Montlahuc *et al. BMC Nutrition* (2016) 2:71
DOI 10.1186/s40795-016-0114-2

BMC Nutrition

RESEARCH ARTICLE Open Access

Association between dietary polyphenols intake and an oxidative stress biomarker: interest of multiple imputation for handling missing covariates and outcomes

Claire Montlahuc[1,2,3*], Chantal Julia[3,4], Mathilde Touvier[3,4], Léopold Fezeu[3,4], Serge Hercberg[3,4], Pilar Galan[3,4], Emmanuelle Kesse-Guyot[3,4] and Sylvie Chevret[1,2,3]

Abstract

Background: We aimed to illustrate the importance of imputation models specifications, based on a study exploring the associations between subclasses of dietary polyphenols and the thiobarbituric-acid-reactive substances (TBARS).

Methods: Data were collected in a long-term cohort study based on a double-blind randomized placebo-controlled nutritional trial (SU.VI.MAX 2 study). The association between polyphenols intakes and TBARS were studied using linear regression models. Missing data were handled using multiple imputation with chained equations.

Results: A total of 4,129 subjects were included in the analysis, 2,116 of whom had an available outcome measure (TBARS). Differences in selected predictors of TBARS according to the handling of missing data on both covariates and outcome (complete case analysis or multiple imputation) were observed. In the complete case analysis, none of the dietary polyphenol subclasses was found to be associated with TBARS while based on multiple imputed datasets, two polyphenol subclasses, namely catechins and hydroxybenzoic acids, could be selected as associated with TBARS. Of note, while there was a positive association between catechins and TBARS, the hydroxybenzoic acids were negatively associated with TBARS.

Conclusions: Adequate modelling of missing data on both covariates and outcome allowed dietary catechins intake to be selected as associated with a biomarker of oxidative stress.

Keywords: Prospective study, Missing data, Polyphenols intake, Multiple imputation, Complete case analysis, Oxidative stress

Background

When analyzing data collected in epidemiologic population-based studies, many statistical methods have been reported to handle missing data, frequently observed in this setting. These missing data may concern baseline predictors though more frequently follow-up data, including the study outcome, due to study dropouts. Excluding these cases from analyses possibly results in so-called attrition bias [1]. Indeed, such a complete case analysis is only valid if the outcome is missing completely at random (MCAR), so that conclusions about the population of followed subjects also apply to the population of those who dropped out [2]. Otherwise, simple imputation methods such as last observation carried forward or including missing data indicators in regression models, may also lead to overly precise or even biased estimates for relationships of interest when the missing data are missing at random (MAR) or even MCAR. To avoid biases and incorporate

* Correspondence: claire.montlahuc@univ-paris-diderot.fr
[1]Service de Biostatistique et d'information médicale, Hôpital Saint Louis, AP-HP, 1 avenue Claude Vellefaux, 75010 Paris, France
[2]ECSTRA (Epidémiologie Clinique et Statistiques pour la Recherche en Santé), UMR 1153 INSERM, Université Paris Diderot, Sorbonne Paris Cité, France
Full list of author information is available at the end of the article

the imprecision due to imputation rather than observation, a widely applicable approach is multiple imputation [3]. Its interests in the handling of missing values from baseline data [4] or longitudinal data in epidemiologic contexts [5–7] have been shown although multiple imputation is still underused and reported [8–10]. Notably, although it is recommended to include the outcome in addition to the covariates in the imputation model [11, 12], it is not so commonly used in epidemiological studies.

We focused on multiple imputation by chained equations (MICE), that has been advocated and developed as the most appropriate, flexible and practical approach to handle missing data – including missing outcomes – in complex surveys under MAR mechanisms [11], though other techniques have been also proposed [13]. When this MAR assumption can be supported by the data collection, it provides asymptotically unbiased estimates and standard errors, and is asymptotically efficient when correct models are specified for the imputation. Notably, the distribution of each variable with missing values should be properly modeled, and the imputation model should include at least all variables of the analysis model namely all the predictors as well as the outcome [14, 15].

Polyphenols represent a complex family of natural plant-based molecules occurring in most plant foods (such as fruits, cereals, vegetables, chocolate, wine…) consisting of more than 500 identified compounds in the human diet, ranging from low-molecular weight phenolic acids to highly polymerized proanthocyanidins [16, 17], with varying levels of bioavailability and biological properties [16, 18, 19]. Many studies (particularly laboratory experiments using animal models or cultured human cell lines) support an antioxidant role of polyphenols. However, very few of them have ever explored the association between polyphenols and oxidative stress in large samples from the general population.

This study aims to report the interests of handling missing values based on MICE, using as an illustrative example, the study of the long-term relationship between subclasses of dietary polyphenols intake and the thiobarbituric-acid-reactive substances (TBARS), a marker of oxidative stress [20]. This original analysis was performed using data from the SU.VI.MAX (SUpplémentation en VItamines et Minéraux AntioXydants) study, nutritional randomized primary prevention trial which aimed to study the association between antioxidant vitamins supplementation at nutritional doses and health events (1994–2005), and its follow-up cohort study (2008–2009), defining the SU.VI.MAX2 study. As it was an ancillary analysis, there were some incomplete data and we decided to provide interests of multiple imputation techniques in the analysis of such data. Thus, we studied the impact on the results of using complete case analysis or multiple imputation models, as well as the impact of the subsample chosen (including or not subjects without outcome available) to apply the multiple imputation method.

Methods

Motivating example

We used data from the SU.VI.MAX study, a randomized, double-blind, placebo-controlled, primary prevention trial that was initially designed to evaluate the effect of daily supplementation with antioxidant vitamins (E, C, and β-carotene) and minerals (selenium and zinc) at nutritional doses on the incidence of cancer and ischemic heart disease (Trial Registration clinicaltrials.gov Identifier: NCT00272428) [21]. Subjects were included between October 1994 and May 1995 for a planned 8-years supplementation. Volunteer subjects had to fulfil the following eligibility criteria: (1) being 35–60 and 45–60 years old respectively for women and men, (2) declare themselves free of any disease that might compromise participation, (3) not be taking supplements with vitamins or minerals provided for the trial, (4) applying protocol constraints, especially that of receiving a placebo; and (5) express no ambiguous motivations or obsessional behavior concerning diet and health [22]. At inclusion in SU.VI.MAX, all the 12,741 participants fulfilled questionnaires related to socio-demographics, smoking status, physical activity and diet. Dietary data were collected through 24h dietary records, and polyphenol intakes were computed for subjects with at least six 24h records available in the first two years of the follow-up, with a specific distribution: at least three in the summer and three in the winter, to account for seasonal variation in intakes. Polyphenols intakes of each participant were estimated from 24-h dietary records through the Phenol-Explorer database. This database is the first comprehensive electronic database on polyphenols contents in foods and beverages, implemented for the first time in 2009, which provides data on a total of 502 different polyphenols from 452 different foods [23]. All data are available using an open access website which enables to identify the correspondence between food and its polyphenols content (type and amount of each individual polyphenols). Overall, seventeen dietary subclasses of polyphenols intakes (namely anthocyanins, chalcones, dihydrochalcones, dihydroflavonols, proanthocyanidins, theaflavins, catechins, flavanones, flavones, flavonols, isoflavonoïds, hydroxybenzoic acids, hydroxycinnamic acids, other phenolics acids than hydroxybenzoic or hydroxycinnamic, stilbenes, lignans, other polyphenols) were identified [24].

At the end of the supplementation, 4,129 subjects with available dietary intake data and who agreed to participate were included in the optional SU.VI.MAX 2 study [25]. Among them, 2,116 (51.2%) patients, who were selected according to geographical criteria, underwent at 12 years following the SU.VI.MAX randomization visit, a measure of thiobarbituric-acid-reactive substance (TBARS). Indeed, due to operative and logistical obligations (TBARS measurement requires specialized laboratories), this plasmatic oxidative biomarker was only measured for a subsample of about one half of subjects.

Epidemiologists were interested in exploring the associations between subclasses of dietary polyphenols and TBARS plasma concentration. Given one-half of the patients had an available outcome measure, statistical analysis should handle the missingness of the outcome, besides that of the potential confounders from subject self-reporting questionnaires at inclusion.

Statistical analysis

Qualitative parameters were expressed as numbers (percentages) and quantitative parameters as median (interquartile range [IQR]). We compared categorical variables using Fisher's exact tests and Chi-square tests as appropriate and continuous variables using Kruskal-Wallis test.

We used the multiple imputation with chained equations algorithm (MICE) to impute the missing data. The key concept of this sequential method is to use the distribution of the observed data to estimate a set of plausible values for the missing data, incorporating random components to reflect the uncertainty of these estimates. Multiple data sets are created, analyzed individually but identically to obtain a set of parameter estimates, and then combined to obtain the overall estimates, variances and confidence intervals, reported as Rubin's rule [26]. Several imputation models, one for each variable with missing values, are to be defined [11].

All potential confounders – selected by univariate analyses at the 0.2 level (this threshold was chosen to insure that most potentially prognostic variables have been selected so far), or identified from the literature – namely age, intervention group, smoking status at inclusion, number of dietary records, vitamin C and β-carotene were included in the imputation models [11], as well as auxiliary variables (year of inclusion, physical activity at inclusion, educational level, smoking status) [11, 27] and the outcome (TBARS) [11, 12, 15].

We used predictive mean matching algorithm (pmm) for quantitative variables, and logistic regression (logreg) for binary variables, and polytomous logistic regression (polyreg) for categorical variables [28]. All multiple imputations used 50 imputed data sets.

Associations between each subclass of dietary polyphenols and plasma TBARS measure were evaluated using univariate and multivariate linear regression models. The relationship between polyphenol subclasses and TBARS was not linear, therefore polyphenols subclasses were divided into quartiles on the whole population in order to assess the relationship between polyphenols and TBARS when avoiding such an assumption. Potential confounders were selected by univariate analyses at the 0.2 level, or identified from the literature - namely age, sex, intervention group, smoking status at inclusion, number of dietary records, energy intake, and alcohol intake at inclusion. A backward selection procedure was used, though the former seven confounders were forced in the final model.

We assessed and compared the selected predictors when dealing with various samples. First, we considered the subsample of the 2,116 subjects with available TBARS, with either no missing confounders (complete case analysis: $n = 1,112$) or with imputed confounders. Then, we considered the whole sample of the 4,129 enrolled subjects of SU.VI.MAX2, handling the missing values in the outcome itself to reach valid inferences on the whole population.

All statistical tests were two-sided, with *P*-values of 0.05 or less denoting statistical significance. Statistical analysis was performed on R 3.2.0 (http://www.R-project.org/) and multiple imputation analysis used the R package "mice" [29].

Results

Figure 1 displays the study Flow Chart. Among the 4,129 included subjects, 2,013 (48.9%) had a missing outcome, including 734 who also had at least one missing covariate (Table 1). Comparison of patients with or without measure of the outcome is reported in Table 2. Actually, though based on geographic criterion, when comparing more specifically the group with measured TBARS and the one with no measured TBARS, the TBARS measure was related to participant's characteristics at a threshold level of $p = 0.10$ such as gender, age, selenium, serum concentration of α-tocopherol and retinol. This suggests that a not missing completely at random (MCAR) but possibly at random (MAR) underlying mechanism.

Univariable analyses

First, based on complete case analysis ($n = 1,112$), there was no evidence of any association between subclasses of dietary polyphenols and the TBARS (Fig. 2). Then, these associations were studied after multiple imputation excluding individuals with missing outcome or not.

While no significant relationship between total dietary polyphenols and TBARS was found, eight subclasses of polyphenols were selected by MICE univariate analyses as associated with TBARS at a 0.20 level, whatever the sample, namely proanthocyanidins, dihydroflavonols,

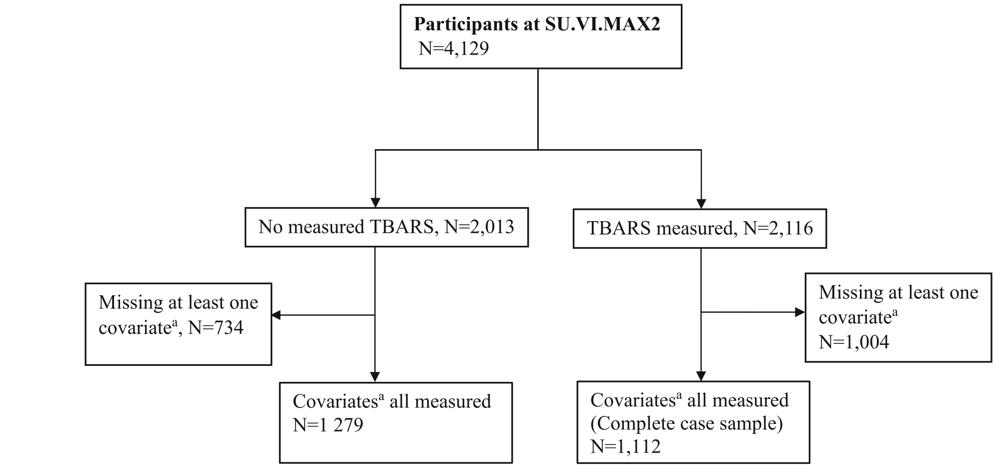
Participants at SU.VI.MAX2
N=4,129
No measured TBARS, N=2,013
TBARS measured, N=2,116
Missing at least one covariate[a], N=734
Missing at least one covariate[a]
N=1,004
Covariates[a] all measured
N=1 279
Covariates[a] all measured
(Complete case sample)
N=1,112

Fig. 1 Flowchart of the study. [a]Covariates considered are those included in the analysis model (but not the "auxiliary" covariates only used in the imputation model)

theflavins, flavones, catechins, flavonols, hydroxybenzoic acids and stilbenes (Fig. 2). Note that standard errors of regression coefficients and thus widths of the 95% confidence intervals were larger for complete case analysis (CCA) than for MICE estimates, that were close when performed in the sample of patients with or without the outcome (Fig. 2). Note also that, besides these polyphenols, nine other variables were selected as having prognostic information, namely weight and height, gender, retinol levels, energy intake, alcohol intakes, α-tocopherol, selenium and zinc. We also incorporated the six other prognostics covariates from literature namely age, intervention group, smoking status at inclusion, number of dietary records, vitamin C, β-carotene. Thus a total of 23 covariates: 8 subclasses of polyphenols exposure variables (namely proanthocyanidins, dihydroflavonols, theaflavins, flavones, catechins, flavonols, hydroxybenzoic acids and stilbenes) and 15 confounding factors, namely weight and height at inclusion, gender, retinol levels, energy intake, alcohol intake, α-tocopherol, selenium and zinc, age, intervention group, smoking status at inclusion, number of dietary records, vitamin C, β-carotene were finally included into the multivariate model.

Multivariable analyses

Table 3 displays the comparison of multivariate models based either on complete cases or after MICE. When excluding patients with missing outcome, only two polyphenol subclasses namely catechins and hydroxybenzoic acids levels, were selected as associated with the outcome, while only the predictive value of catechins was observed after imputing the outcome itself. Of note, while higher the catechins higher the expected TBARS ($p = 0.0033$), the association between hydroxybenzoic acids and TBARS was negative: higher the hydroxybenzoic acids lower the expected TBARS ($p = 0.027$). Finally, as compared to the complete case analysis, two other variables were selected by the multiple imputation procedures only, namely energy intake and β-carotene level while in contrast, the predictive value of age disappeared; these findings were similar whatever we imputed the outcome or not.

Discussion

Missing data are a common burden of epidemiological studies. When faced to such missing data, the assumptions with regards to their underlying mechanisms need to be considered cautiously and correctly to avoid biases and/or inefficiency in estimates of any exposure on the outcome. Accordingly, Little and Rubin's classification of missing data is widely used to segregate their mechanisms, distinguishing three main types of missing data [30]. When the assumption of MCAR mechanism is violated but the MAR can hold, the use of multiple imputation approaches has been shown a valid and simple approach to deal with missing covariates and even outcome [11]. However, it is still seldom used and reported in epidemiological settings. Indeed, in addition to the incompressible time between any statistical innovation and its practical use, multiple imputation techniques require a good understanding and knowledge to be correctly applied. Notably, its implementation is not straightforward. Finally, complete case analysis method (i.e. excluding all subjects with missing value on at least one covariate) is the default method used by most of the statistical softwares, causing it easier to use and largely implemented in epidemiologic studies. We thus attempted to illustrate their use and the importance to be

Table 1 Amount of covariates missing values among subjects with missing outcome and those without missing outcome

	Missing outcome (TBARS)			Available outcome (TBARS)		
	n = 2,013			*n* = 2,116		
Covariates introduced in the imputation models (*n* = 37)[a]	Total of subjects included (covariates all measured)	Total of subjects excluded due to missing covariates	% of missing values	Total of subjects included (covariates all measured)	Total of subjects excluded due to missing covariates	% of missing values
Potential predictors[b]						
Gender	2,013	0	0	2,116	0	0
Age	2,013	0	0	2,116	0	0
Intervention group	2,013	0	0	2,116	0	0
Number of dietary records	2,013	0	0	2,116	0	0
Energy intake	2,013	0	0	2,116	0	0
Weight at inclusion	1,999	14	0.7	2,093	23	1.1
Height at inclusion	1,999	14	0.7	2,089	27	1.3
Zinc	1,952	61	3.1	2,024	92	4.5
Smoking status at inclusion	1,951	62	3.2	2,045	71	3.5
Selenium	1,944	69	3.5	2,024	92	4.5
Alcohol intake at inclusion	1,875	138	7.4	1,976	140	7.1
Alpha-tocophérol	1,729	284	16.4	1,824	292	16.0
Beta carotene	1,729	284	16.4	1,824	292	16.0
Retinol levels	1,729	284	16.4	1,824	292	16.0
Vitamin C	1,651	362	21.9	1,441	675	46.8
Total	1,279	734		1,112	1,004	
Number of missing values						
1		391	19.4		617	29.2
2		47	2.3		71	3.4
3		268	13.3		276	13.0
> 5		28	1.4		40	1.9
Auxiliary covariates						
Years of inclusion	2,013	0	0	2,116	0	0
Physical activity at inclusion	1,975	38	1.9	2,081	35	1.7
Educational level	1,979	34	1.7	2,080	36	1.7
Smoking status at SU.VI.MAX2	2,000	13	0.7	2,095	21	0.01
Total	1,112	901				

[a] Beside the 15 predictive covariates included in the analysis (and thus in the imputation model) and the 4 auxiliary covariates only included in the imputation model, the imputation model also included the 17 subclasses of polyphenols (with no missing value) and the outcome (TBARS)
[b] Covariates included in the analysis model (*n* = 15)

explicit about the multiple imputation procedure when assessing the impact of polyphenols on oxidative stress which is indeed a subject of major concern for epidemiologists. Actually, an adverse effect role of oxidative stress has been previously suggested in brain injury [31], carcinogenesis process [32], or cardiovascular diseases [33], and a protective role of polyphenols has been shown in cancers [34], or dementia [35].

We used data from SU.VI.MAX 2, which is a large cohort study conducted in the general population. The outcome was the TBARS concentration, which has also been recently used as an endpoint in clinical trials of selenium supplementation in patients with Type 2 diabetes [36], of exercise training in hemodialysis patients [37]. The association between catechins and biomarkers of oxidation measured with TBARS has been the focus of

Table 2 Baseline characteristics of subjects according to availability of TBARS measure and the 15 selected predictors

Total (*n* = 4,129)	Measured TBARS (*n* = 2,116)		Non measured TBARS (*n* = 2,013)		*P* value[a]
No. of subjects (%)	Incomplete cases	Complete cases	Incomplete cases	Complete cases	
Median [IQR]	1,004	1,112	734	1,279	
Gender					<0.0001
Male	467 (46.5)	538 (48.4)	283 (38.6)	515 (40.3)	
Female	537 (53.5)	574 (51.6)	451 (61.4)	764 (59.7)	
Age (years)	50 (46 to 55)	49 (46 to 54)	48 (44 to 53)	48 (44 to 53)	<0.0001
Intervention group					0.54
Placebo	471 (46.9)	545 (49.0)	357 (48.6)	593 (46.4)	
Supplementation	533 (53.1)	567 (51.0)	377 (51.4)	686 (53.6)	
Number of dietary records	12 [10–13]	12 [10–13]	12 [10–13]	12 [10–13]	0.24
Energy intake[b] (kcal/d)	1,994 [1,664–2,395]	2,017 [1,660–2,369]	1,965 [1,635–2,326]	1,988 [1,669–2,315]	0.34
Alcohol intake at inclusion (log) (g/d)	2.68 [0.00–3.38]	2.68 [0.00–3.38]	2.68 [0.00–3.09]	2.01 [0.00–3.18]	0.006
Missing	140	0	138	0	
Smoking status at inclusion					0.18
Never smoker	475 (50.9)	549 (49.4)	350 (52.1)	665 (52.0)	
Former smoker	375 (40.2)	424 (38.1)	246 (36.6)	475 (37.1)	
Current smoker	83 (8.9)	139 (12.5)	76 (11.3)	139 (10.9)	
Missing	71	0	62	0	
Weight at inclusion (kg)	66 [57–75]	67 [58–77]	64 [56–74]	65 [57–74]	0.004
Missing	23	0	14	0	
Height at inclusion (cm)	167 [161–173]	168 [161–174]	166 [160–172]	166 [160–172]	<0.0001
Missing	27	0	14	0	
Plasma selenium (μmol/L)	1.08 [0.96–1.20]	1.09 [0.98–1.20]	1.09 [0.98–1.20]	1.11 [1.00–1.23]	0.0003
Missing	92	0	69	0	
Plasma zinc (μmol/L)	13.20 [12.10–14.30]	13.30 [12.00–14.4]	12.80 [11.70–14.00]	13.10 [11.90–14.35]	<0.0001
Missing	92	0	61	0	
Plasma α-tocopherol (μmol/L) (log)	3.46 [3.31–3.60]	3.41 [3.25–3.56]	3.41 [3.27–3.57]	3.41 [3.25–3.55]	<0.0001
Missing	292	0	284	0	
Plasma β-carotene (μmol/L) (log)	−0.64 [−1.05- -0.21]	−0.71 [−1.17- -0.29]	−0.65 [−1.12- -0.21]	−0.66 [−1.11- -0.28]	0.009
Missing	292	0	284	0	
Retinol (μmol/L)	2.31 [1.90-2.74]	2.16 [1.76-2.55]	2.27 [1.90-2.62]	2.12 [1.78- 2.54]	<0.0001
Missing	292	0	284	0	
Vitamin C (μmol/L)	9.63 [7.77 -11.71]	9.64 [7.44-11.63]	10.18 [7.76-12.19]	9.81 [7.33-12.04]	0.13
Missing	675	0	362	0	

Abbreviation: *IQR* interquartile range

[a] pvalues were obtained comparing the 4 groups

[b] Energy intake excluding from alcohol

interest of many studies, particularly because of the abundance of catechins in the human diet. The impact of catechins on TBARS was mostly evaluated on animal studies while studies in humans were based on small samples, with conflicting results [38]. Studies on the relationship between catechins and oxidative stress are inconsistent [39] possibly related to the different amount of catechins ingested by subjects according to the study design. For example, Tinahones et al., in a study of 14 healthy women, showed that the consumption of green tea extract for five weeks was associated with a significant 37.4% reduction in the concentration of oxidized LDL (TBARS) ($p = 0.017$) [40]. Nantz et al. [41] showed, in a study on 111 healthy volunteers that after 3 weeks taking Camellia sinensis compounds twice a day serum malondialdehyde levels was 11.9% lower compared to baseline levels in the intervention group compared to the placebo group. On the contrary, Gomikawa et al.

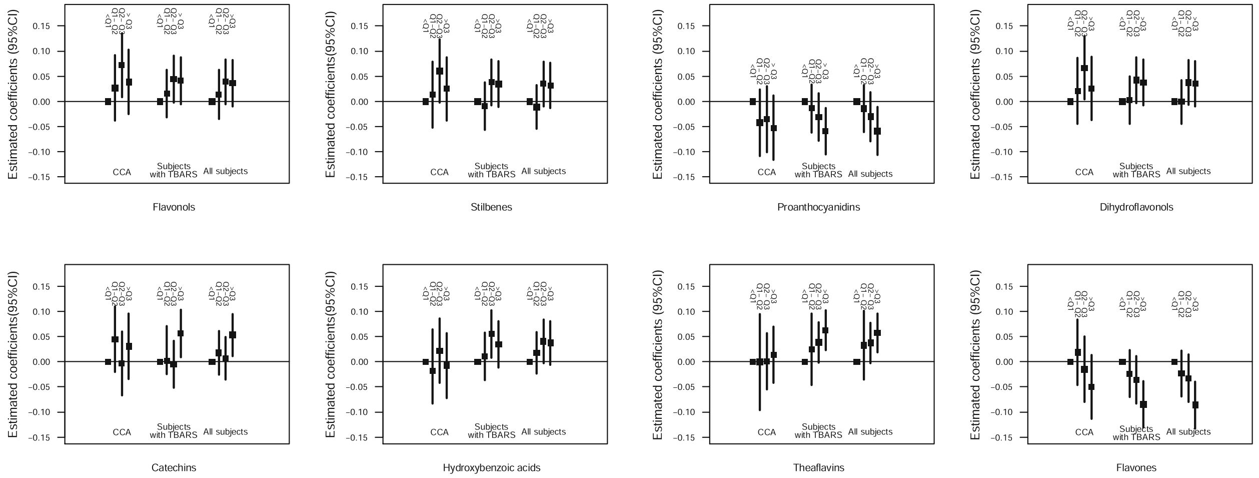

Fig. 2 Comparison of estimated regression coefficient (with 95% CI) of the subclasses of polyphenols (namely flavonols, stilbenes, proanthocyanidins, dihydroflavonols, catechins, hydroxybenzoics acids, theaflavins, flavones) that had univariable prognostic value on TBARS based on at least one sample according to the handling of missing values. Each subclasses was defined as a 4-class variable according to the quartiles of polyphenols subclasses, namely below the First quartile (Q1), between Q1 and the median (Q2), between Q2 and the third quartile (Q3) and above Q3; note that the first class (<Q1) was used as the reference one. Abbreviations: *CCA*: complete case analyses; *subjects with TBARS*: multiple imputation with chained equations of missing covariates ($n = 2116$); *all subjects*: multiple imputation of missing covariates and outcome ($n = 4129$)

showed that TBARS contents in plasma were not changed after green tea consumption [42]. In our study, the relationships between catechins or acids hydroxybenzoic and TBARS were not linear, neither in univariate nor in multivariate analyses whatever the imputation model. Based on multivariate models from imputed datasets, high acids hydroxybenzoic intakes above the third quartile (Q3) were negatively associated with TBARS, while catechins intake higher than Q3 were selected as positively associated with TBARS; thus while the former subclass of polyphenols appears antioxidant, the latter appears to increase the oxidative stress. This result is quite unexpected since polyphenols have usually been reported as antioxidants. However, most of these studies were in vitro experimental or in vivo animal models studies with polyphenols at pharmacological doses, while the current study examines the role of polyphenols at nutritional doses in humans. Moreover, at high doses, polyphenols have also been shown to exert pro-oxidative effects (e.g., increased expression of phosphorylated histone 2AX and metallothionein, markers of DNA damage and response to oxidative stress, respectively). These prooxidative activities may be involved in hepatic and gastrointestinal toxicities observed in animals and humans [43, 44]. Nevertheless, the shape of the relationship between catechins and TBARS, should be stressed out. These results are driving interest in further explorations of the association between polyphenols intake and TBARS in humans.

Of note, in the complete case analysis, none of the polyphenol subclasses was found to be associated with TBARS. Moreover, the strength of association as measured by the estimated regression coefficients were affected by the handling of missing value approach, with for instance coefficients divided by two for the randomized supplementation group and the number of dietary records, while two-fold for others such as energy intake. At last, some participants' characteristics such as sex, age, alcohol intake, zinc or selenium, differed according to the availability of a TBARS measure, suggesting that the probability of data being missing may depend to the observed data (that is, MAR). This mechanism excludes the complete case analysis as the basis of conclusive findings. However, it is not possible to further distinguish between MAR and MNAR (missing not at random, i.e. when the probability of missing data depends on unobserved data) from the observed data alone, although the MAR assumption may appear more plausible in this series due to the large collection of many explanatory variables included in the analysis. Otherwise, when the MNAR appear likely, other statistical approaches should be used [45–47].

According to previous recommendations, the imputation models included all variables planned to be in the analysis model including the outcome [11]. While it has been recommended to include the outcome in the imputation model, this point is somewhat delicate: as underlined by Moons et al. [48] it could seem, on the contrary, a self-fulfilling prophecy to use the outcome to impute data studying the existence of a potential association between the covariates and the outcome. As recommended, the TBARS outcome was systematically

Table 3 Multivariate[a] model of the association of polyphenol intakes with TBARs according to the method of handling missing values

Final models		Complete case analysis			Multiple imputation for covariates			Multiple imputation for covariates and outcome		
		(*n* = 1,112)			(*n* = 2,116)			(*n* = 4,129)		
		Estimation	se	*p*	Estimation	se	*p*	Estimation	se	*p*
Female gender		0.159	0.0282	**<0.0001**	0.134	0.0215	**<0.0001**	0.131	0.022	**<0.0001**
Age (/10 years)		0.044	0.020	**0.027**	0.023	0.0144	0.12	0.023	0.015	0.12
Intervention group		−0.028	0.022	0.20	−0.0146	0.0158	0.35	−0.015	0.017	0.40
Smoking status at inclusion	Never smoked	ref		0.13	ref		0.27	ref		0.18
	Former smoker	0.029	0.0241		0.025	0.0175		0.0261	0.017	
	Smoker	0.069	0.035		0.032	0.027		0.0369	0.026	
Number of dietary records		0.0013	0.0054	0.81	0.0007	0.0039	0.87	0.0007	0.0037	0.83
Energy intake (/1000, kcal/d)		−0.021	0.024	0.38	−0.04	0.017	**0.023**	−0.040	0.018	**0.025**
Alcohol intake at inclusion (log) (g/d)		0.0241	0.0082	**0.0031**	0.026	0.0061	**<0.0001**	0.023	0.0058	**<0.0001**
Plasma α-tocophérol (log) (μmol/L)		0.149	0.0414	**0.00033**	0.172	0.0358	**<0.0001**	0.177	0.0346	**<0.0001**
Plasma β-carotene (log) (μmol/L)					−0.034	0.0153	**0.025**	−0.0341	0.0149	**0.024**
Plasma Selenium (μmol/L)		0.1320	0.0598	**0.027**	0.103	0.045	**0.021**	0.106	0.041	**0.010**
Plasma Zinc (μmol/L)		0.0151	0.0055	**0.0058**	0.014	0.0043	**0.0016**	0.0129	0.0044	**0.004**
Catechins intake (mg/d)	> Q3	-	-	-	0.0798	0.0272	**0.0033**	0.034	0.017	**0.047**
Hydroxybenzoic acids intake (mg/d)	> Q3				−0.059	0.0268	**0.0272**	-		

[a]The multivariable models first included eight subclasses of polyphenols (Dihydroflavonols, Proanthocyanidins, Theflavins, Catechins, Flavones, Flavonols, Hydroxybenzoic, Stilbenes) age, gender, intervention group, smoking status at inclusion, height and weight at inclusion, number of dietary records, energy intake, vitamin C, selenium, β-carotene, zinc, α-tocopherol, retinol, and alcohol intake at inclusion
Bold data indicate statistical significance

included in the imputation models whatever the sample of interest. Whatever the imputation model, the same predictors (except the hydroxybenzoic acids) were selected with close estimated effects, while differences in patient characteristics suggested some selected population from geographic criterion (Table 2). Thus, the close estimations of the two models can also be explained by the fact that these models were adjusted on those discrepancies. Moreover, the hydroxybenzoic acids were not selected when the imputation also applied to the outcome. The high proportion of missing outcome may have influenced these results, with about 50% of the outcomes that had to be imputed. However, Moons et al. [48] showed that imputation of such a high proportion of missing values still provided less biased results compared to complete case analysis. Thus, further studies are necessary to infirm or confirm this observation and the shape of the relationship.

Some limitations of our study could be advocated. First, the measurements of dietary intakes were performed at baseline whereas the outcome was measured about 12 years thereafter. Thus, it could be difficult to highlight the associations between polyphenols and TBARS due to changes in dietary intakes that may have confounded the estimation. Moreover, we used TBARS measure to explore the relationship between polyphenols subclasses and oxidative stress, which was the only oxidative stress biomarker collected in SU.VI.MAX2. Measure of oxidative stress is complicated and it is difficult to declare which oxidative stress biomarker is the best. TBARS represent an interesting marker of oxidative stress since they have been hypothesized to represent a composite number of oxidative damage products [49], however the assay of serum or urine isoprostanes as an oxidative biomarker is now frequently used as a gold standard. Further researches dealing with this association and using another oxidative stress biomarker than TBARS would be of great interest to confirm our results. Besides, the assessment of polyphenols intakes through dietary records is subject to self-reporting bias, despite the fact that repeating 24-h dietary records constitutes an accurate and efficient measurement of polyphenols intakes [50]. However, taking into account missing data with multiple imputations limit the over-selection bias issue analyzing subjects, regardless of the availability of missing covariates. Then, multiple imputation with chained equation procedure has some limitations among which are it lack of a theoretical justification, it MAR assumption sensitiveness increasing with the number of missing data, the possible non convergence insofar as it is an iterative procedure [11]. However, multiple imputation by chained equations, compared to other methods (CCA, median imputation…) remain a less biased methods to handle missing values. At last, the multiplicity issue should be pointed out that some false positive results could have occurred.

Conclusions

In summary, we have provided some results on the association of dietary polyphenols subclasses intake and a validated biomarker of oxidative stress, taking into account missing data on both the covariates and the outcome. The likely missing at random underlying mechanisms allowed using multiple imputation approach, allowing to suggest the only predictive value of catechins among the 17 subclasses of polyphenols on the oxidative stress.

Abbreviations

CCA: Complete case analyses; IQR: Interquartile range; MAR: Missing at random; MCAR: Missing completely at random; MICE: Multiple imputation by chained equations; MNAR: Missing not at random; RCT: Randomized clinical trial; SU.VI.MAX: Supplémentation en Vitamines et Minéraux AntioXydants; TBARS: Thiobarbituric-acid-reactive substances

Acknowledgements

We thank Younes Esseddik, Gwenael Monot, Paul Flanzy, Mohand Ait Oufella, Yasmina Chelghoum, and Than Duong Van (computer scientists), Rachida Mehroug (Logistic assistant) and Nathalie Arnault, Véronique Gourlet, Fabien Szabo, Laurent Bourhis, and Stephen Besseau (statisticians) for their technical contribution to the SU.VI.MAX study. We thank Dr François Laporte for biological analyses.

Funding

This work was supported by the ANR (National Research Agency, grant n° ANR-05-PNRA-010) and DGS (Ministry of Health). The funders had no role in the design, implementation, analysis, or interpretation of data.

Availability of data and materials

Estimation of polyphenols content of foods was based on the Phenol-Explorer database available at http://www.phenol-explorer.eu. Personal data underlying the findings of our study are not publicly available due to legal reasons related to data privacy protection. However, the data are available upon request to all interested researchers after authorization of the French "Commission nationale de l'informatique et des libertés". The person to contact is e.kesse@eren.sbmh.univ-paris13.fr.

Authors' contributions

CM analysed the data, conducted the literature review and wrote the initial draft of the manuscript. CJ was responsible for editing the data. CJ, MT and LF interpreted the data and critical revised the manuscript for important intellectual content. SH, EKG and PG were responsible for developing the design and protocol of the study, interpreted the data and critically revised the manuscript for important intellectual content. SC planned the outline of the present study, conducted the literature review, wrote the initial draft of the manuscript, supervised the statistical analysis and the study and critically revised the manuscript for important intellectual content. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

The SU.VI.MAX and SU.VI.MAX 2 studies were conducted according to the Declaration of Helsinki guidelines and were approved by the Ethics Committee for Studies with Human Subjects of Paris-Cochin Hospital (CCPPRB n° 706 and n° 2364, respectively) and the Comission Nationale de l'Informatique et des Libertés (CNIL n° 334641 and n° 907094, respectively). All individuals gave a written informed consent.

Author details

[1]Service de Biostatistique et d'information médicale, Hôpital Saint Louis, AP-HP, 1 avenue Claude Vellefaux, 75010 Paris, France. [2]ECSTRA (Epidémiologie Clinique et Statistiques pour la Recherche en Santé), UMR 1153 INSERM, Université Paris Diderot, Sorbonne Paris Cité, France. [3]CRESS (Centre de Recherche en Epidémiologie et Statistiques Sorbonne Paris Cité), UMR 1153 INSERM, 75004 Paris, France. [4]EREN (Equipe de Recherche en Epidemiologie Nutritionnelle), Inserm, Université Paris 13, Inra, Cnam, 93017 Bobigny, France.

Received: 26 April 2016 Accepted: 19 November 2016
Published online: 28 November 2016

References

1. Twisk J, de Vente W. Attrition in longitudinal studies. How to deal with missing data. J Clin Epidemiol. 2002;55:329–37.
2. Wood AM, White IR, Hillsdon M, Carpenter J. Comparison of imputation and modelling methods in the analysis of a physical activity trial with missing outcomes. Int J Epidemiol. 2005;34:89–99.
3. Rubin D. Multiple imputation after 18+ years. J Am Stat Assoc. 1996;91:473–89.
4. Arnold AM, Kronmal RA. Multiple imputation of baseline data in the cardiovascular health study. Am J Epidemiol. 2003;157:74–84.
5. Taylor JMG, Cooper KL, Wei JT, Sarma AV, Raghunathan TE, Heeringa SG. Use of multiple imputation to correct for nonresponse bias in a survey of urologic symptoms among African-American men. Am J Epidemiol. 2002; 156:774–82.
6. Barzi F, Woodward M, Marfisi RM, Tognoni G, Marchioli R, GISSI-Prevenzione Investigators. Analysis of the benefits of a Mediterranean diet in the GISSI-Prevenzione study: a case study in imputation of missing values from repeated measurements. Eur J Epidemiol. 2006;21:15–24.
7. Nur U, Longford NT, Cade JE, Greenwood DC. The impact of handling missing data on alcohol consumption estimates in the UK women cohort study. Eur J Epidemiol. 2009;24:589–95.
8. Eekhout I, de Boer RM, Twisk JWR, de Vet HCW, Heymans MW. Missing data: a systematic review of how they are reported and handled. Epidemiol Camb Mass. 2012;23:729–32.
9. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. Am J Epidemiol. 2008;168:355–7.
10. Groenwold RHH, Donders ART, Roes KCB, Harrell FE, Moons KGM. Dealing with missing outcome data in randomized trials and observational studies. Am J Epidemiol. 2012;175:210–7.
11. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med. 2011;30:377–99.
12. van Buuren S. Flexible imputation of Missing Data. Chapman & Hall/CRC editors, Boca Raton, FL. 2012.
13. Burgette LF, Reiter JP. Multiple imputation for missing data via sequential regression trees. Am J Epidemiol. 2010;172:1070–6.
14. Schafer JL. Analysis of incomplete multivariate data. Chapman & Hall, London.1997.

15. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009;338:b2393.
16. Pérez-Jiménez J, Neveu V, Vos F, Scalbert A. Systematic analysis of the content of 502 polyphenols in 452 foods and beverages: an application of the phenol-explorer database. J Agric Food Chem. 2010;58:4959–69.
17. Neveu V, Perez-Jiménez J, Vos F, Crespy V, du Chaffaut L, Mennen L, et al. Phenol-Explorer: an online comprehensive database on polyphenol contents in foods. Database J Biol Databases Curation. 2010;2010:bap024.
18. Manach C, Williamson G, Morand C, Scalbert A, Rémésy C. Bioavailability and bioefficacy of polyphenols in humans. I. Review of 97 bioavailability studies. Am J Clin Nutr. 2005;81:230S–42S.
19. Nielsen ILF, Chee WSS, Poulsen L, Offord-Cavin E, Rasmussen SE, Frederiksen H, et al. Bioavailability is improved by enzymatic modification of the citrus flavonoid hesperidin in humans: a randomized, double-blind, crossover trial. J Nutr. 2006;136:404–8.
20. Williamson G, Manach C. Bioavailability and bioefficacy of polyphenols in humans. II. Review of 93 intervention studies. Am J Clin Nutr. 2005;81:243S–55S.
21. Hercberg S, Galan P, Preziosi P, Bertrais S, Mennen L, Malvy D, et al. The SU.VI.MAX Study: a randomized, placebo-controlled trial of the health effects of antioxidant vitamins and minerals. Arch Intern Med. 2004;164:2335–42.
22. Hercberg S, Preziosi P, Briançon S, Galan P, Triol I, Malvy D, et al. A primary prevention trial using nutritional doses of antioxidant vitamins and minerals in cardiovascular diseases and cancers in a general population: the SU.VI.MAX study–design, methods, and participant characteristics. SUpplementation en VItamines et Minéraux AntioXydants. Control Clin Trials. 1998;19:336–51.
23. Phenol-Explorer database. http://www.phenol-explorer.eu. Accessed 26 June 2016.
24. Manach C, Scalbert A, Morand C, Rémésy C, Jiménez L. Polyphenols: food sources and bioavailability. Am J Clin Nutr. 2004;79:727–47.
25. Kesse-Guyot E, Amieva H, Castetbon K, Henegar A, Ferry M, Jeandel C, et al. Adherence to nutritional recommendations and subsequent cognitive performance: findings from the prospective Supplementation with Antioxidant Vitamins and Minerals 2 (SU.VI.MAX 2) study. Am J Clin Nutr. 2011;93:200–10.
26. Rubin D. Multiple imputation for nonresponse in surveys. New York: Wiley; 1987.
27. Romaniuk H, Patton GC, Carlin JB. Multiple imputation in a longitudinal cohort study: a case study of sensitivity to imputation methods. Am J Epidemiol. 2014;180:920–32.
28. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res. 2007;16:219–42.
29. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45. http://www.jstatsoft.org/v45/i03. Accessed 26 June 2016.
30. Little RJARD. Statistical analysis with missing data. 2nd ed. Hoboken: Wiley; 2002.
31. Lin W-M, Chen M-H, Wang H-C, Lu C-H, Chen P-C, Chen H-L, et al. Association between peripheral oxidative stress and white matter damage in acute traumatic brain injury. BioMed Res Int. 2014;2014:340936.
32. Valko M, Rhodes CJ, Moncol J, Izakovic M, Mazur M. Free radicals, metals and antioxidants in oxidative stress-induced cancer. Chem Biol Interact. 2006;160:1–40.
33. Walter MF, Jacob RF, Jeffers B, Ghadanfar MM, Preston GM, Buch J, et al. Serum levels of thiobarbituric acid reactive substances predict cardiovascular events in patients with stable coronary artery disease: a longitudinal analysis of the PREVENT study. J Am Coll Cardiol. 2004;44:1996–2002.
34. Touvier M, Druesne-Pecollo N, Kesse-Guyot E, Andreeva VA, Fezeu L, Galan P, et al. Dual association between polyphenol intake and breast cancer risk according to alcohol consumption level: a prospective cohort study. Breast Cancer Res Treat. 2013;137:225–36.
35. Commenges D, Scotet V, Renaud S, Jacqmin-Gadda H, Barberger-Gateau P, Dartigues JF. Intake of flavonoids and risk of dementia. Eur J Epidemiol. 2000;16:357–63.
36. Faghihi T, Radfar M, Barmal M, Amini P, Qorbani M, Abdollahi M, et al. A randomized, placebo-controlled trial of selenium supplementation in patients with type 2 diabetes: effects on glucose homeostasis, oxidative stress, and lipid profile. Am J Ther. 2014;21:491–5.
37. Wilund KR, Tomayko EJ, Wu P-T, Ryong Chung H, Vallurupalli S, Lakshminarayanan B, et al. Intradialytic exercise training reduces oxidative stress and epicardial fat: a pilot study. Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc. 2010;25:2695–701.
38. Higdon JV, Frei B. Tea catechins and polyphenols: health effects, metabolism, and antioxidant functions. Crit Rev Food Sci Nutr. 2003;43:89–143.
39. Ellinger S, Müller N, Stehle P, Ulrich-Merzenich G. Consumption of green tea or green tea products: is there an evidence for antioxidant effects from controlled interventional studies? Phytomedicine Int J Phytother Phytopharm. 2011;18:903–15.
40. Tinahones FJ, Rubio MA, Garrido-Sánchez L, Ruiz C, Gordillo E, Cabrerizo L, et al. Green tea reduces LDL oxidability and improves vascular function. J Am Coll Nutr. 2008;27:209–13.
41. Nantz MP, Rowe CA, Bukowski JF, Percival SS. Standardized capsule of Camellia sinensis lowers cardiovascular risk factors in a randomized, double-blind, placebo-controlled study. Nutr Burbank Los Angel Cty Calif. 2009;25:147–54.
42. Gomikawa S, Ishikawa Y, Hayase W, Haratake Y, Hirano N, Matuura H, et al. Effect of ground green tea drinking for 2 weeks on the susceptibility of plasma and LDL to the oxidation ex vivo in healthy volunteers. Kobe J Med Sci. 2008;54:E62–72.
43. Lambert JD, Sang S, Yang CS. Possible controversy over dietary polyphenols: benefits vs risks. Chem Res Toxicol. 2007;20:583–5.
44. Halliwell B. Are polyphenols antioxidants or pro-oxidants? What do we learn from cell culture and in vivo studies? Arch Biochem Biophys. 2008;476:107–12.
45. Heckman J. Sample selection bias as a specification error. Econometrica. 1979;47:153–62.
46. Mealli F, Imbens GW, Ferro S, Biggeri A. Analyzing a randomized trial on breast self-examination with noncompliance and missing outcomes. Biostat Oxf Engl. 2004;5:207–22.
47. Mattei A, Mealli F, Pacini B. Identification of causal effects in the presence of nonignorable missing outcome values. Biometrics. 2014;70:278–88.
48. Moons KGM, Donders RART, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59:1092–101.
49. Trevisan M, Browne R, Ram M, Muti P, Freudenheim J, Carosella AM, et al. Correlates of markers of oxidative status in the general population. Am J Epidemiol. 2001;154:348–56.
50. Ma Y, Olendzki BC, Pagoto SL, Hurley TG, Magner RP, Ockene IS, et al. Number of 24-hour diet recalls needed to estimate energy intake. Ann Epidemiol. 2009;19:553–9.