Prevalence and predictors of high nicotine dependence among adult smokers in Botswana,2017

Betelihem Getachew

4/9/23

Warning: package 'dplyr' was built under R version 4.2.2

Warning: package 'skimr' was built under R version 4.2.2

Warning: package 'png' was built under R version 4.2.2

# 1. Introduction

## 1.1 General Background Information

For countries like Botswana, and other African countries the prevalence of smoking is relatively low compared to other middle and high income countries. However, while the smoking prevalence might be low, not much has been said about smokers nicotine dependence in the region. Understanding nicotine dependence among smokers provides an opportunity for a targeted intervention strategies that are more efficient and effective.

Under the Global Tobacco Surveillance system Data (GTSSData) at CDC, the Global Adult Tobacco Survey (GATS) is the global standard to systematically monitor adult tobacco use and track key tobacco control indicators. The survey is a nationally representative household survey of adults 15 years of age or older, using a standard protocol. It is intended to generate comparable data within and across countries. GATS enhances countries’ capacity to design, implement and evaluate tobacco control interventions ([GATS Botswana Survey 2017](https://nccd.cdc.gov/GTSSDataSurveyResources/Ancillary/DataReports.aspx?CAID=2&Survey=4&WHORegion=3&Country=123&Site=27000)). Using GATS protocol, a nationally representative sample of 4,643 participants was collected in Botswana in 2017 using a stratified cluster sample design.

The aim of this paper is therefor to;

1. Describe the prevalence of nicotine dependence in Botswana using the Heavy Smoking Index(HSI) by Socio-demographics and other factors
2. Explore and Identify predictors of nicotine dependence in Botswana.

# 2. Methods

This study will conduct a secondary data analysis using the Global Adult Tobacco Survey (GATS) Botswana as described above. The survey data collection was completed in 2017. The sample design is a multi-stage, geographically clustered probability sample design to produce nationally representative data. First, households are randomly selected, then one individual is randomly chosen from each selected household to participate in the survey. The random selection of households and participants allows for an unbiased, randomly selected, and nationally representative sample of the larger population. The cluster sampling allows representation in gender and urbanicity. More details on the Global Adult Tobacco Survey implementation process can be found elsewhere ([GATS Implementation Protocol](https://nccd.cdc.gov/GTSSDataSurveyResources/Ancillary/Documentation.aspx?SUID=4&DOCT=4)).

A total of 608 variables covering topics of socio-demographic, tobacco smoking, smokeless tobacco, cessation both smoking and smokeless, economics, media and knowledge, attitude and perceptions of tobacco were collected in this survey. While the outcome measure of interest “Nicotine dependence” is not directly collected, the measure for nicotine dependence, also known as the Heavy Smoking Index (HSI), will be calculated as a score using two question items from the tobacco smoking section on the survey questionnaire ([Heaviness of Smoking Index | Data Share 2.0 (nih.gov)](https://datashare.nida.nih.gov/instrument/heaviness-of-smoking-index). Based on the HSI scores, smokers will be categorized into low addiction (score 0-2), medium addiction (score 3-4), high addiction (5-6). For this analysis, in consultation with other literature, scores 0 to 3 are categorized as low addiction and scores 4 to 6 are categorized as high addiction.

## 2.1 Measurements

The outcome of interest is HSI: Heavy Smoking Index. Independent variables explored in this analysis include AGE: Current age of smokers, A01: Gender A04:Education level, A05: Employment status, A11: Marital Status, RESIDENCE: Residence of participants, B04: Age of smoking initiation, D01: Smoking quit attempt, D08: Smoking intent to quit

## 2.2 Data import and cleaning

An SPSS data file was [downloaded from GTSS Info website](https://nccd.cdc.gov/GTSSDataSurveyResources/Ancillary/DataReports.aspx?CAID=2&Survey=4&WHORegion=3&Country=123&Site=27000) and imported into R. The code for importing and cleaning the dataset is documented in the R script file titled “processingcode.R”. The raw dataset contained 4643 observations and 608 variables, of which 591 variables were filtered out. Only 17 variables were considered for further examination.The variables of interest included socio-demographics, smoking status, smoking behaviors, cessation and media exposure. The 18th variable was computed by summing two variables (B01 + B07) to create the Heavy Smoking Index (HSI) Score, a proxy for examining nicotine dependence. The number of observations was further reduced by including only daily and less than daily smokers leaving a total of 631 observations. The dataset was examined for outliers, distribution, class appropriation. Recatagorization of response options was conducted when appropriate. After the removal of missing and outlier values, a total of 416 observations and 8 variables were saved for data exploration.

|  |
| --- |
|  |

—- exploredata ——–

## 2.3 Statistical analysis

*Explain anything related to your statistical analyses.*

# 3. Results

## 3.1 Exploratory/Descriptive analysis

In this sample of smokers, males with a median age of 35 years, those with secondary education and employed and belonging to the lowest wealth index had high nicotine addiction. There was also higher proportion of smokers who are single, live in rural for high addiction. Interms of behavior, more than half of those in the high addiction catagory made a quit attempt. similarly , the majoirty of smokers had no intent to quit within the next year . All codes related to data exploration are saved under code/analysis\_code with a file name of exploratorycodeR.R. All tables and figures listed below are saved under results folder

(**Table1\_Demographics?**) Shows the demographics of the sample population

result\_table1 <- readRDS(here("results", "Table1\_Demographics.rds"))  
result\_table1

|  | High Addiction | Low Addiction | Overall |
| --- | --- | --- | --- |
|  | (N=192) | (N=226) | (N=418) |
| **factor(A01)** | | | |
| Female | 26 (13.5%) | 33 (14.6%) | 59 (14.1%) |
| Male | 166 (86.5%) | 193 (85.4%) | 359 (85.9%) |
| **AGE** | | | |
| Mean (SD) | 40.1 (14.2) | 40.6 (15.5) | 40.4 (14.9) |
| Median [Min, Max] | 35.0 [19.0, 77.0] | 38.0 [16.0, 97.0] | 37.0 [16.0, 97.0] |
| **B04** | | | |
| Mean (SD) | 22.4 (8.95) | 20.2 (6.54) | 21.2 (7.81) |
| Median [Min, Max] | 20.0 [1.00, 70.0] | 19.0 [5.00, 52.0] | 20.0 [1.00, 70.0] |
| **factor(A04)** | | | |
| College and above | 27 (14.1%) | 32 (14.2%) | 59 (14.1%) |
| No Formal Education | 32 (16.7%) | 48 (21.2%) | 80 (19.1%) |
| Primary Education | 37 (19.3%) | 43 (19.0%) | 80 (19.1%) |
| Secondary Education | 96 (50.0%) | 101 (44.7%) | 197 (47.1%) |
| Missing | 0 (0%) | 2 (0.9%) | 2 (0.5%) |
| **factor(A05)** | | | |
| Employed | 108 (56.3%) | 122 (54.0%) | 230 (55.0%) |
| Unemployed | 84 (43.8%) | 104 (46.0%) | 188 (45.0%) |
| **factor(Wealth)** | | | |
| High | 21 (10.9%) | 28 (12.4%) | 49 (11.7%) |
| Higher | 34 (17.7%) | 40 (17.7%) | 74 (17.7%) |
| Low | 17 (8.9%) | 16 (7.1%) | 33 (7.9%) |
| Lowest | 86 (44.8%) | 99 (43.8%) | 185 (44.3%) |
| Middle | 34 (17.7%) | 43 (19.0%) | 77 (18.4%) |
| **factor(A11)** | | | |
| Married | 21 (10.9%) | 34 (15.0%) | 55 (13.2%) |
| Single | 171 (89.1%) | 192 (85.0%) | 363 (86.8%) |
| **factor(RESIDENCE)** | | | |
| Rural | 109 (56.8%) | 123 (54.4%) | 232 (55.5%) |
| Urban | 83 (43.2%) | 103 (45.6%) | 186 (44.5%) |
| **factor(D01)** | | | |
| No | 85 (44.3%) | 111 (49.1%) | 196 (46.9%) |
| Yes | 107 (55.7%) | 115 (50.9%) | 222 (53.1%) |
| **factor(D08)** | | | |
| No | 128 (66.7%) | 173 (76.5%) | 301 (72.0%) |
| Yes | 63 (32.8%) | 53 (23.5%) | 116 (27.8%) |
| Missing | 1 (0.5%) | 0 (0%) | 1 (0.2%) |

The data also shows the mean age of those with high addiction to be lower than the mean age of those with low addiction. This is contrary to the belief that older smokers have higher addiction when compared to younger smokers.

(**figure1?**) shows a boxplot showing HSI Group mean by age

|  |
| --- |
| HSI By Age. |

Interms of age of smoking initiation, those in the high addiction group have a mean age of smoking initiation slightly higher than those in the low addicition group. This again is contrary to the literature out there that suggests starting smoking early leads to high addiction in adulthood.

(**figure2?**) shows a boxplot showing HSI Group mean age of smoking initiation

|  |
| --- |
| HSI By Age of smoking Initiation. |

## 3.2 Basic statistical analysis

## 3.3 Full analysis

A binary 5 fold cross validation logisitc regression was performed. Model 1 used all predictors, Model 2 used only thet age of initiation as the predictor and model three used cessation behaviors (quit attempt and quit intention) as predictors. The tables below represent the results for all three models

(**Model1?**) All predictors

Model\_Results1 <- readRDS(here("results", "model1.rds"))  
Model\_Results1

# A tibble: 1 × 3  
 .metric .estimator .estimate  
 <chr> <chr> <dbl>  
1 roc\_auc binary 0.409

(**Model2?**) Age of initiation as predictor

model2 <- readRDS(here("results", "model2.rds"))  
model2

# A tibble: 1 × 3  
 .metric .estimator .estimate  
 <chr> <chr> <dbl>  
1 roc\_auc binary 0.432

(**Model3?**) Quit attempt and quit intent as predictors

model3 <- readRDS(here("results", "model3.rds"))  
model3

# A tibble: 1 × 3  
 .metric .estimator .estimate  
 <chr> <chr> <dbl>  
1 roc\_auc binary 0.432

### 3.3.1 Model Evaluation

Model 1 ROC Curve with all predictors

|  |
| --- |
|  |

Model 2 ROC Curve with age of smoking initiation only as the predictor

|  |
| --- |
|  |

Model 3 ROC Curve with quit attempt and quit intent as predcitors

|  |
| --- |
|  |

# 4. Discussion

## 4.1 Summary and Interpretation

All three models used in this anlaysis had an roc\_auc estimate of less than .5. This means all three models do not do a good job of predicting addiction with any more accuracy than a guess with a 50/50 outcome. This also implies further literature review is required to narrow down some of the most notable predictors.

## 4.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 4.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

These papers (McKay, Ebell, Billings, et al., 2020; McKay, Ebell, Dale, Shen, & Handel, 2020) are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 5. References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*(6228), 1314–1315. <https://doi.org/10.1126/science.aaa6146>

McKay, B., Ebell, M., Billings, W. Z., Dale, A. P., Shen, Y., & Handel, A. (2020). Associations Between Relative Viral Load at Diagnosis and Influenza A Symptoms and Recovery. *Open Forum Infectious Diseases*, *7*(11), ofaa494. <https://doi.org/10.1093/ofid/ofaa494>

McKay, B., Ebell, M., Dale, A. P., Shen, Y., & Handel, A. (2020). Virulence-mediated infectiousness and activity trade-offs and their impact on transmission potential of influenza patients. *Proceedings. Biological Sciences*, *287*(1927), 20200496. <https://doi.org/10.1098/rspb.2020.0496>