# Imputation in missing not at random SNPs data using EM algorithm

Mahmood Alipour Heidari[1], Hamid Alavi Majd[2,*], Ebrahim Hajizadeh[1], Kamal Azam[3] , Mohammad Reza Zali[4]

[1] Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
[2] Department of Biostatistics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[3] Department of Epidemiology and Biostatistics, School of Public health, Tehran University of Medical Sciences, Tehran, Iran
[4] Research Institute for Gastroenterology and Liver Disease, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[*] Corresponding Author: e-mail address: alavimajd@gmail.com (H. Alavi - Majd)

## ABSTRACT

The relation between single nucleotide polymorphisms (SNPs) and some diseases has been concerned by many researchers. Also the missing SNPs are quite common in genetic association studies. Hence, this article investigates the relation between existing SNPs in DNMT1 of human chromosome 19 with colorectal cancer. This article aims is to presents an imputation method for missing SNPs not at random. In this case-control study, 100 patients suffering from colorectal cancer consulting with the Research Institute for Gastroenterology and Liver Disease of Shahid Beheshti University of Medical Sciences were considered as the case group and 100 other patients consulting with the same research institute were considered as the control group and the genetic test was applied in order to identify the genotype of the 6 SNPs of the DNMT1 of chromosom 19 for all the patients under investigation. The obtained data were analyzed using logistic regression, then a fraction of the data was eliminated both at random and not at random and the imputation was done through the EM algorithm and the logistic regression coefficients variation before and after the imputation was compared. The results of this study implied that in both methods, at random and not at random missing SNPs, the estimation of the logistic regression coefficients after the imputation through EM algorithm has a greater correspondence to the results obtained from the complete data in comparison with the method of eliminating the missing values.

**Keywords:** EM algorithm; Single Nucleotide Polymorphisms (SNPs); colorectal cancer; DNMT1; Human's 19th chromosome; Logistic regression; Missing Value

## INTRODUCTION

Single nucleotide polymorphisms (SNPs) are alterations in the DNA which are caused by variation in a single nucleotide (A, T, C, or G). For example, the DNA sequence may differ from AAGGCTAA to ATGGCTAA. To be considered as an SNP, the alteration should be observed at least in one percent of the society. The SNPs which are responsible for about 90 percent of the genetic alterations in human body occur per 100 to 300 bases. Two third of the SNPs are brought about by a change from T to C. The SNPs can be found in coding and noncoding areas [14].

Many of the SNPs do not affect the cell's operation, but some of them are capable of preparing the person for catching diseases or impacts their healing process. Although more than 99 percent of human DNA is common to all human beings, a change in the DNA can influence greatly the human's reaction to diseases, external invasions as bacteria, viruses, toxins, chemicals and also the required treatments [2]. As a result, the SNPs are so valuable in biomedical researches and manufacturing of chemicals and also in medical diagnoses. SNPs are hereditary and do not change from one generation to the other generations and this fact simplifies the population studies. The researchers believe that SNP maps is a great help for them in finding the effective genes in complicated diseases as cancer, diabetes, vascular diseases, and some mental diseases. Indentifying these associations through common methods is a difficult task, since a single gene may affect the trend of pathogenesis only slightly [10].

People with specific SNPs or a number of SNPs may be more susceptible when exposed to carcinogenic substances or radiations. A single SNP may be capable of increasing the risk of cancer, but considering the overlap rate and multiplicity in the DNA repair path, a single SNP can not affect the final result of the cancer greatly on its own. Even though, a number of polymorphic areas can increase the risk of cancer [9].

DNA (cytosine-5)-methyltransferase is an enzyme that in humans is encoded by the *DNMT1* gene [16]. DNA (cytosine-5-)-methyltransferase has a role in the establishment and regulation of tissue-specific patterns of methylated cytosine residues. Aberrant methylation patterns are associated with certain human tumors and developmental abnormalities [11].

The colorectal cancer involves the growth of cancerous cells in the large intestine, rectum and the epenthesis. With a worldwide annual mortality rate of 655 thousand people, this disease is the fourth widespread cancer in the U.S. and the third lethal cancerous disease in the western world [1,6]. The colorectal cancer is brought about by the growth of a gland in the large intestine. These tumors are fungous and are usually benign, but in some cases they turn to cancer by the time passing. The local colorectal cancer is usually diagnosed by colonoscopy.

Active cancers which are confined in the large intestine's walls (first and second stages of TNM) are remediable by a surgery. In the case the disease is not cured in this stage, they spread in the lymphatic glands of the same area (third stage). In this stage there is 70 percent chance of remedy through operation and chemotherapy. Cancers extended to more distant areas (forth stage) are usually untreatable, although the chemotherapy can increase the length of life and in rare cases operation along with chemotherapy has lead to the therapy of the patient. Radiography is also used in the therapy of rectum cancer [7].

Logistic regression is an analytic device which is widely utilized in medical and epidemiologic researches [13]. The objective of logistic regression is to acquire the best fitting and the most economical models for describing the relation between the binary or multi-mode ordinal response variable with one or a collection of independent variables [5].

In many of the medical data we encounter cases in which a part of them is not reported, e.g., answer avoidance, not completing the questionnaires or the records, incomplete research framework, etc. In such cases we should deal with missing data which cause many problems in the analyzing process. This fact has attracted much attention during the recent years. Missing data can exist in the covariates and the response variables. In this study it is deemed that the missing data occur in the covariates and the response variables are observed fully. Until now, many different methods have been proposed for analyzing conditional and unconditional category regression models with missing data in covariates. Satten and Carroll (2000), have estimated parameters based on the maximum likelihood method for binary response variables with missing values [11]. For a valid analysis, the knowledge about missing data mechanism is the key for the analysis. Hence, it is needed to clarify the missing data mechanism in order to choose the proper analytic method. In a univariate sample if the missing variables are in a random subsample of the main sample, we have missing at random (MAR). Missing at random is an ignorable mechanism. In the case the missing possibility depends on the value of the variable the mechanism is missing not at random (MNAR) and disregarding in such cases causes bias [3].

Since the missing SNPs are common in genetic studies and the statistical inference based on such data considering missing mechanism which may be at random or not at random, it is essential to consider some observations. From among methods we can deal with missing values, we can refer to first simply ignoring them and second the method of imputation. There are different methods proposed for imputation, the most important of which is Expectation-Maximization (EM) algorithm which aims at obtaining maximum likelihood estimation. In any iteration of EM algorithm E step aims at Expectation and M step aims at Maximization. In addition to data with missing values, the EM algorithm can be used for broken distributions, categorical observations or censored data, etc.

This study aims at investigating the appropriate methods for imputation in data with missing SNPs in a missing not at random mechanism (MNAR).

## MATERIALS AND METHODS

In this case-control study 6 SNPs from the DNMT1 gene of the 19th human's chromosome were investigated in two groups which are namely; rs61750053, rs62621087, rs16999358, rs61750052, rs16999593 and rs2228613. These two groups were patients consulting with the Research Institute for Gastroenterology and Liver Disease of Shahid Beheshti University of Medical Sciences in 2008. The case group patients were selected among patients suffering from colorectal cancer and the control group patients were selected among the rest of patients. The sample size for each group was 100 cases. The genetic test was applied for identifying the genotype of the patients in the laboratory of the research institute. The statistical analysis of the data was implemented using the logistic regression by the *haplo.stats* package which is a sub-package of R software. First, using the EM algorithm the imputation of the missing SNPs was implemented by the haplo.em function. Then the related coefficients were estimated by logistic regression and the statistical analyses were carried out for defining the significance level. After that, a fraction of the data was eliminated once at random and once at not random and each time the related coefficients were estimated by logistic regression for defining the significance level. Later, the imputation of the missing values was done using EM algorithm and again the estimation of the logistic regression coefficients was implemented. The statistical analysis for defining the significant level of them was also implemented and the required comparisons were carried out for analyzing the efficiency of the applied method.

The data of this research include a response and 6 independent variables. Y is a response binary variable in which Y=1 represents case group and includes patients suffering from colorectal cancer and Y=0 represents the control group. The independent variables included 6 SNPs as explained above. Each of these SNPs is a triple mode variable. In the first mode, there is no change in any of the paternal and maternal alleles. In the second mode, there is a change in one of the paternal or maternal alleles. And in the third mode the change exists in both paternal and maternal alleles.

The logistic regression formula in this study is:

$$\text{Logit}(P(Y=1|SNPs))=$$

$$\beta_0 + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \beta_5 S_5 + \beta_6 S_6$$

In the above formula $\beta_0$ to $\beta_6$ represent the logistic regression coefficients and $S_1$ to $S_6$ represent the SNP1 to SNP6. It should also be noted that in the above formula the interactions are ignored due to their little significance in the obtained results.

## RESULTS

In the first stage the related coefficients of logistic regression was estimated and the statistical test was implemented for identifying their significance level. The results of this stage are shown in table 1. As shown in table 1, all logistic regression coefficients, except the SNP5 coefficient, were proved to be significant. This implies that 5 SNPs have a significant relation to the colorectal cancer risk. Moreover, the Hosmer and Lemeshow Test for goodness of fit proved the model with a *P=0.929*.

Table1. Estimation of logistic regression coefficient for complete data

| Factor | β | SE | Z_value | P_value |
|---|---|---|---|---|
| Intercept | -22.39818 | 4.33931 | -5.162 | 2.45e-07 |
| SNP1 | 2.95896 | 0.65203 | 4.538 | 5.68e-06 |
| SNP2 | 2.94461 | 0.68161 | 4.320 | 1.56e-05 |
| SNP3 | 3.20642 | 0.69537 | 4.611 | 4.01e-06 |
| SNP4 | 3.89093 | 1.65556 | 2.350 | 0.0188 |
| SNP5 | 0.04331 | 0.74549 | 0.058 | 0.9537 |
| SNP6 | 2.54605 | 0.61218 | 4.159 | 3.20e-05 |

In the next stage, 10 percent of the SNPs were eliminated at random and considered as the missing values. Again the related coefficients were estimated using logistic regression after eliminating the missing values and the statistical test was implemented again for identifying the significance level of them. The results are shown in table 2. As shown in table 2, after eliminating the missing SNPs, in addition to the SNP5 coefficient, the SNP4 coefficient is not significant too. Moreover, the increase in SE values indicates a decrease in the accuracy of estimations and a decreased efficiency of the model.

Table2. Estimation of logistic regression for omitted missing data at random (MAR)

| Factor | β | SE | Z_value | p_value |
|---|---|---|---|---|
| Intercept | -20.80570 | 9.45546 | -2.200 | 0.02778 |
| SNP1 | 3.59483 | 1.14685 | 3.135 | 0.00172 |
| SNP2 | 3.68693 | 1.24776 | 2.955 | 0.00313 |
| SNP3 | 3.50666 | 1.12306 | 3.122 | 0.00179 |
| SNP4 | 0.05652 | 7.40516 | 0.008 | 0.99391 |
| SNP5 | -0.50530 | 1.10243 | -0.458 | 0.64670 |
| SNP6 | 2.96382 | 1.08223 | 2.739 | 0.00617 |

In the next stage, the imputation of the missing SNPs with an EM algorithm was implemented and again the related coefficients were estimated using logistic regression. As shown in table 3, all logistic regression coefficients, except the SNP5 coefficient, were proved to be significant. This implies that the imputation of the missing values was implemented with a high level of accuracy and the obtained results did not differ with the results of the complete data. The decrease in the SE values in the obtained model from the imputation indicates the increase of the estimation accuracy. The Hosmer and Lemeshow Test for goodness of fit also confirmed the results with *P=0.114.*

Table3. Estimation of logistic regression coefficient after the imputation by EM algorithm in random missing SNPs

| Factor | β | SE | Z_value | p_value |
|---|---|---|---|---|
| Intercept | -16.5672 | 2.9384 | -5.638 | 1.72e-08 |
| SNP1 | 1.8161 | 0.4450 | 4.082 | 4.47e-05 |
| SNP2 | 2.5603 | 0.5080 | 5.040 | 4.65e-07 |
| SNP3 | 2.6425 | 0.5013 | 5.272 | 1.35e-07 |
| SNP4 | 3.4068 | 1.4395 | 2.367 | 0.018 |
| SNP5 | -0.6013 | 0.6206 | -0.969 | 0.333 |
| SNP6 | 1.8173 | 0.4480 | 4.056 | 4.98e-05 |

In the next stage, 20% of the third mode, 15% of the second mode and 5% of the first mode of the primary SNP data were eliminated, to create not at random missing SNPs in a way that the missing mechanism depend on the value of the variable. Once again the related coefficients were estimated using logistic regression for the remaining cases. The results are presented in Table 4. As can be observed in table 4, after eliminating missing data, in addition to SNP5coefficient, the SNP4 coefficient was not also proved significant. The increase in SE values, like the situation in table 2, indicates a decrease in the accuracy of the estimations and the decreased efficiency of the model.

Table4. Estimation of logistic regression for omitted missing data not at random (MNAR)

| Factor | β | SE | Z_value | p_value |
|---|---|---|---|---|
| Intercept | -20.2969 | 5.3116 | -3.821 | 0.000133 |
| SNP1 | 3.0157 | 0.7497 | 4.022 | 5.76e-05 |
| SNP2 | 2.3809 | 0.7422 | 3.208 | 0.001337 |
| SNP3 | 2.8608 | 0.7999 | 3.577 | 0.000348 |
| SNP4 | 3.5146 | 3.1651 | 1.110 | 0.266818 |
| SNP5 | 0.3369 | 1.1787 | 0.286 | 0.775023 |
| SNP6 | 1.8753 | 0.6571 | 2.854 | 0.004317 |

In the next stage, the imputation of the missing not at random SNPs was implemented using EM algorithm and again the related coefficients were estimated using logistic regression. See the results in table 5. As can be seen in table 5, all logistic regression coefficients, except SNP4 and SNP5 coefficients are proved to be significant. This indicates that the imputation of the missing not at random values is less accurate in comparison with the status in which the missing mechanism was at random. The comparison of the estimated coefficients showed that after the imputation the estimations were more accurate than when the missing values were eliminated. The decrease of the SE values in table 5 in comparison with table 4 shows that the accuracy of the estimations and their efficiency is increased.

Table5. Estimation of logistic regression coefficient after the imputation by EM algorithm in not at random missing SNPs

| Factor | β | SE | Z_value | P_value |
|---|---|---|---|---|
| Intercept | -13.4132 | 2.1087 | -6.361 | 2.00e-10 |
| SNP1 | 2.1129 | 0.4393 | 4.809 | 1.51e-06 |
| SNP2 | 1.4526 | 0.3940 | 3.687 | 0.000227 |
| SNP3 | 2.0439 | 0.4357 | 4.691 | 2.72e-06 |
| SNP4 | 1.2840 | 0.9889 | 1.298 | 0.194167 |
| SNP5 | 0.9857 | 0.5934 | 1.661 | 0.096707 |
| SNP6 | 1.6694 | 0.3969 | 4.206 | 2.59e-05 |

The *P value* obtained in the Hosmer and Lemeshow Test for goodness of fit equated to *0.703* which certified the results obtained from the model.

## DISCUSSION

In a research carried out in 2003, William Grady stated that colorectal cancer is the third mortality cause in the U.S. and also concluded that 20 to 30 percent of people suffering from this disease have a provable hereditary factor [15].

In another study, James Dai, et al. (2006) applied some imputation methods for missing values and compared the results with the traditional method of eliminating missing values and concluded that "imputation generally improves efficiency over the standard practice of ignoring missing data". They also concluded that results obtained using the EM or WEM algorithms prove to be more reliable in comparison with other imputation methods [5].

In 2010, Martha Slattery, et al. investigated 561 cases and 721controls to show the relation between re4464148 with colorectal cancer. The result of this study showed that the odds ratio of catching colorectal cancer through a comparison with TT is OR=1.06 (95%CI:0.82-1.38) for CT and OR=1.86 (95%CI:1.17-2.96) for CC which is significant in a 0.04 level of significance [8].

Grittner, et al. (2011) investigated five different methods of imputation for missing not at random data and came up with the result that the "Bayesian approach yielded the most unbiased estimates for imputation" (p. 50) [3].

This study was an attempt to investigate the efficiency of different mechanisms of missing data and it was shown that using EM algorithm for the imputation of the missing data yields better statistical results in comparison with the standard method of eliminating missing data.

## ACKNOWLEDGEMENT

This research project would not have been possible without the sincere cooperation of the respected personnel of the research institute for gastroenterology and liver disease of Shahid Beheshti University of Medical Sciences, especially Dr. Mahdi Montazar Haghighi whose insightful guidance in genetics was a great help to this study. Moreover we would like to extend our gratefulness to Ms. Narges Alipour Heidari who Helped with the preparation of the English Version of the article.

## REFERENCES

1. Astin, M; Griffin, T, Neal, RD, Rose, P, Hamilton, W (2011 May). "The diagnostic value of symptoms for colorectal cancer in primary care: a systematic review.". *The British journal of general practice : the journal of the Royal College of General Practitioners* 61 (586): 231-43.

2. Bruce Carlson, (2008) "SNPs A Shortcut to Personalized Medicine". Genetic Engineering & Biotechnology News (Mary Ann Liebert, Inc.): p. 12.

3. Grittner, U., et al. (2011). Missing value imputation in longitudinal mearures of alcohol consumption. *International Journal of Methods in Psychiatric Research, 20(1),* 50-61.

4. Hosmer D.W. Lemeshow S. (1998), "Encyclopedia of biostatistics", vol. 3 : 2316-2333, John Wiley, New York.

5. James Y. Dai, Ingo Ruczinski, Michael LeBlanc, Charles Kooperberg, (2006)" Imputation Methods to Improve Inference in SNP Association Studies", Genetic Epidemiology 30: 690–702

6. Levin KE, Dozois RR (1991). "Epidemiology of large bowel cancer". *World J Surg* 15 (5):562–7.

7. Markowitz SD, Bertagnolli MM (2009)"Molecular basis of colorectal cancer". N. Engl. J. Med. 361 (25): 2449–60.

8. Martha L. Slattery, et al., (2010) " Increased Risk of Colon Cancer Associated with a Genetic Polymorphism of SMAD7 ", Cancer Research , Vol. 31 (4 ) : 70, 1479

9. Morita, A; Nakayama, T; Doba, N; Hinohara, S; Mizutani, T; Soma, M(2007) "Genotyping of triallelic SNPs using TaqMan PCR". Molecular and Cellular Probes 21 (3): 171–176.

10. Sachidanandam, R. et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 409:928–933

11. Rountree, M R; Bachman K E, Baylin S B (Jul. 2000). "DNMT1 binds HDAC2 and a new

co-repressor, DMAP1, to form a complex at replication foci". *Nat. Genet.* (UNITED STATES) 25 (3): 269–77.

12. Satten G.A. Carroll R.J. *(*2000*),* Conditional and unconditional categorical regression models with missing covariates, Biometrics, 56, 384-388.

13. Stuart R.L. Michael P. Marium E. (1998) "Inference using conditional logistic regression with missing covariates", Biometrics, vol. 54: 295-303

14. Väli, U; Brandström; Johansson; Ellegren (2008) "Insertion-deletion polymorphisms (indels) as genetic markers in natural populations" BMC genetics vol. 9(8): 10.1186/1471-2156-9-8

15. William M. Grady ,"Genetic testing for high-risk colon cancer patients", American Gastroenterological Association, Vol. 124 (6) 1574-1594 (2003)

16. Yen RW, Vertino PM, Nelkin BD, Yu JJ, el-Deiry W, (1992) Cumaraswamy A, Lennon GG, Trask BJ, Celano P, Baylin SB "Isolation and characterization of the cDNA encoding human DNA methyltransferase". Nucleic Acids Res 20 (9): 2287–91