THE CURIOSITY CUP 2023
A Global SAS® Student Competition

Smokeless tobacco consumption and non-fatal stroke: a cross sectional study with multiple imputation

Luigi Annicchiarico, Elia Gonzato, Università degli Studi di Milano-Bicocca
Team SASNuff

ABSTRACT

This paper aims to investigate the association between smokeless tobacco products consumption and non-fatal stroke event, through a cross-sectional study, using the 2020 BRFSS database. We accomplish this task by estimating a logistic model. As the percentage of missing values is significant, multiple imputation is performed and a second logistic regression is carried out. Then, we compare parameters from the two analyses, and comment results.

INTRODUCTION

Cardiovascular diseases, such as heart attack and stroke, are the first cause of death worldwide. Indeed, factors that might increase the probability of event of these acute pathologies have been widely studied. Stroke is a health issue in which brain cells death is caused by poor blood flow. Stroke is a variegated syndrome, and risk factors can be categorized as modifiable and nonmodifiable. Age, gender, and ethnicity are nonmodifiable risk factors, while hypertension, diet, physical inactivity and smoking, are among some of the most commonly reported modifiable risk factors. The harmful cardiovascular effects of smoking are well documented. However, health hazards associated with the use of smokeless tobacco remain controversial[1,2]. The purpose of this study is to explore whether the use of smokeless tobacco products is associated with non-fatal stroke. The analyses are performed using SAS® Studio on SAS® OnDemand for Academics.

DATA

We conduct our study using data taken from the 2020 Behavioral Risk Factor Surveillance System (BRFSS) [3], which is a cross sectional state-based telephone survey coordinated by the Centers for Disease Control and Prevention (CDC). Information about chronic conditions and health risk behaviours, which are self-reported, is gathered during the year using telephone surveys in all 50 U.S. states, Washington DC, Guam, Puerto Rico and the US Virgin Islands. More than 400,000 adults participate in the survey each year, contributing to make the BRFSS the largest telephone survey in the world[11]. Since data is taken from different states, weighting is taken into account for demographic and geographic variables.

PROBLEM

According to the National Cancer Institute, Smokeless tobacco is "a type of tobacco that is not smoked or burned. It may be used as chewing tobacco or moist snuff, or inhaled through the nose as dry snuff"[4]. So far, health hazards associated with the use of smokeless tobacco remain controversial. In the questionnaire proposed to people included in the database, the following sentence is asked: "Do you currently use chewing tobacco, snuff, or snus every day, some days, or not at all?". The answer is collected in a variable named "*snus*".

In order to address the presence of missing data, which could impact the magnitude of our findings, a second analysis with imputed data is carried out and its results are compared with the previous one.

DATA CLEANING/VALIDATION

Records with missing information regarding stroke (which are 1,186 in total) are excluded from the whole study. In the first analysis, observations which have at least one missing value for one of the variables that are to enter the model are 110,802 in total, and are initially excluded. So, in the first analysis our cohort consists of 289,970 participants, about 72% of the initial number.

Through a **DATA STEP** we create a second dataset, containing only variables of interest and with an **ARRAY** statement we clean the data. Then, we use **PROC FORMAT** for creating formats and we apply them to our new dataset. We dichotomize the *snus* variable in the following manner: people who answered "not at all" are categorized as "Non-users", while all the others are classified as "Users".

The other variables included in the model are:

1. **Gender**, with levels "Male" and "Female".
2. **Age**, arranged in categories from 18 to 84.
3. **Bmi**, reported as continuous variable.
4. **Ethnicity**, with levels "White", "Black", "Asian", "American Indian", "Hispanic", and "Other".
5. **Income**, with levels:
    a. "Less than $10,000 per year"
    b. "$10,000 to $15,000 per year"
    c. "$15,000 to $20,000 per year"
    d. "$20,000 to $25,000 per year"
    e. "$25,000 to $35,000 per year"
    f. "$35,000 to $50,000 per year"
    g. "$50,000 to $75,000 per year"
    h. "More than $75,000 per year".
6. **Alcohol**, with levels "yes" for people who had at least one drink per week during the past 30 days, and "no" otherwise.
7. **Smoke**, with levels "Never smoker", "Former smoker", "Someday smoker", "Everyday smoker".

The following table, obtained using **PROC MEANS**, reports the number of missing values for each variable:

Variable	Label	N Miss
Stroke	Stroke	0
Gen	Gender	0
Ethnicity	Ethnicity	0
Age	Age	8085
Snus	Smokeless tobacco	19561
Smoke	Smoke	20915
Alc	Alcohol	26565
Bmi	Bmi	41050
Income	Income	79543

ANALYSIS

DESCRIPTIVE ANALYSIS

As a first step, we conduct a descriptive analysis regarding our variables of interest. The following table, obtained using **PROC TABULATE**, reports numbers and percentages of the covariates, stratified by the outcome variable (*Stroke*). Univariate logistic regression is performed for each variable and Odds Ratios are added manually to the table. In this table only, Bmi is categorized following WHO guidelines[10]. Only observations without missing values are included in the table.

	Stroke				Total		OR (CI 95%)
	No		Yes				
	N	% Row	N	% Row	N	% Col	
Total	278696	96.1	11274	3.9	289970	100.0	
Gender							
Female	143665	96.1	5845	3.9	149510	51.6	REF
Male	135031	96.1	5429	3.9	140460	48.4	0.99 (0.95 - 1.03)
Ethnicity							
White	213249	96.1	8607	3.9	221856	76.5	2.10 (1.77 - 2.49)
Black	20041	94.4	1183	5.6	21224	7.3	3.07 (2.56 - 3.68)
Asian	6972	98.1	134	1.9	7106	2.5	REF
American Indian	4695	93.9	304	6.1	4999	1.7	3.37 (2.74 - 4.14)
Hispanic	24174	97.7	575	2.3	24749	8.5	1.24 (1.02 - 1.5)
Other	9565	95.3	471	4.7	10036	3.5	2.56 (2.11 - 3.11)
Income							
Less than $10k	10217	91.8	908	8.2	11125	3.8	4.87 (4.49 - 5.28)
$10k< income <$15k	11040	90.9	1099	9.1	12139	4.2	5.45 (5.05 - 5.89)
$15k< income <$20k	17136	92.6	1364	7.4	18500	6.4	4.36 (4.06 - 4.68)
$20k< income <$25k	22960	93.7	1539	6.3	24499	8.4	3.67 (3.43 - 3.93)
$25k< income <$35k	26705	94.9	1449	5.1	28154	9.7	2.97 (2.77 - 3.19)
$35k< income <$50k	38020	96.0	1587	4.0	39607	13.7	2.29 (2.14 - 2.45)
$50k< income <$75k	46321	97.1	1388	2.9	47709	16.5	1.64 (1.53 - 1.76)
More than $75k	106297	98.2	1940	1.8	108237	37.3	REF
Alcohol							
No	126137	94.6	7228	5.4	133365	46.0	REF
Yes	152559	97.4	4046	2.6	156605	54.0	0.46 (0.45 - 0.48)
Smoke							
Every day	28724	94.5	1684	5.5	30408	10.5	2.00 (1.89 - 2.12)
Some day	10754	94.7	607	5.3	11361	3.9	1.93 (1.77 - 2.1)
Former	76601	94.8	4219	5.2	80820	27.9	1.88 (1.8 - 1.96)
Never	162617	97.2	4764	2.8	167381	57.7	REF
Smokeless tobacco							
User	10195	96.2	401	3.8	10596	3.7	0.97 (0.88 - 1.08)
Non-user	268501	96.1	10873	3.9	279374	96.3	REF
Age							
18 to 24	16471	99.7	46	0.3	16517	5.7	REF
25 to 34	32998	99.4	209	0.6	33207	11.5	2.27 (1.65 - 3.12)
35 to 44	39200	98.8	475	1.2	39675	13.7	4.34 (3.2 - 5.87)
45 to 54	43663	97.6	1072	2.4	44735	15.4	8.78 (6.54 - 11.81)
55 to 64	56292	95.7	2501	4.3	58793	20.3	15.9 (11.87 - 21.29)
65 to 74	54886	94.3	3348	5.7	58234	20.1	21.83 (16.31 - 29.21)
75 to 84	35186	90.7	3623	9.3	38809	13.4	36.84 (27.53 - 49.3)

	Stroke				Total		OR (CI 95%)
	No		Yes				
	N	% Row	N	% Row	N	% Col	
Bmi							
Normal weight and underweight	86822	96.6	3082	3.4	89904	31.0	REF
Pre-obesity	100200	96.2	3961	3.8	104161	35.9	1.11 (1.06 - 1.17)
Obesity	91674	95.6	4231	4.4	95905	33.1	1.3 (1.24 - 1.36)

The univariate OR for smokeless tobacco consumption is found to be 0.97 (CI 95% 0.88 – 1.08). Hence, no significant association between smokeless tobacco and stroke is reported; however, confounders are not considered, and further investigation is needed.

Figure 1, which is obtained using **PROC GMAP**, shows the prevalence of people who had stroke and survived for each state. Prevalence ranges approximately from 1.50% to 7%, and countries with higher frequency of events are located in the southeast of the USA. On the other hand, *Figure 2* shows prevalence of Snus users in each state and it ranges approximately from 0.50% to 7.50%. Although we cannot identify a region with higher prevalence, we can say that Montana, Alaska and West Virginia are the states with the highest consumption of smokeless tobacco.
After doing this, we compare the prevalence of stroke stratified by ethnicity and income and, as shown in *Figure 3*, low income and ethnicities such as "Black" and "American Indian" are associated with higher risk of stroke. In addition, *Figure 4* shows a clear association between ethnicity and smokeless tobacco consumption, even though there is no important difference in smokeless tobacco consumption between income levels.

In order to assess the association between use of smokeless tobacco products and stroke risk, multiple logistic regression has been carried out using **PROC LOGISTIC**. Variables that entered the model were age, gender, bmi, ethnicity, income, alcohol, smoke and snus.

REGRESSION WITH EXCLUSION OF MISSING DATA

The flowchart in figure A summarizes the analysis strategy and the procedure we followed. We obtained this flowchart by using graph template language (**GTL**), with **TEXTPLOT**, **VECTORPLOT**, and **DRAWLINE** statements.

In the first analysis, where observations which have at least one missing value are excluded, the odds ratio of stroke for Snuff users is found to be 1.08 (95% confidence interval: 1.07 - 1.09) compared with non-users. This result suggests that smokeless tobacco does not have role as important as smoke[5] in increasing the risk of stroke. Furthermore, a ROC Curve (*Figure 5*) is produced and its area under the curve (AUC), which is a parameter that is widely used to evaluate the goodness of fit of the model, is 0.77, which is an acceptable parameter according to literature[12].

REGRESSION WITH MULTIPLE IMPUTATION

To comprehend whether this result is influenced by the presence of missing data, another logistic regression is conducted, and missing values are imputed before running the analysis. The imputation is performed using **PROC MI**: SAS offers plenty of options for multiple data imputation, based on the missing data pattern. In this case, as there were no notes regarding missing values, we assume that missingness is due to chance and not because of a particular reason. To impute missing data of categorical variables, the **FCS** statement is used, with the *logistic* option for dichotomous variables and the *discrim* option for the other categorical variables; for continuous variables, *reg* option is chosen. The option *classeffects=include* is used, as we were interested in considering class effects of the independent variables to determine missing data values. No trend is found in the trace plot regarding imputation of BMI, as shows *Figure 6*.

Multiple imputation is motivated by the Bayesian framework, hence, a prior is needed for the process to start. In the *discrim* option default is Jeffreys (which is a non-informative prior), but other priors are available. We built a **MACRO PROGRAM** (reported in the appendix) with the aim to determine if changing the prior would affect the imputation efficiency, which is an estimator of the quality of the process. However, differences in efficiency were not significant, probably because the amount of information available is elevated and the likelihood plays a major role; then we decided to use the *proportional*[6] prior.

We imputed five times, then, for each of them a logistic regression (**PROC LOGISTIC**) is conducted. Afterwards, we proceed using **PROC MIANALYZE**, which serves as a useful tool to summarize parameter estimates obtained from each logistic regression.

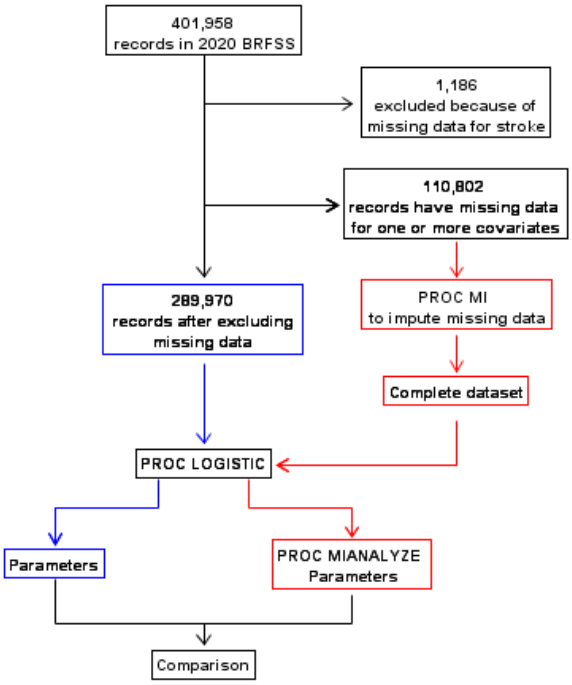

Figure A

Results are coherent with the previous logistic regression, as the odds ratio remains equal to 1.08 (CI 95%: 0.99 – 1.18) and no longer statistically significant. The confidence interval is wider than in the previous analysis; as we are imputing data, we expect a narrower interval. However, as we imputed five times, we should keep in mind that a new source of variability is introduced (*variance between imputations*). This leads to an increasement of variability which translates into a wider confidence interval. Parameters estimated after imputation were similar to the ones that were produced in the first model and this was a good result as, in particular in clinical research, imputing data is widely used as a sensitivity analysis, to verify robustness of both model and results. Anyway, caution is needed in drawing conclusions, as this method increased the standard error.

SUGGESTION FOR FUTURE STUDIES

Our study has several limitations: as the BRFSS is based on telephone interviews, we are not able to account for the effect of smokeless tobacco on people who died because of stroke. In fact, we are only able to evaluate the association with non-fatal stroke. This means that the real association between stroke and Snuff might be greater than what found in our model. Moreover, information upon type of stroke is not collected. Furthermore, the cross-sectional nature of the dataset does not allow us to investigate a causal effect between factors, but only to evaluate presence and strength of association. Then, non-response bias could impact the representativeness of the population. At last, several confounding factors could impact the estimates and results: the logistic regression performed by us accounts for some confounding factors only. Therefore, we suggest leading longitudinal studies, accounting for several confounders, to further investigate the association between use of smokeless tobacco and both non-fatal and fatal stroke.

CONCLUSION

Our findings are consistent with results obtained in several studies conducted in the past[1,6]. Therefore, we can conclude that smokeless tobacco does not have a role that is as impactful as smoke, in increasing the risk of non-fatal stroke. Also, multiple imputation showed robustness of the model.

REFERENCES

1) Hansson J, Galanti MR, Hergens MP, Fredlund P, Ahlbom A, Alfredsson L, Bellocco R, Engström G, Eriksson M, Hallqvist J, Hedblad B, Jansson JH, Pedersen NL, Trolle Lagerros Y, Ostergren PO, Magnusson C. 2014. "Snus (Swedish smokeless tobacco) use and risk of stroke: pooled analyses of incidence and survival." J Intern Med. 276(1):87-95.
2) Hergens MP, Lambe M, Pershagen G, Terent A, Ye W. 2008. "Smokeless tobacco and the risk of stroke." Epidemiology. 19(6):794-9.
3) Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/annual_data/annual_2020.html.
4) National Cancer Institute. Accessed January 18, 2023. https://www.cancer.gov/publications/dictionaries/cancer-terms/def/smokeless-tobacco.
5) Shah RS, Cole JW. 2010. "Smoking and stroke: the more you smoke the more you stroke." Expert Rev Cardiovasc Ther. 8(7):917-32.
6) Asplund K, Nasic S, Janlert U, Stegmayr B. 2003."Smokeless tobacco as a possible risk factor for stroke in men: a nested case-control study." Stroke. 34(7):1754-9.
7) Hajat C, Stein E, Ramstrom L, Shantikumar S, Polosa R. 2021. "The health impact of smokeless tobacco products: a systematic review." Harm Reduct J. 18(1):123.
8) SAS Help Center. "The MI Procedure." Accessed December 10, 2022. https://documentation.sas.com/doc/en/statug/15.2/statug_mi_syntax01.htm.
9) SAS Help Center. "The MIANALYZE Procedure." Accessed December 20, 2022. https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.3/statug/statug_mianalyze_syntax01.htm.
10) World Health Organization. Accessed January 23, 2023. https://www.who.int/europe/news-room/fact-sheets/item/a-healthy-lifestyle---who-recommendations.
11) Pierannunzi C, Hu SS, Balluz L. 2013. "A systematic review of publications assessing reliability and validity of the Behavioral Risk Factor Surveillance System (BRFSS), 2004-2011." BMC Med Res Methodol. 13:49.
12) Mandrekar JN. 2010. "Receiver operating characteristic curve in diagnostic test assessment." J Thorac Oncol. 5(9):1315-6.

APPENDIX

```
%macro efficiency(dist,k,imp,burn);
%do i=1 %to &k;
%let prior=%scan(&dist,&i);

proc mi data=prova1a nimpute=&imp out=ris_discrim_&prior seed=545;
var gen ethnicity age snus bmi income alc smoke;
class gen age ethnicity  snus income alc smoke;
fcs nbiter=&burn reg(bmi/details)
       logistic(snus/details link=glogit)
       logistic(alc/details link=glogit)
       discrim(age/details classeffects=include prior=&prior)
       discrim(income/details classeffects=include prior=&prior)
       discrim(smoke/details classeffects=include prior=&prior);
run;

proc logistic data=ris_discrim_&prior plots=roc;
class gen(ref='Male') ethnicity(ref='Asian') snus(ref='No') age(ref='18 to 24')
alc(ref='Yes') income(ref='More than $75k') smoke=(ref='Never') /  param=ref;
model stroke(ref='0')=ethnicity gen snus age bmi alc income/ link=glogit;
weight weight_analysis;
by _imputation_;
ods output ParameterEstimates=lgsparms_&prior;
run;

proc mianalyze parms(link=glogit)=lgsparms2_&prior;
modeleffects Intercept Female Snus Bmi Alcol
                    AmericanIndian Black Hispanic Other White
                    Age25to34 Age35to44 Age45to54
                    Age55to64 Age65to74 Age75to84
                    Between10and15 Between15and20 Between20and25 Between25and35
                    Between35and50 Between50and75 Less10 EveryDay SomeDay Former;
ods output VarianceInfo=vi_&prior;
run;

proc means data=vi_&prior mean;
var releff;
ods output summary=s_&prior;
run;

data s_&prior;
set s_&prior;
Prior="&prior";
run;

%end;
%mend;

%efficiency(dist=equal jeffreys proportional ridge,k=4,imp=5,burn=20)
```

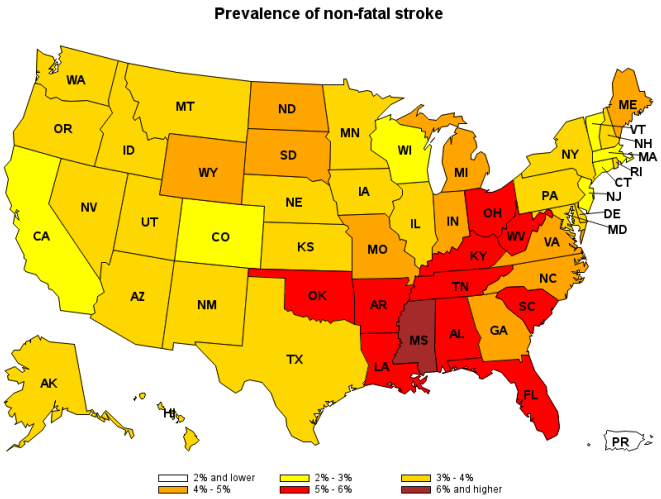

Figure 1

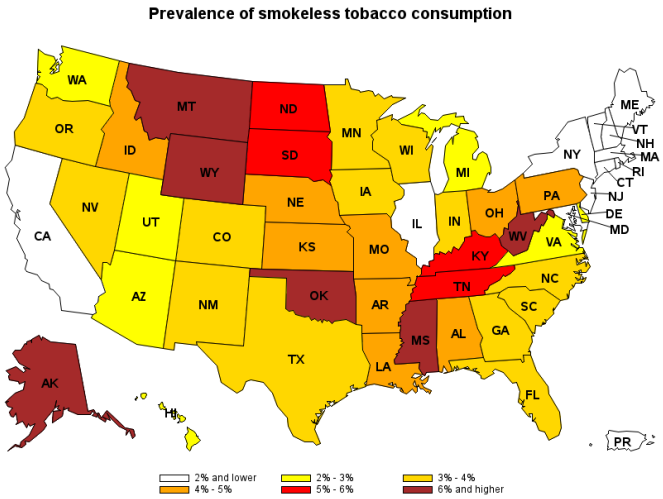

Figure 2

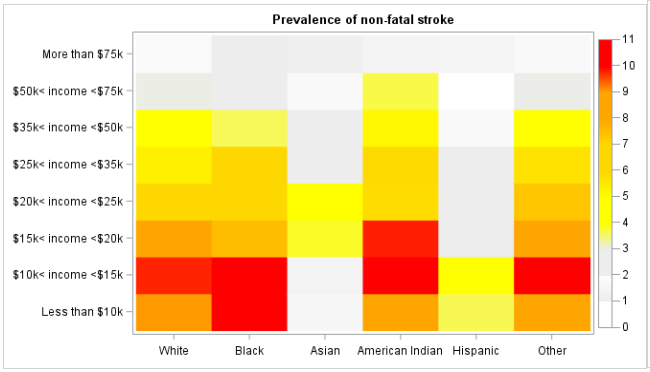

Figure 3

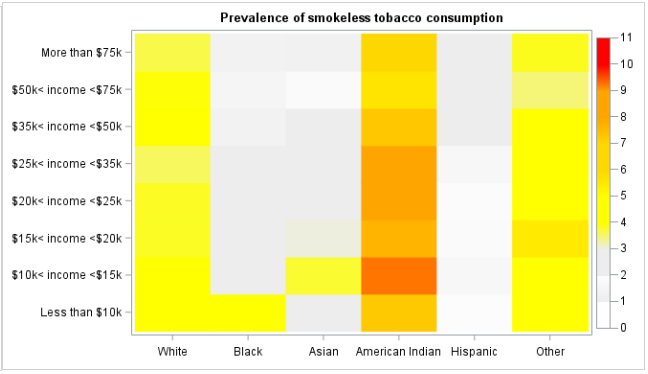

Figure 4

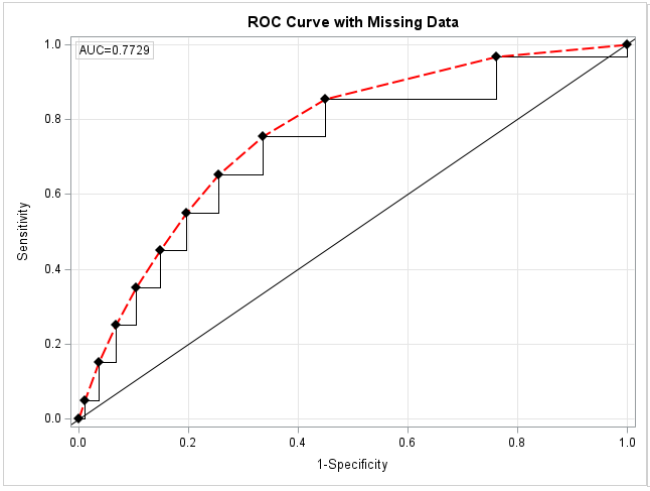

Figure 5

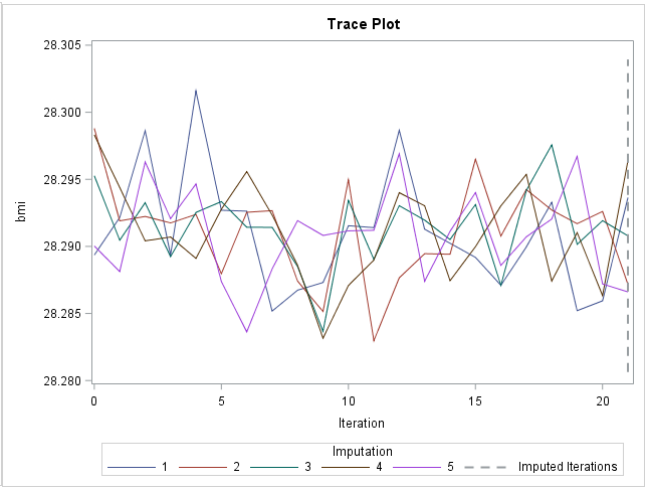

Figure 6