Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Qingqi Ren Date

Comparing Methods to Handle Missing Data in the Estimation of Population Attributable Factors of Anemia in Preschool Children

By

Qingqi Ren

Master of Public Health

Department of Biostatistics and Bioinformatics

Yi-An Ko, Ph.D. (Thesis Advisor)

Melissa Young, Ph.D. (Thesis Reader)

Comparing Methods to Handle Missing Data in the Estimation of Population Attributable Factors of Anemia in Preschool Children

By

Qingqi Ren

B.A.
Northwest A&F University

University of Nebraska Lincoln

2019

Yi-An Ko, Ph.D.

Melissa Young, Ph.D.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics and Bioinformatics

2021

Abstract

Comparing Methods to Handle Missing Data in the Estimation of Population Attributable Factors of Anemia in Preschool Children

By Qingqi Ren

Introduction: Researchers frequently ignore missing data and include only subjects with complete data in analysis. However, ignoring missing data can lead to systemic bias in the effect estimation and inference as well as loss of power. The challenges of handling missing data include the lack of a general method to handle missing data, unknown missing data mechanisms, and applying appropriate methods corresponding to the missing data mechanism.

Objectives: The objective of the study is to evaluate the impact of missing data on the estimation of the population attributable fraction (PAF) of anemia in children based on three commonly used approaches.

Methods: The prevalence, relative risks, and PAF for proximal risk factors of anemia were estimated in preschool children accounting for complex survey design using national survey data from Nicaragua (NI2005), United States (US2006), and Pakistan (PK2011). Three approaches were used to handle missing data: 1) complete case analysis, 2) inverse probability weighting, and 3) multiple imputation.

Results: In this study, 32.75%, 13.49% and 4.48% were missing SF in NI2005, US2006, and PK2011, respectively. The estimates of PAF were similar across different methods in US2006 and PK2011. The estimated PAF values were substantially smaller using multiple imputation in NI2005 compared to those using complete case and inverse probability weighting. Specifically, the estimated PAF for inflammation, iron deficiency, and vitamin A deficiency were respectively 3%, 29%, and 2-3% using complete case and inverse probability weighting; however, they were 1%, 7%, and 1%, respectively, using multiple imputation. Overall, the estimates using complete case were similar to those using inverse probability weighting,

Conclusions: Different ways of handling missing data can affect the estimate of PAF. Greater impact is observed with a larger proportion of missing data (e.g., >30%). The findings were based on three national surveys and may not be generalized to other PAF estimations. Although the results of inverse probability weighting method and complete case analysis are similar, we recommend to use multiple imputation method in this study.

Comparing Methods to Handle Missing Data in the Estimation of Population Attributable Factors of Anemia in Preschool Children

By

Qingqi Ren

B.A.
Northwest A&F University

University of Nebraska Lincoln

2019

Yi-An Ko, Ph.D.

Melissa Young, Ph.D.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Biostatistics and Bioinformatics

2021

Acknowledgements

Frist, I would like to thank the Biostatistics Department from Rollins School of Public Health at Emory University for the support in all areas of my education. To the faculty and staff at Emory, thank you for the lectures, support, and help.

I would especially say thank you to Advisor Yi-An Ko, Ph.D. who gives me the direction and guidance since I get into Emory. She provided me several opportunities to explore the work in BRINDA project and mentored me throughout this process. Working with her was a privilege. Moreover, her patient and diligence encourage me to accomplish this process.

I am grateful for my Thesis reader, Melissa Young, Ph.D. who is also the chair of BRINDA project. She supported me so much during the working at BRINDA with her patients and enthusiastic during the work. Also, really grateful for the coworkers in BRINDA project. Thank you so much for their help and support.

I would also like to be grateful for my parents and families in China, for their sustainable support. Also, I am really grateful to my friends here in Atlanta. With these relationships, I gained to be a better role of individual and team member of the society.

Table of Contents

1. Background....1
   1.1 Missing data in survey data analysis....1
   1.2 Missing data mechanisms and assumptions....1
   1.3 Current approaches....3
   1.3.1 Complete case analysis....3
   1.3.2 Available case analysis....3
   1.3.3 Missing indicator method....4
   1.3.4 Inverse probability weighting method....4
   1.3.5 Single imputation method....5
   1.3.6 Multiple imputation method....5
   1.4 Anemia in children....6
   1.5 Population attributable fractions....7
2. Introduction....8
   2.1 Reasons and significance of missing data....8
   2.2 Challenges of handling missing data....9
   2.3 Objective....10
3. Method….....11
   3.1 Data selection....11
   3.2 Primary outcomes....11
   3.3 Potential risk factors of anemia and covariates....12
   3.4 Population attributable fractions....13
   3.5 Complete case analysis....13
   3.6 Inverse probability weighting method....14
   3.7 Multiple imputation method....15
   3.8 Statistical analysis....16
4. Results…….....17
   4.1 Population characteristics....17
   4.2 Comparisons of variables with or without missing pattern....17
   4.3 Summary of select anemia risk factors....18
5. Discussion...…….....19
6. Conclusion..…….....22
7. Reference.…….....24
8. Appendix…….....27

1. Background

1.1 Missing data in survey data analysis

Missing data is a common problem in all kinds of survey studies[1]. Researchers often ignore missing values and include only subjects with complete records in the analysis. However, ignoring missing data can lead to many problems, including biased results (overestimates and underestimates of treatment effectiveness) and a loss of power[1].

1.2 Missing data mechanisms and assumptions

There are three types of missing data classifications, which describe the hypothetical mechanisms that lead to missing data. First, missing completely at random (MCAR) is a condition in which the probability of a value missing in a study is the same for all subjects and does not depend on observed or unobserved subject characteristics in the study. In this case, the missing has nothing to do with a particular value missing from the data or an observed value. If the missingness is a random process and no systematic difference between with or without a missing value, this is missing completely at random[2]. If we assume X is the input data, and M is a binary covariate for missing data (M=1 means missingness; M=0 means exist). ξ is the parameter, which cannot be determined by certainty[3]. Under MCAR mechanism, M is completely unrelated with X (1), and the missingness only depends on ξ, which means the probability of missingness is completely random.

$$p(M = 1|\, X, \xi) = p(M = 1|\xi) \quad (1)$$

Second, missing at random (MAR) is when a value is equally likely to be missing in groups of subjects, the groups, which are defined based on observed data. Under this situation, the missingness depends on the observed rather than the unobserved characteristics of the subject in the study, including the missing specific value. In the conclusion, if the missingness is a random process at the level of the observed variable, then it is called Missing at random (MAR)[2]. Under MAR, the probability of missingness depends is equal to the probability of missingness of observed information (X_{obs}) instead of the probability of missingness (X_{miss}) (2) [3].

$$p(M = 1|\, X, \xi) = p(M = 1|X_{obs}, \xi) \quad (2)$$

Lastly, when the missing data are neither MCAR nor MAR, missing data are missing not at random (MNAR). If the missingness is not a random process but depends on the unobserved or unmeasured variable, it is called MNAR[2]. In other words, MNAR means the probability of missingness depends on the exact value of being missing and the characteristics observed by the subjects. Under this mechanism, the missingness depend on both X_{obs} (observed information) and X_{miss} (missing information) (3) [3].

$$p(M = 1|\, X, \xi) = p(M = 1|X_{obs}, X_{miss}, \xi) \quad (3)$$

MCAR and MAR can be distinguished based upon the observed data. However, only using observed data cannot make a distinction between MCAR/MAR and MNAR, since the difference between MCAR/MAR and MNAR depends on unobserved data. Therefore, hypotheses about the mechanism of missing data can be accompanied by data analysis. It

cannot be definitely confirmed by data alone, as data does not tell whether the missing data mechanism is at work or not.

1.3 Current approaches

1.3.1 Complete case analysis

Various ways have been proposed to deal with missing data. Simple but commonly used methods include complete case analysis, available case analysis (pairwise deletion), missing indicator method, inverse probability weighting method, single imputation method, multiple imputation method and overall mean imputation method. The complete case analysis is a standard statistical method that ignores the missing data, which means the subjects with missing values will not contribute to the analysis[2]. Subjects with missing values on any of the variables in a multivariate model will be automatically excluded by complete case (CC) analysis[4]. In addition, the analysis based on complete case may be biased, if the data is missing at random (MAR) rather than completely at random, and the researchers may not be aware of this[4].

1.3.2 Available case analysis

Same as complete case analysis, available case analysis is simple and frequently used in handling missing data. All available data are used in available case analysis to estimate parameters in the model[5]. This method is preferred to use when a study looks for missing

observations in univariate descriptive statistics of data, in order to examine the means and variances of variables[5]. Both complete case analysis and available case analysis can lead to inefficient analysis and highly biased correlation estimations. More complex techniques for handling missing data will provide better results.

1.3.3 Missing indicator method

Another method that is popular for dealing with missing values is called the missing indicator method. A new dummy or indicator (0/1) variable is created, for each independent variable with a missing value, where "1" represents the missingness of the original variable and "0" represents the observed value. In the case of the original variable, the missingness is recorded as "0", and in the case of the original categorical variable, this essentially means that an additional value category is created for the missingness[2]. When the relationship between the independent variable and the outcome is estimated in a multivariate analysis, this indicator is always included with the original variable[2]. The advantage of this indexing approach is, in the multivariate analysis, all subjects are used.

1.3.4 Inverse probability weighting method

In addition, inverse probability weighting (IPW) method is also a commonly used method to reduce bias caused by complete case method and available case method by re-adjusting weights. In this method, the weights of the complete cases are the inverse of the probability.

Additionally, it can also be used to adjust unequal sample scores in sample surveys simultaneously[6].

1.3.5 Single imputation method

A simulation study was conducted based on the diagnostic example. The missing value of these test results is imputed by the overall sample mean of the test results of the observed object, for example, the non-diseased object is calculated together with the diseased object, which significantly increases the amount of overlap. For single imputation method, if the number of study variables is limited, it may be feasible to directly replace subjects from source populations based on observed subject characteristics[2].

1.3.6 Multiple imputation method

In general, the number of covariables is large. If the missing test result is dependent on all known covariates (MAR), the replacement of the subject should be a randomly selected subject from the source population. It is cumbersome to use observations from other subjects or available data to estimate the distribution of test results in the source population. Multiple imputation (MI) method is to obtain correct estimates, the imprecision from the distribution of variables of missing values is estimated. Each analysis generates a correlation with the standard error, resulting in multiple regression coefficients or odds ratios and the corresponding standard error[2]. The estimates can simply be averaged to get a summary

estimate of the correlation, because the correlation of each estimate is unbiased, by assuming the data is MAR. These averages generally reduce the variance of estimates. A single imputation produces an unbiased estimate with a too narrow confidence interval, while multiple imputations method do produce an unbiased estimate with the correct standard error and p-values. With these techniques, the missing data is imputed into a value that is predicted using other known characteristics of the object. These complex technologies are readily available and used in standard software such as SAS, STATA, R, SPSS, PYTHON, MPLUS, MATLAB, and S-Plus.

1.4 Anemia in children

For children's health and development, low hemoglobin concentrations and anemia are important risk factors[7], which adversely affect the development of cognitive and motor, and contribute to fatigue and low productivity[8]. Compared to the rest of the population, the incidence of anemia in preschoolers (PSC) remains the highest, declining by only 4 percentage points from 1995 to 2011. The etiology of anemia is complex and contextual which may vary between mild and severe anemia[9]. The proximal and distal determinants of anemia pathways have many conceptual models that have been developed, including a framework for specific Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia (BRINDA) project[10]. The BRINDA project is a multi-agency international collaboration established in 2012 aim to improve global micronutrient assessment and the

characterization of anemia. BRINDA 's overall goal is to improve methods for assessing nutritional status, thus improving the goals, design and effectiveness of nutrition research and programs[11].

1.5 Population attributable fractions

To examine risk factors that contribute to anemia at the population level, the population attributable fraction (PAF) framework is often used to estimate the relative contributions of multiple known risk factors to anemia. PAF is defined as the fraction of the cases in one disease or a condition in one population which attribute to a specific exposure. The estimation is relevant only when the exposure is in the causal pathway to the outcome and when the exposure can be modified by interventions12. The attributable fractions (AFs) assess the proportion of cases attribute to certain risk factors in the population, but the attributable fractions are rarely reported and are mostly calculated without considering potential confounders12. Definition of population attributable ratio is the reduction of the population's rate of disease or mortality when exposure to one risk factor is reduced to another ideal exposure scenario12. Since many diseases are caused by multiple risk factors and the effects of a single risk factor on the disease may interact, the PAFs for different risk factors of one patient can add up to more than 100%. Thus, PAF can be used to estimate the relative contribution of several known risk factors to anemia. In this equation, P is the

prevalence of the particular exposure or predictor variable and RR is the relative risk or odds ratio (OR) comparing the risk of the outcome in the exposed (RR) to unexposed (OR) (4).

$$PAF = \frac{P*(RR-1)}{P*(RR-1)+1} \quad (4)$$

2. Introduction

2.1 Reasons and significance of missing data

Missing data can occur for many reasons, such as subjects dropouts, loss to follow-up, no response to survey items, and a result of data entry errors[13]. Fail to handle missing data appropriately in the analysis can lead to systemic bias in the association or effect estimation, including direction and magnitude. System errors occur when missing values are improperly handled by researchers. It can have an impact, causing people to overestimate or underestimate the association of diseases with certain risk factors[14]. The systematic error also leads to a decrease in precision and research power[15]. Missing data on a unit basis, such as missing participants, reduce the validity and accuracy of trial results, as well as the external validity of trial results if certain groups of participants. For example, those who performed poorly dropped out of the study. In addition, fail to handle missing data appropriately could pose a significant threat to the internal validity of the results. The participants with missing data are systematically different with those with complete data, for example, the data are loss to follow-up because of the death of the participants[16]. In MCAR, fail to handle missing data appropriately can threaten the internal validity, because when we consider the participants

without missing data which are a random sample of the full study population. Therefore, the power of the study will decrease[17]. Considering MAR mechanism, although imputation can be produced, biases can still be introduced to the analysis. If we do not handle missing data appropriately under MNAR mechanism, a serious risk of biased results will be showed in the study[17]. In addition, fail to handle missing data appropriately also cause a loss of precision or efficiency in analysis[18].

2.2 Challenges of handling missing data

The challenges of handling missing data include the lack of a general method to handle missing data, unknown missing data mechanism, and applying appropriate methods corresponding to the mechanism of missing data. The primary reason of no general method to handle missing data is that the process of missing observations on each subject is usually unknown. The data cannot inform the process alone, and there can be different forms of missing data. For example, missing outcomes, missing covariates, missing both outcomes and covariates, and even missing one variable of interest for all the observations[19]. In most cases, simple techniques for dealing with missing data, for example the complete case studies, overall mean imputation, and the missing-indicator method, may produce biased results[2]. When the missing data is MCAR, complete and valid case studies provide valid but inefficient results. One commonly used method to deal with missing data as a whole means that the Missing-Indicator method provides biased results when missing data is MCAR. Also,

if the sample weights are omitted by the researchers during the analysis, the estimate of parameters can be biased. Furthermore, the findings from the analysis will not be representative for the whole population of interest[20]. When missing data mechanism is MAR and parameters from MAR are independent from the parameters in the analysis, then the missing is ignorable. However, if the missing data under MAR are not ignorable, the MAR mechanism will still produce valid result but will cause the problem of losing efficiency[21].

Given the challenges of handling missing data, many nutrition assessment studies do not account for missing data in the analysis, for example, the analysis of the micronutrient deficiencies, nutritional status and the determinants of anemia in 0-59 months age in children and in the non-pregnant women of reproductive age in Gambia and the study focused on the nutritional status and disease severity in children visit a primary health clinic in rural Gambia acutely [22,23]. Both of these two studies excluded the participants with missing data when estimating PAF of risk factors for anemia.

2.3 Objective

The objective of the study is to evaluate the impact of missing data on the estimation of PAF of anemia in children based on three commonly used approaches: 1) complete case analysis, 2) inverse probability weighting method, and 3) multiple imputation method. The current study evaluates the impact of missing data on the estimation of PAF of anemia using the

BRINDA data sets and provides recommendations of dealing with missing data in nutritional assessment studies.

3. Methods

3.1 Data source

The data for this analysis were extracted from the most recent BRINDA project dataset (www.brinda-nutrition.org), which included surveys that were conducted after 2004 and included at least one biomarker of inflammation (C-reactive protein [CRP] or α-1-acid glycoprotein [AGP]) and at least one measure of anemia (hemoglobin), biomarkers of iron (serum ferritin [SF] or soluble transferrin receptor [sTfR])[24]. Three surveys (Nicaragua 2005 (NI2005), U.S.A 2006 (US2006), and Pakistan 2011 (PK2011) in pre-school children (age range: 6–59 month) (PSC) were considered for the current analysis[25-27]. These three surveys were chosen because they had complete data in most variables of interest except SF (a primary determinant of anemia). Additionally, there were distinct missing data proportions of SF. Specifically, 32.75%, 13.49% and 4.48% were missing SF in NI2005, US2006, and PK2011, respectively.

3.2 Primary outcomes

The primary outcome was hemoglobin concentration (grams per liter). Anemia was defined as hemoglobin concentrations adjusted for altitude < 11.0g/dL[25].

3.3 Potential risk factors of anemia and covariates

The expected relationship between anemia and the selected potential risk factors for anemia is summarized in the introductory methodology article[28,29], which was used as the guideline of covariate selection. The most consistent predictors of anemia were child age, iron deficiency, inflammation and stunting, in multivariable pooled models[8]. Potential predictors of anemia were selected for our current study based on Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia (BRINDA) project to identify risk factors and their relative contributions to anemia in preschool children. These potential predictors included age, sex, hemoglobin, socioeconomic status, SF, CRP, and AGP. Age, hemoglobin, SF, CRP, and AGP were continuous variables, and sex, and socioeconomic status were categorical variables. SF was analyzed individually and combined with the use of a meta-analysis to apply arithmetic correction factors and use a regression correction approach[30]. To reduce confounding between indicators of iron deficiency with indicators of inflammation, regression-adjusted ferritin was used to estimate individual iron[8,31]. The adjusted concentration of SF was derived from linear regression models with the use of regression coefficients. The demographic characteristics including the age of children defined as a continuous variable in months, sex, and household Socioeconomic Status (SES). Household SES was defined by each survey on the basis of the poverty-index ratio in the United States,

and an asset score in Nicaragua and Pakistan as a categorical variable. Household SES scores were dichotomized to be used in bivariate analysis[8].

Inflammation defined as either CRP > 5mg/L or AGP >1 g/dL, iron deficiency (ID) defined as inflammation-adjusted ferritin < 12 μg/L, vitamin A deficiency (VAD) defined as inflammation-adjusted Retinol binding protein (RBP) or retinol <0.7 μmol/L, folate deficiency (FD) defined as folate red blood cell folate < 226.5 nmol/L or serum or plasma folate < 6.8 nmol/L, vitamin B12 deficiency (VB12D) defined as serum or plasma B12 < 150 pmol/L, and current or recent malaria defined as a positive diagnosis during assessing[25-27].

3.4 Population attributable fractions

The prevalence ratios and relative risks were calculated accounting for complex survey design and PAF for proximal risk factors of anemia were estimated. We chose Relative Risks (RR) instead of Odds Ratios (OR) because anemia was common (the incidence is 10% or more). Under this situation, to estimate an RR was more desirable because there was an increase in differences between RR and OR with increasing incidence rates[31-33].

3.5 Complete case analysis

First, we used the complete case analysis to examine the association between anemia and other covariates under the MCAR assumption. For NI2005, 1423 participants with 957 SF

records took part of this study. Moving to US2006, 1312 people participated this survey, however, SF had 1135 records and SES had 1259 records. Concerning to PK2011, 7477 participants took part in this survey. The missingness of SF in this country was the lowest with 7142 records.

3.6 Inverse probability weighting method

Second, we used the inverse probability weighting method under MCAR missing data assumption. Specifically, in US2006, 4% of SES were missing. To fully utilize all available data, we used single imputation to impute the missing SES of based on age, sex, hemoglobin, and CRP. Next, a missing indicator for SF was created. The probability of missing SF was predicted based on age, sex, hemoglobin, CRP, AGP, and SES (when available) using a logistic regression model. We fitted this logistic regression model to find which variables were associated with SF for each country. To adjust for selection bias, we computerized one set of weights, together with the probability of the dataset have complete data, for each country, in the numerator and other variables in this logistic regression model with or without missing data[34]. The numerator was calculated as the probability, which was directly from the data for each country[34]. Logistic regression was used to calculate the denominator with or without the missing data as the outcome and factors, which was associated to missing data, as independent variables[34]. The new weight was the getting from the numerator divided by the denominator. PROC LOGISTIC procedure was used to help to get the new weights of inverse

probability methods. Logistic regression described the relationship between the binary variable we built (use 0 to describe the observations with missing SF, and 1 to describe the observations without missing data) and a set of predictor variables (age, sex, hemoglobin, CRP, AGP, and SES). We found out the predicted values of the binary variables in order to build up new weight. After getting the new weight, we used the new weight to fit into our original model to get the new prevalence, relative risk, and PAF, for each country.

3.7 Multiple imputation method

We applied an iterative Markov chain Monte Carlo method. The original existing complete data (age, sex, hemoglobin, SES, AGP, and CRP) were included as predictors in each survey, to predict and imputed the value of the missing SF using regression models. We used PROC MI to impute the missing data. PROC MI procedure was used to obtain multiple imputation method results using the existing complete variables (age, sex, hemoglobin, SES, AGP, and CRP). We generated 5 imputations for each missing measurement for missing values (SF for NI2005, SF and SES for US2006, and SF for PK2011), SES and CRP in NI2005, AGP in US2006, and CRP in PK2011 were not used due to 100% missingness. The final PAFs for each country were the average of the results from 5 imputations. Furthermore, 95% confidence intervals were gotten for each prevalence and RR to compare the prediction of the multiple imputation method.

3.8 Statistical analysis

All the analyses were conducted in SAS version 9.4 software (SAS Institute). Initially, we compared subject characteristics between those with and without SF using chi-square test for categorical variables (i.e., sex, and socioeconomic status) and two-sample t-test for continuous variables (i.e., age, hemoglobin, CRP, and AGP). The prevalence of each risk factor was estimated using PROC SURVEYMEANS incorporating survey strata, cluster, and sampling weights. The sampling weights for each survey was provided by survey representatives. The weight used in this analysis was sampling weight for hemoglobin analysis. All these three surveys were complex surveys with both Cluster variable and Strata variable. Cluster variable was applicable if randomization is performed at the cluster level instead of the subject level. Strata variable was applicable based on geographic designation (rural/urban or regions or others). Survey-specific analysis was performed during this case. PROC GENMOD procedure was used to get the Relative Risk of each risk factor, in order to get PAF. A logistic regression model was built to analyze the binary outcomes. A log-poisson model was used to get relative risk. Relative risks or risk ratios (RR) were received for comparing groups with different sets of characteristics[34-36]. We estimated the RR with a Poisson regression model with a robust error variance. Age, sex, inflammation, vitamin A deficiency, and iron deficiency were used as covariates for NI2005, and cluster, strata, and anemia were used as class level information. For PK2011, age, sex, inflammation, vitamin A deficiency, and iron deficiency were used as covariate, and cluster, strata, anemia, and SES

were used as class level information. For US2006, age, sex, inflammation, and iron deficiency were used as covariates, and cluster, strata, anemia, and SES were used as class level information. Weight for hemoglobin was used as the weight in this regression model for each country.

4. Results

4.1 Population characteristics

For NI2005, there were 1423 participants with 67.25% (957) SF records. The SES and CRP records were fully missing in this country. Otherwise, all the other variables were not missing. Moving to US2006, 1312 people participated this survey, however, both SF and SES were partially missing with 86.51% (1135) observations and 95.96% (1259), separately. AGP record was fully missing and other variables did not have any missingness. Concerning to PK2011, 7477 participants took part in this survey. The missingness of SF in this country was the lowest with only approximately 5% missing. Also, there was only one variable fully missing, which is CRP. All the other variables do not have any missingness. (Table1).

4.2 Comparisons of variables with or without missing SF

In order to explore whether the data sets followed the MCAR mechanism or not, we compared the risk factors and covariates between those with and without missing data of hemoglobin (Table 2). For NI2005, 466 SF records were missing. There was no significant

difference in age (p=0.67) and sex (p=0.98) between those with and without missing SF in NI2005. However, significant differences in hemoglobin (p<0.0001) and AGP (p=0.0172) were found. In US2006, 177 SF records were missing. There was no significant difference in hemoglobin (p=0.1037), CRP (p=0.214), sex (p=0.1415), and SES (p=0.2003) between those with and without missing SF in NI2005. However, significant difference in age (p<0.0001) was found. In PK2011, 97 SF records were missing. There was no significant difference in age (p=0.8679), sex (p=0.2003), and SES (p=0.7261) between those with and without missing SF in NI2005. However, significant differences in hemoglobin (p<0.0001) and AGP (p<0.0001) were found.

4.3 PAF of anemia

In general, all the results of these three methods were similar. The results of complete case analysis were closer to inverse probability weighting method rather than multiple imputation method, generally. For NI2005, because the proportion of missing data was the largest comparing to other two countries, the differences between three methods were the largest. When we estimate the result to 2 decimal points, all the result for complete case analysis and inverse probability weighting method were the same (PR, RR and PAF), except the PAF for inflammation (with only 0.06% difference) and vitamin A deficiency (with only 0.1% difference). The result of multiple imputation method was larger in RR of inflammation than the result for complete case analysis and inverse probability weighting method . Otherwise, all

the other results of multiple imputation method were smaller than the results of other two methods. The other two countries, US2006 and PK2011, as their proportions of missing data were less than NI2005, the estimated results to 2 decimal points are probably the same, mostly. Some of the results from US2006 showed huger differences, such as RR and PAF of iron deficiency in three methods, and the RR and PAF of inflammation from complete case analysis and inverse probability weighting method compared with multiple imputation method (with 0.80 and 7.08% differences, separately). Because PK2011 included more observations, the results of this survey were the most similar among these three methods. The only differences of the result were the prevalence, RR, and PAF of iron deficiency from multiple imputation method and PAF of vitamin A deficiency from all these three methods. In general, the results of iron deficiency (both RR and PAF) were frequently different values, because of the missing SF values. (Table 3).

5. Discussion

We used three surveys (NI2005, US2006, and PK2011) from the BRINDA working group to estimate PAF of each potential risk factor for anemia in PSC. The findings of this work allow us to compare PAF estimates across three different methods for missing data analysis, including complete case analysis , inverse probability weighting method, and multiple imputation method. Overall, we found little difference in the estimation of PAF. Specifically,

complete case analysis and inverse probability weighting method yielded similar results, whereas some differences were found using multiple imputation method.

A similar previous study assessed selection bias in the estimation of childhood obesity prevalence using data from electronic health records using inverse probability weighting method and multiple imputation method[1]. The findings of the comparison with these three methods in this study are quite similar with our study, both of the analysis by these three methods got similar result. According to the result of our analysis, the results from complete case analysis and inverse probability weighting method are similar, because some covariates have significantly different distributions under the pattern of with or without missing SF. When some of the distributions of incomplete covariates are highly skewed compares with complete covariates, the multiple imputation method involves extension from complete cases to incomplete cases implicitly, because the variables in the complete cases is used to imputes the variables in the incomplete cases with missing values[6].

Comparing the characteristics between those with and without missing SF can provide insights into the appropriateness of the MCAR assumption. All of the three surveys had one to two potential risk factors that showed significant differences, including hemoglobin and AGP in NI2005, age in US2006, and hemoglobin and AGP in PK2011.

In previous studies, complete case analysis has been considered as a non-preferred approach especially when the excluded individuals were significantly different from the included ones. On the other hand, inverse probability weighting method and multiple imputation method are considered as preferred approaches that can reduce the bias resulting from missing data. The advantage of multiple imputation method is the efficiency over inverse probability weighting method. Inverse probability weighting method uses variables without missing data to derive the weights. When the distributions are skewed of some covariates under the missing pattern compared with complete covariates, the multiple imputation method can produce smaller Standard Errors compared to complete case analysis and inverse probability weighting method. In addition, one condition which inverse probability weighting is more preferable is when one specific individual with missing data has many missing values rather than one or two variables[6]. Therefore, although the results of inverse probability weighting method and complete case analysis are similar, we recommend to use multiple imputation method in this study.

The study has several limitations. One limitation of this study comes from the highly skewed distribution under missing pattern on some existing variables compared with the full variables without missing values. If the distributions of incomplete covariates are highly skewed compared with complete covariates, the missingness mechanisms can be MCAR, MAR, or combined mechanisms. Therefore, even if we use multiple imputation method to handle the

missing value, it still produces biased regression coefficient estimates for the incomplete covariates with skewed distributions compared with complete covariates[37]. In addition, this study is only based on a small number of surveys from the original BRINDA data sets. Therefore, the conclusion of the similarity across methods may not be generalized to all nutrition assessment studies. Moreover, we did not conduct a comprehensive comparison using all existing methods to deal with missing data. Other approaches may include available case analysis and Missing Indicator method. Similarly, we considered a limited number of covariates in the estimation of PAF. Potential unmeasurable confounding could exist. Both iron deficiency and vitamin A deficiency were considered as risk factors of anemia, but there remains uncertainty in the status of iron or vitamin A when inflammation is present[8]. Lastly, there may be uncertainty in PAF estimates and differences between population subgroups[38]. For current larger PAF analysis and implications, researchers should try different methods, in order to find out the best fitted model for their missing mechanisms in the data and the studies.

6. Conclusion

In conclusion, this research is the very first study to consider the missing data from the BRINDA project dataset. Indeed, after concerning different methods to handle the missing data, the result of PAF are slightly different using different methods as expected. In addition, the preferred method (multiple imputation method) showed different results from the

previous method from the BRINDA project, the complete case analysis. For future direction, researchers probably can use multiple imputation method applying regression switching with predictive mean matching (PMM) to improve the analysis focusing on the skewed distributions of missing covariates. Ultimately, researchers may want to pay attention to investigate the potential impact of missing data on their studies.

Reference

1. Groenwold R, Dekkers OM. Missing data: the impact of what is not there. European journal of endocrinology 2020; 183(4), E7–E9.
2. Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. Journal of clinical epidemiology 2006; 59(10), 1087–1091.
3. Santos MS, Pereira RC, Costa AF, Soares JP, Santos J, Abreu PH. Generating Synthetic Missing Data: A Review by Missing Mechanism. IEEE Access 2019; vol. 7, pp. 11651-11667, 2019.
4. Martín-Merino E, Calderón-Larrañaga A, Hawley S, Poblador-Plou B, Llorente-García A, Petersen I, Prieto-Alhambra D. The impact of different strategies to handle missing data on both precision and bias in a drug safety study: a multidatabase multinational population-based cohort study. Clin Epidemiol 2018; 10:643-654.
5. Pigott TD. A Review of Methods for Missing Data. Educational Research and Evaluation 2001; 7(4), 353–383.
6. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research 2011; *22*(3), 278–295.
7. Stevens GA, Finucane MM, De-Regil LM, Paciorek CJ, Flaxman SR, Branca F, Peña-Rosas JP, Bhutta ZA, Ezzati M. Global, regional, and national trends in haemoglobin concentration and prevalence of total and severe anaemia in children and pregnant and non-pregnant women for 1995–2011: a systematic analysis of population-representative data. The Lancet Global Health 2013; 1(1): e16-25.
8. Engle-Stone R, Aaron GJ, Huang J, Wirth JP, Namaste SM, Williams AM, Peerson JM, Rohner F, Varadhan R, Addo OY, Temple V, Rayco-Solon P, Macdonald B, Suchdev PS. Predictors of anemia in preschool children: Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia (BRINDA) project. The American journal of clinical nutrition 2017; 106(Suppl 1), 402S–415S.
9. Foote EM, Suchdev PS, Williams TN, Sadumah I, Sullivan KM, Oremo J, & Ruth, LJ (2013). Determinants of Anemia among Preschool Children in Rural, Western Kenya. The American Journal of Tropical Medicine and Hygiene 2013; 88(4), 757–764.
10. Namaste SM, Aaron GJ, Varadhan R, Peerson JM, Suchdev PS, BRINDA Working Group. Methodologic approach for the Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia (BRINDA) project. The American journal of clinical nutrition 2017; 106(Suppl 1), 333S–347S.
11. Suchdev PS, Namaste SM, Aaron GJ, Raiten DJ, Brown KH, Flores-Ayala R. Overview of the Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia (BRINDA) Project. Advances in Nutrition 2016; 7(2), 349–356.
12. Rückinger S, von Kries R, Toschke AM. An illustration of and programs estimating attributable fractions in large scale surveys considering multiple risk factors. BMC medical research methodology 2009; 9:7.

13. Chowdhry AK, Gondi V, Pugh SL. Missing Data in Clinical Studies. International Journal of Radiation Oncology*Biology*Physics 2021.
14. Myers KO. Missing Data and Systematic Bias. American Journal of Public Health 2017; 107(9), e14.
15. Hussain JA, White IR, Langan D, Johnson MJ, Currow DC, Torgerson DJ, Bland M. Missing data in randomized controlled trials testing palliative interventions pose a significant risk of bias and loss of power: a systematic review and meta-analyses. Journal of Clinical Epidemiology 2016; 74, 57–65.
16. Ferreira, Juliana Carvalho, & Patino, Cecilia Maria. Loss to follow-up and missing data: important issues that can affect your study results. Jornal Brasileiro de Pneumologia 2019; 45(2), e20190091.
17. Kristman V, Manno M, Côté P. Loss to Follow-Up in Cohort Studies: How Much is Too Much? European Journal of Epidemiology 2003; 19(8), 751–760.
18. Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. International Journal of Epidemiology 2019; 48(4), 1294–1304.
19. Zhou XH. Challenges and strategies in analysis of missing data. Biostatistics & Epidemiology 2019; 4(1), 15–23.
20. Missing data and complex samples: The impact of listwise deletion vs. subpopulation analysis on statistical bias and hypothesis test results when data are MCAR and MAR. Proceedings of the Joint Statistical Meetings, Survey Research Methods Section 2009.
21. Statistical analysis with missing data. 2nd Ed. Hoboken 2002.
22. Petry N, Jallow B, Sawo Y, Darboe MK, Barrow S, Sarr A, Ceesay PO, Fofana MN, Prentice AM, Wegmüller R, Rohner F, Phall MC, Wirth JP. Micronutrient Deficiencies, Nutritional Status and the Determinants of Anemia in Children 0–59 Months of Age and Non-Pregnant Women of Reproductive Age in The Gambia. Nutrients 2019; 11(10):2275.
23. Mark H, Been JV, Sonko B. Nutritional status and disease severity in children acutely presenting to a primary health clinic in rural Gambia. BMC Public Health 2019; 19, 668.
24. Iron deficiency anemia: assessment, prevention, and control. A guide for programme managers 2001.
25. Nicaragua Integrated Surveillance System of Nutrition Interventions Progress Report 2003–2005. Managua, Nicaragua: Ministry of Health (Nicaragua) 2008.
26. Pakistan National Nutrition Survey 2011 2012.
27. National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention 2006.
28. Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity 2011.
29. Namaste SM, Aaron GJ, Varadhan R, Peerson JM, Suchdev PS, BRINDA Working Group. Methodologic approach for the Biomarkers Reflecting Inflammation and Nutritional

Determinants of Anemia (BRINDA) project. The American journal of clinical nutrition 2017; 106(Suppl 1), 333S–347S.

30. Namaste SM, Rohner F, Huang J, Bhushan NL, Flores-Ayala R, Kupka R, Mei Z, Rawat R, Williams AM, Raiten DJ, Northrop-Clewes CA, Suchdev PS. Adjusting ferritin concentrations for inflammation: Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia (BRINDA) project. The American journal of clinical nutrition 2017; 106(Suppl 1), 359S–371S.
31. Serum ferritin concentrations for the assessment of iron status and iron deficiency in populations 2011.
32. Larson LM, Namaste S., Williams AM, Engle-Stone R, Addo OY, Suchdev PS, Wirth JP, Temple V, Serdula M, Northrop-Clewes CA. Adjusting retinol-binding protein concentrations for inflammation: Biomarkers Reflecting Inflammation and Nutritional Determinants of Anemia (BRINDA) project. The American journal of clinical nutrition 2017; 106(Suppl 1), 390S–401S.
33. McNutt LA. Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes. American Journal of Epidemiology 2003; 157(10), 940–943.
34. Sayon-Orea C, Moreno-Iribas C, Delfrade J, Sanchez-Echenique M, Amiano P, Ardanaz E, Gorricho J, Basterra G, Nuin M, Guevara M. Inverse-probability weighting and multiple imputation for evaluating selection bias in the estimation of childhood obesity prevalence using data from electronic health records. BMC Medical Informatics and Decision Making 2020; 20(1).
35. Zou G. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. American Journal of Epidemiology 2004; 159(7), 702–706.
36. Greenland S. Model-based Estimation of Relative Risks and Other Epidemiologic Measures in Studies of Common Outcomes and in Case-Control Studies. American Journal of Epidemiology 2004; 160(4), 301–305.
37. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Medical Research Methodology 2010; 10(1), 7.
38. Arriaga ME, Vajdic CM, Canfell K. The burden of cancer attributable to modifiable risk factors: the Australian cancer-PAF cohort consortium. BMJ Open 2017;7:e016178.

Appendix

Table 1. Numbers of observations with complete data for selected anemia risk factors by survey

Country	Total N	Serum Ferritin	Iron Deficiency	Vitamin A Deficiency	Folate Deficiency	Age	Sex	Socioeconomic Status	Hemoglobin	C-reactive protein	α-1-acid glycoprotein
NI2005	1423	957 (67.25%)	957 (67.25%)	1419 (99.72%)	n/a	1423 (100%)	1423 (100%)	n/a	1423 (100%)	n/a	1423 (100%)
US2006	1312	1135 (86.51%)	1135 (86.51%)	0 (0%)	1312 (100%)	1312 (100%)	1312 (100%)	1259 (95.96%)	1312 (100%)	1312 (100%)	n/a
PK2011	7477	7142 (95.52%)	7142 (95.52%)	7239 (96.82%)	n/a	7477 (100%)	7477 (100%)	7477 (100%)	7477 (100%)	n/a	7477 (100%)

Table 2. Comparisons of variables among subjects with or without missing serum ferritin (SF)

	NI2005		US2006		PK2011	
	missing SF (n=466)	non-missing SF (n=957)	missing SF (n=177)	non-missing SF (n=1135)	missing SF (n=335)	non-missing SF (n=7142)
Age, month	33.46 (14.93)	33.25 (15.21)	28.29 (7.92)	35.79 (14.07)	26.69 (14.37)	27.40 (15.32)
	34.40	34.17	29	36	25	25
	(6.18, 59.83)	(6.11, 59.89)	(12,55)	(12,59)	(6,59)	(6,59)
	p=0.6702		p<0.0001		p=0.8679	
Hemoglobin, g/dL*	117.21 (11.67)	120.32 (11.65)	126.23 (8.57)	125.10 (8.46)	96.84 (17.42)	103.35 (18.02)
	117.5(60,152)	121(69,160)	127(100,148)	125(84,151)	98(42,137)	105(41,178)
	p<0.0001		p=0.1037		p<0.0001	
C-reactive protein, mg/L *	.	.	2.20 (8.42)	1.43 (4.37)	.	.
	.	.	0.3(0.1,74.7)	0.20(0.1,81.1)	.	.
	.		p=0.214		.	
α-1-acid glycoprotein, g/L *	0.82 (0.28)	0.87 (0.32)	.	.	0.88 (0.36)	0.96 (0.41)
	0.76(0.4, 2.34)	0.81(0.11,3.44)	.	.	0.82(0.04,2.64)	0.88(0.03, 7.89)
	p=0.0172		.		p<0.0001	
Boys	235(50.43%)	482(50.37%)	79(44.63%)	574(50.57%)	186(55.52%)	3710(51.95%)
Girls	231(49.57%)	475(49.63%)	98(55.37%)	561(49.43%)	149(44.48%)	3432(48.05%)
Sex	p=0.9821		p=0.1415		p=0.2003	
Low Socioeconomic Status	.	.	90 (52.63%)	618 (56.80%)	139 (41.49%)	2924 (40.94%)
Medium Socioeconomic Status	.	.	66 (38.60%)	342 (31.43%)	142 (42.39%)	2946 (41.25%)
High Socioeconomic Status	.	.	15 (8.77%)	128 (11.76%)	54 (16.12%)	1272 (17.81%)
Socioeconomic Status	.		p=0.2003		p=0.7261	

*Mean (standard deviation), median (lower quartile, upper quartile), frequency count (%), and P-value (log transformation T-test and chi-square) were reported.
* CRP and SES were fully missing in NI2005; AGP was fully missing in US2006; CRP was fully missing in PK2011.

Table 3. Comparisons of selected anemia risk factors by three common methods by survey

			NI2005	US2006	PK2011
Prevalence	Inflammation	Complete Case Analysis	23.97%	6.00%	35.50%
		Inverse Probability Weighting	23.43%	6.38%	35.48%
		Multiple Imputation	23.97%	6.00%	35.50%
	Iron Deficiency	Complete Case Analysis	44.87%	13.31%	51.17%
		Inverse Probability Weighting	44.87%	13.31%	51.17%
		Multiple Imputation (95% CI[1])	35.73%	13.46%	49.77%
			(33.64%,38.01%)	(12.89%,14.02%)	(49.66%,49.88%)
	Vitamin A Deficiency	Complete Case Analysis	1.87%	.	52.18%
		Inverse Probability Weighting	1.54%	.	51.17%
		Multiple Imputation (95% CI[1])	1.87%	.	52.18%
Relative Risk	Inflammation	Complete Case Analysis	1.12	1.36	1.09
		Inverse Probability Weighting	1.12	1.38	1.09
		Multiple Imputation (95% CI[1])	1.45	1.27	1.09
			(1.36,1.54)	(1.22,1.32)	(1.09,1.09)
	Iron Deficiency	Complete Case Analysis	1.92	6.09	1.43
		Inverse Probability Weighting	1.92	7.22	1.43
		Multiple Imputation (95% CI[1])	1.22	6.58	1.4
			(1.48,1.57)	(5.90,7.27)	(1.39,1.40)
	Vitamin A Deficiency	Complete Case Analysis	2.58	.	0.97
		Inverse Probability Weighting	2.58	.	0.97
		Multiple Imputation (95% CI[1])	1.52	.	0.97
			(0.93,1.51)		(0.971,0.975)
	Inflammation	Complete Case Analysis	2.68%	2.12%	3.10%
		Inverse Probability Weighting	2.62%	2.26%	3.10%
		Multiple Imputation	9.70%	1.59%	3.09%

	Iron	Complete Case Analysis	29.30%	40.40%	18.06%
PAF	Deficiency	Inverse Probability Weighting	29.30%	45.32%	18.06%
		Multiple Imputation	7.00%	42.81%	16.44%
	Vitamin A	Complete Case Analysis	2.28%	.	-1.50%
	Deficiency	Inverse Probability Weighting	2.38%	.	-1.50%
		Multiple Imputation	0.97%	.	-1.44%

[1] CI: Confidence Interval.

Anemia defined as Hemoglobin adjusted for altitude <11.0 g/dL, inflammation defined as either CRP > 5mg/L or AGP >1 g/dL, Iron Deficiency defined as inflammation-adjusted ferritin < 12 µg/L, Vitamin A Deficiency defined as inflammation-adjusted RBP or retinol <0.7 µmol/L, Folate Deficiency defined as folate red blood cell folate < 226.5 nmol/L or serum or plasma folate < 6.8 nmol/L, Vitanin B12 deficiency serum or plasma B12 < 150 pmol/L.