menarche									
<12 years	147	17.4	29.9	22.4	11.6	11.6	11.56	11.6	11.6*
>=12 years	676	80.1	29.0	25.0	9.6	9.6	9.62	9.6	9.7
Unknown	21	2.5	28.6	14.3	23.8	23.8	23.81	23.81	28.6
Health Care									
No	68	8.1	33.8*	_	14.7*	14.7*	14.7*	14.7*	14.7*
Yes	571	67.6	34.7	_	11.4	11.4	11.4	11.4	11.4
Unknwown	205	24.3	12.2	_	5.9	5.8	5.8	5.8	5.8
N° of providers									
One	364	43.1	34.1*	23.1*	0.0	0.0*	0.0*	0.0*	0.5*
Two or three	379	44.9	27.9	28.0	0.0	0.3	0.3	0.3	0.3
More than three	14	1.7	21.4	21.4	0.0	0.0	0.0	0.0	0.0
Unknown	87	10.3	14.9	13.8	100.0	98.8	98.8	98.8	98.8
Computer									
No	293	34.7	28.0*	32.8*	0.0*	_	0.0	0.0	0.3*
Yes	464	55.0	32.5	20.9	0.2	_	0.0	0.0	0.2
Unknown	87	10.3	14.9	13.8	98.8	_	100.0	100.0	100.0
Internet									
No	346	41	27.5*	32.7*	0.3*	0.0	_	0.0	0.6*
Yes	411	48.7	33.6	19.5	0.0	0.0	_	0.0	0.0
Unknown	87	10.3	14.9	13.8	98.8	100.0	_	100.0	100.0
Debit Card									
No	354	41.9	28.8*	26.0*	0.3*	0.0	0.0	_	0.6*
Yes	403	47.8	32.5	25.1	00	0.0	0.0	_	0.0
Unknown	87	10.3	14.9	13.8	98.8	100.0	100.0	_	100
Cars									
None	361	42.8	35.2*	19.7*	0.3*	0.0	0.0	0.0	_
One	332	39.3	26.5	30.4	0.0	0.0	0.0	0.0	_
Two	58	6.9	29.3	36.2	0.0	0.0	0.0	0.0	_
Three or more	4	0.5	25.0	0.0	0.0	0.0	0.0	0.0	_
Unknown	89	10.5	14.6	13.5	96.6	97.7	97.7	97.7	_

*statistically significant differences ($p<0.05$).

Table 3. Association Measurements (Odds Ratio), Confidence Intervals and P-Values, Breast Cancer Case-Control Study Córdoba, Argentina 2008-2015

	Complete Case Analysis (n=265)			Multiple Imputation Analysis (n=703)		
	Odds Ratio	95% CI	p value	Odds Ratio	95% CI	p value
Traditional dietary Pattern	1.3	1.0-1.8	0.039	1.4	1.2-1.6	0.00
BMI	1.1	0.9-1.1	0.262	1.1	1.0-1.1	0.03
Breastfeeding	0.6	0.3-1.1	0.087	0.5	0.4-0.8	0.003
SES						
Low-low	0.4	0.1-1.8	0.260	1.1	0.5-2.9	0.766
Upper-low	0.9	0.3-2.9	0.985	1.3	0.6-2.8	0.454
Middle	1.1	0.4-3.3	0.872	1.9	0.9-4.1	0.112
Upper middle	1.5	0.5-4.1	0.443	1.6	0.8-3.2	0.213
Upper	1.3	0.4-3.7	0.673	1.4	0.7-2.9	0.369
Menopause	1.2	0.6-2.6	0.633	1.5	0.9-2.5	0.102
Physical Activity	0.9	0.9-1.0	0.694	1.0	0.9-1.0	0.510
Age	0.9	0.9-1.0	0.961	0.9	0.9-1.0	0.571
Menarche	0.9	0.8-1.1	0.458	1.0	0.9-1.1	0.653
Having Children	1.5	0.6-3.7	0.328	1.6	0.9-2.7	0.102

be related in part to the disease (Karimi et al., 2014).

The association between BMI and breast cancer has long been reported in literature (Chan and Norat, 2015). Three hormonal candidate mechanisms have been proposed for the adiposity–cancer link (related to sex hormone, insulin and insulin-like growth factor 1 (IGF1), and adipokine pathophysiology) (Renehan et al., 2015). In a recent study, the risk ratio of incidence per 5 kg/m^2 increase in BMI showed a significantly stronger trend of association between BMI and breast cancer incidence in Asia–Pacific patients than in European–Australian and North-American patients (Wang et al., 2016). Similarly, the present study showed a 3.0% increased risk of breast cancer per 1kg/m^2 increase in BMI. Tumas et al (Tumas et al., 2014) have already observed an association between BMI and breast cancer in the same population of South America.

Several studies have reported an inverse association between breastfeeding and breast cancer (Anothaisintawee et al., 2013; Collaborative Group on Hormonal Factors in Breast Cancer, 2002). Various mechanisms have been proposed such as decreased frequency and intensity of ovulation, mobilization of endogenous carcinogens from the ductal and lobular epithelial cell environment and facilitating the excretion of organochlorides (Lodha et al., 2011). Recently a meta-analysis showed a strong protective effect of ever breastfeeding against hormone receptor-negative breast cancers (Islami et al., 2015). In our study breast cancer was not clasified by hormone-receptors; however breastfeeding effect became signifcantly associated with breast cancer risk only in MI analysis. While Table 1 shows that more than 60.0% of women have breastfed six months or more, the CC model had not success in identifying this effect. When MI was applied a noticeable improvement of estimates precision was obtained resulting in a significant effect.

Modeling risk factors in epidemiologic studies is always a multidimensional assesment. We utilized models that includes life style variables asociated with breast cancer ocurrence. BMI and breastfeeding were two of these relevant variables reported in literature and they only became significantly asociated with the disease after applying MI in other covariates. This highlights the importance of MI to elucidate effects or associations arising from not so large studies that through conventional methods may not be observed.

The protective role of PA has been documented (Goodwin et al., 2015). Potential anticancer effects of PA include reductions in endogenous sex hormone concentrations, insulin resistance, and chronic low-grade inflammation (Harvie et al., 2015). A recent meta-analysis has identified a significant reduction of breast cancer incidence in European and American patients, and in pre or/and postmenopausal women as well. Furthermore there was a significant non linear dose-effect relationship: the more the PA the lower breast cancer incidence (Liu et al., 2016). In our study PA resulted not associated with breast cancer risk neither in the CC nor MI analysis. Participants were mainly sedentary or presented a moderate activity and was imputed in around 30% of women. In theory, the observed and imputed distributions should not differ from each other, thus the imputation mechanism must have imputed homogeneously in all PA categories. At a population level, almost 60% of argentinean women declare practicing low PA (Instituto Nacional de Estadísticas y Censos, 2013).

Missing values are frequent in epidemiological studies and a problem in statistical analyses. Although using only CC is simpler, estimates obtained may be affected if participants with missing values are omitted. Excluding observations that have missing values also ignores the possibility of systematic differences between complete

cases and incomplete cases, thus the resulting inference might not apply to the entire population, especially when the number of complete cases is small (National Research Council, 2010). The present work analyses reliability of the case-control study on breast cancer conducted in order to identify risk factors of disease. Its sample size is not very large and only a third of subjects are included in CC analysis, thus this issue must be taken into account. When results obtain from CC analysis are compare with those achieve through the MI method, unreliable p values may be obtained in the first case and assessment of the importance of covariates may be inaccurate (Ibrahim et al., 2012). Furthermore, in some cases MI is likely to be advantageous for the coefficient of a relative complete covariate when other covariates are incomplete (White and Carlin, 2010). Misleading results may be obtained regarding the exposure effect. Besides, time and resources invested in collecting information will be wasted, because some will be discarded at the moment of the analysis.

It is noteworthy that we did not use MI to estimate each missing value through simulated values, but rather to represent a random sample of the missing values. This process results in valid statistical inferences when the mechanism chosen is suitable for dataset (Molenberghs and Kenward, 2007). Our work assumed that the information was missing at random (MAR), that is, for a variable X, the probability that an observation is missing depends only on the observed values of other variables, not on the unobserved values of X. Unfortunately, MAR assumption cannot be verified, since missing values are not observed; yet the RVI diagnostic measure, calculated after fitting, indicated good performance of modeling approach.

In Latin America, a few health studies have applied MI to address missing data (Benjet et al., 2008; Camargos at el., 2011; Fries et al., 2013; Nunes et al., 2009; Rubinstein et al., 2010). We did not find any nutritional and cancer epidemiologic study that proposes the MI approach to deal with this information bias in the region. In Argentina, only one study related to cardiovascular diseases (Rubinstein et al., 2010) was found addressing missing data. Even though in the last few years MI has been utilized in the region, to our knowledge none cancer epidemiologic paper applying this method have been published in Latin America. Moreover, none of these studies presented any information about the quality of the imputation models proposed. The small average RVI declared in our study is an estimate of the average relative inflation in variance of the estimates caused by the missing values. Ideally, this estimate should be close to zero (Acock, Alan C., 2014). In our opinion, efforts should be made to strengthen the quality of studies in the region, mainly in Southern Cone territory. Here, epidemiological studies on cancer are not very large, and the possibility missing data may be biasing results should be evaluated.

Some limitations identified were the study size, making imperative to use as much information as possible, and lack of information regarding tumor classification by hormone-receptors.

This study has shown that Traditional dietary Pattern, BMI and breastfeeding are associated with the occurrence of breast cancer in this argentinean population when MI is appropriately performed. This study additionally shows the benefits of performing MI on cancer epidemiology datasets with high proportions of missing data in covariates.

Acknowledgements

Financial Support

We would like to thank the Science and Technology National Agency, FONCyT grant PICT 2012-1019 for financial support and the National Scientific and Technical Research Council (CONICET) for JBC and NT fellowships.

Conflicts of interest none.

Authorship contribution

MPD and JBC designed, drafted and revised critically the article. NT participated in the acquisition of data and making novel contributions in discussion. JBC explored missing data mechanism, analyzed dataset and interpreted results. MPD, ARO, MT and IF revised the article critically for intellectual content. All authors approved the final version of the article.

Ethical Standards Disclosure

This study was conducted according to the guidelines laid down in the Declaration of Helsinki and all procedures involving human subjects were approved by the Ethical Committee of the Faculty of Medical Sciences, University of Córdoba. Written informed consent was obtained from all subjects.

References

Acock AC (2005). Working with missing values. *J Marriage Fam*, **67**, 1012–28.

Acock AC (2014). Working with missing values-multiple imputation. In A Gentle Introduction to Stata (Fourth edition.). College Station, Texas, USA: Stata Press.

Allison PD (2009). 4 Missing Data. In The SAGE Handbook of Quantitative Methods in Psychology (pp. 72–90). 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd.

Amadou A, Torres-Mejía G, Hainaut P, et al. (2014). Breast cancer in Latin America: global burden, patterns, and risk factors. *Salud Publica Méx*, **56**, 547–54.

Ambrosini GL, Fritschi L, de Klerk, NH, et al. (2008). Dietary Patterns Identified Using Factor Analysis and Prostate Cancer Risk: A Case Control Study in Western Australia. *Ann Epidemiol*, **18**, 364-70.

Anothaisintawee T, Wiratkapun C, Lerdsitthichai P, et al. (2013). Risk Factors of Breast Cancer A Systematic Review and Meta-Analysis. *Asia Pac J Public Health*, **25**, 368–87.

Asociación Argentina de Marketing. (2002). Índice de Nivel Socioeconómico 2002.

Benjet C, Borges G, Medina-Mora ME. (2008). DSM-IV personality disorders in Mexico: results from a general population survey. *Rev Bras Psiquiatr*, **30**, 227–34.

Bennett DA. (2001). How can I deal with missing data in my study? *Aust N Z J Public Health*, **25**, 464–69.

Benz CC. (2008). Impact of aging on the biology of breast cancer. *Crit Rev Oncol Hematol*, **66**, 65–74.

Bigby J, Holmes MD. (2005). Disparities across the breast cancer

continuum. *Cancer Causes Control*, **16**, 35–44.
Camargos VP, César CC, Caiaffa WT, et al. (2011). Imputação múltipla e análise de casos completos em modelos de regressão logística: uma avaliação prática do impacto das perdas em covariáveis. *Cad Saude Publica*, **27**, 2299–313.
Chan DSM, Norat T. (2015). Obesity and Breast Cancer: Not Only a Risk Factor of the Disease. *Curr Treat Options Oncol*, **16**, 22.
Collaborative Group on Hormonal Factors in Breast Cancer. (2002). Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. *Lancet*, **360**, 187–95.
Dirección de Estadísticas e Información de Salud. (n.d.). Estadísticas Vitales. Información Básica. Año 2013 (No. 5(57)). Ministerio de Salud. Presidencia de la Nación.
Eddings W, Marchenko Y. (2012). Diagnostics for multiple imputation in Stata. *Stata J*, **12**, 353–67.
Fries L, Grogan-Kaylor A, Bares CB, et al. (2013). Gender differences in predictors of self-reported physical aggression: Exploring theoretically relevant dimensions among adolescents from Santiago, Chile. *Int Perspect Psychol*, **2**, 255–68.
Goodwin P J, Ambrosone CB, Hong CC. (2015). Modifiable Lifestyle Factors and Breast Cancer Outcomes: Current Controversies and Research Recommendations. In P.
A. Ganz (Ed.), Improving Outcomes for Breast Cancer Survivors (Vol. 862, pp. 177–92). Cham: Springer International Publishing.
Harvie M, Howell A, Evans DG. (2015). Can diet and lifestyle prevent breast cancer: what is the evidence? *J Clin Oncol*, **66**, 73-90.
Ibrahim, JG, Chu H, Chen MH. (2012). Missing Data in Clinical Studies: Issues and Methods. *J Clin Oncol*, **30**, 3297–303.
Instituto Nacional de Estadísticas y Censos. Direcciones Provinciales de Estadísticas. Ministerio de Salud de la Nación. (2013). Tercer Encuesta Nacional de Factores de Riesgo para Enfermedades No Transmisibles. Argentina: Ministerio de Salud. Presidencia de la Nación.
Islami F, Liu Y, Jemal A, et al. (2015). Breastfeeding and breast cancer risk by receptor status-a systematic review and meta-analysis. *Ann Oncol*, mdv379.
Karahalios A, Baglietto L, Lee KJ, et al. (2013). The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: a simulation study. *Emerg Themes Epidemiol*, **10**, 75-83.
Karimi Z, Jessri M, Houshiar-Rad A, Mirzaei HR, Rashidkhani B. (2014). Dietary patterns and breast cancer risk among women. P*ublic Health Nutr*, **17**, 1098–106.
Klebanoff MA, Cole SR. (2008). Use of Multiple Imputation in the Epidemiologic Literature. *Am J Epidemiol*, **168**, 355–7.
Lajous M, Willett W, Lazcano-Ponce E, et al. (2005). Glycemic Load, Glycemic Index, and the Risk of Breast Cancer Among Mexican Women. *Cancer Causes Control*, **16**, 1165–9.
Lash TL, Fox MP, Fink AK. (2009). Applying Quantitative Bias Analysis to Epidemiologic Data. New York, NY: Springer New York.
Liu L, Shi Y, Li T, et al. (2016). Leisure time physical activity and cancer risk: evaluation of the WHO's recommendation based on 126 high-quality epidemiological studies. *Br J Sport Med*, **50**, 372–8.
Lodha R, Paul D, Nahar N, et al. (2011). Association between reproductive factors and breast cancer in an urban set up at central India: A case-control study. *Indian J Cancer*, **48**, 303–7.
Molenberghs G, Kenward MG. (2007). Missing Data in Clinical Studies. John Wiley and Sons, Ltd.
National Research Council. (2010). The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials. Retrieved April 14, 2016, from http://www.cytel.com/hs-fs/hub/1670/file-2411099288-pdf/Pdf/MissingDataNationalAcademyof-Medicine.2010.pdf
Navarro A, Cristaldo P, Andreatta MM, et al. (2007). Atlas de alimentos. Córdoba (Argentina): Universidad Nacional de Córdoba, UNC.
Navarro A, Díaz MP, Muñoz SE, Lantieri MJ, Eynard, A. R. (2003). Characterization of meat consumption and risk of colorectal cancer in Cordoba, Argentina. *Nutrition*, **19**, 7–10.
Navarro A, Osella AR, Guerra V, et al. (2001). Reproducibility and validity of a food-frequency questionnaire in assessing dietary intakes and food habits in epidemiological cancer studies in Argentina. *J Exp Clin Cancer Res: CR*, **20**, 365–70.
Niclis C, Román MD, Osella AR, Eynard AR, Díaz MP. (2015). Traditional Dietary Pattern Increases Risk of Prostate Cancer in Argentina: Results of a Multilevel Modeling and Bias Analysis from a Case-Control Study. *J Cancer Epidemiol*, 2015. http://doi.org/10.1155/2015/179562
Nunes LN, Klück MM, Fachel JMG. (2009). Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos Multiple imputations for missing data: a simulation with epidemiological data. *Cad Saude Publica*, **25**, 268–78.
Pou SA, Niclis C, Aballay LR, et al. (2014). Cáncer y su asociación con patrones alimentarios en Córdoba (Argentina). *Nutr Hosp*, **29**, 618–28.
Renehan AG, Zwahlen M, Egger M. (2015). Adiposity and cancer risk: new mechanistic insights from epidemiology. *Nat Rev Cancer*, **15**, 484–98.
Román MD, Niclis C, Tumas N, et al. (2014). Tobacco smoking patterns and differential food effects on prostate and breast cancers among smokers and nonsmokers in Córdoba, Argentina: *Eur J Cancer Prev*, **23**, 310–318.
Ronco AL, De Stefani E, Deneo-Pellegrini H, et al. (2010). Dietary patterns and risk of ductal carcinoma of the breast: a factor analysis in Uruguay. *Asian Pac J Cancer Prev*, **11**, 1187–93.
Rubin DB. (1976). Inference and missing data. *Biometrika*, **63**, 581–92.
Rubinstein A, Colantonio L, Bardach A, et al. (2010). Estimación de la carga de las enfermedades cardiovasculares atribuible a factores de riesgo modificables en Argentina. *Rev Panam Salud Publica*, **27**.
Sisti JS, Bernstein JL, Lynch CF, et al. (2015). Reproductive factors, tumor estrogen receptor status and contralateral breast cancer risk: results from the WECARE study. Springerplus, 4. http://doi.org/10.1186/s40064-015-1642-y
Sterne JAC, White IR, Carlin JB, et al. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*, **338**, b2393–b2393.
The International Physical Activity Questionnaire. (n.d.). International Physical Activity Questionnaire-Short Format. Retrieved April 14, 2016, from https://sites.google.com/site/theipaq
Tumas N, Niclis C, Aballay LR, Osella AR, Díaz MP. (2014). Traditional dietary pattern of South America is linked to breast cancer: an ongoing case–control study in Argentina. *Eur J Nutr*, **53**, 557–66.
Von Elm E, Altman DG, Egger M, et al. (2014). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Int J Surg*, **12**, 1495–9.
Wang J, Yang DL, Chen ZZ, Gou BF. (2016). Associations of body mass index with cancer incidence among populations,

genders, and menopausal status: A systematic review and meta-analysis. *Cancer Epidemiol*, **42**, 1–8.
White IR, Carlin JB. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med*, **29**, 2920–31.
White IR, Royston P, Wood AM. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*, **30**, 377–99.
Wood AM, White IR, Thompson SG. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials*, **1**, 368–76.
Zhou Y, Chen J, Li Q, et al. (2015). Association Between Breastfeeding and Breast Cancer Risk: Evidence from a Meta-analysis. *Breastfeed Med*, **10**, 175–82.