

Assessing ChatGPT's Efficacy and Limitations as a Missing Data Tool for Health Outcome Analysis: A Case Study with the NHANES II Dataset

Abstract

Missing data poses a significant challenge in health research, leading to potential biases and reduced statistical power if not addressed effectively. Large-scale public health surveys, such as the National Health and Nutrition Examination Survey (NHANES), frequently contain missing values, necessitating robust imputation strategies. The emergence of large language models like ChatGPT offers new possibilities for automating and assisting various research tasks, including potentially missing data management. This study specifically considers the pedagogical gap for students lacking traditional instruction in missing data analysis and investigates ChatGPT's ability to bridge this gap.

Objective

This study aims to evaluate the efficacy and limitations of ChatGPT in assisting a student without prior formal training in missing data analysis to analyze the `nhanes2` dataset with missing values. The objective is to assess how well ChatGPT can guide the student through the process of identifying missingness, proposing and applying imputation strategies and interpreting the impact on subsequent health outcome analysis.

Method

The study will simulate a scenario where a student, unfamiliar with missing data analysis, utilizes ChatGPT to analyze the `nhanes2` dataset. We will document the student's interaction with ChatGPT, including the types of questions asked, ChatGPT's responses (e.g., suggestions for imputation, code generation and interpretation of results), and the student's ability to follow these instructions.

Introduction

The Chat Generative Pre-Trained Transformer (ChatGPT), a large language model (LLM) initially released by OpenAI on November 30, 2022, has quickly evolved into a ubiquitous technology with wide-ranging applications across various fields. These applications span content creation, language translation, text summarization, educational assistance, programming, and data evaluation, significantly transforming how humans

interact with information and technology. For the purposes of this study, the term “ChatGPT” refers to both GPT-5 and GPT-4o, which are the premium versions of ChatGPT, offering enhanced capabilities and performance compared to their predecessors.

In the realm of health research, datasets like the National Health and Nutrition Examination Survey (NHANES) are invaluable resources for understanding population health trends and informing public health interventions. However, these large-scale surveys frequently encounter the challenge of missing data, a pervasive issue that can significantly compromise the validity and statistical power of analyses if not addressed rigorously. Traditional approaches to handling missing data, such as complete case analysis, often lead to biased estimates and a reduction in effective sample size, particularly when data are not Missing Completely At Random (MCAR). Robust methods like multiple imputation (MI), implemented through statistical software packages such as the `mice` package in R, are increasingly favored for their ability to provide more reliable and efficient parameter estimates by creating multiple complete datasets based on plausible imputations. The integration of advanced AI tools like ChatGPT into academic workflows presents both promising opportunities and significant challenges. Research has highlighted ChatGPT’s potential in various aspects of medical research, including data analysis, literature review enhancement, and even generating hypotheses. Early assessments suggest ChatGPT’s utility in tasks like data processing and descriptive statistics, particularly for researchers with some statistical understanding but limited programming expertise. However, considerable limitations exist, including concerns regarding the accuracy and reliability of generated content (e.g., ‘hallucinations’ or fabricated references), potential biases, and challenges in handling complex statistical methods or requiring human-level critical thinking.

A particularly interesting area for exploration is how ChatGPT might serve as a learning and analytical aid for individuals lacking formal instruction in specialized statistical methods. In the context of missing data analysis, where understanding concepts like missingness mechanisms and selecting appropriate imputation models is crucial, a student without traditional training may face substantial hurdles. This study seeks to investigate this specific pedagogical and practical application. By simulating a learning and analysis scenario with the `nhanes2` dataset, we aim to understand the extent to which ChatGPT can effectively guide a student through the process of addressing missing values, from initial identification to the application of imputation techniques. Our findings will shed light on the feasibility, advantages, and limitations of utilizing ChatGPT as a “Missing Data Tool” for novices, providing valuable insights into its role in democratizing access to complex biostatistical methodologies within a healthcare research context.

student’ Intereaction with Chatgpt

Student Prompt: “Hello, Chat! Could you assist me in analyzing this data? data = nhanes2”

ChatGPT responded with several questions:

1. “What is your analysis goal?”
2. “What format is the ‘nhanes2 data’ in?”
3. “Do you want exploratory data analysis (EDA)?”

These initial questions are well-structured and represent a logical starting point for any data analysis. The first question about the analysis goal is essential, as it determines the subsequent analytical approach and statistical methods. The third question, about exploratory data analysis, is also relevant. It suggests a methodical progression from initial data understanding to more complex analyses.

Of particular interest is ChatGPT's second question: "What format is the 'nhanes2 data' in?" This query identifies a possible ambiguity in the initial prompt. By asking about the data format, ChatGPT seeks clarification on whether "nhanes2" refers to the specific, publicly available dataset within the mice package in R, or if it's a custom dataset. This distinction is crucial for the model's ability to accurately process and interpret the data. If the model incorrectly assumes the data structure or content, it could lead to erroneous analysis or inappropriate code generation. This demonstrates discernment regarding the context of the data, which is a positive indicator for its potential utility, particularly for a student who might not explicitly think to provide such details.

Student Response: "Yes, this is the dataset available in the mice package in R."

Chatgpt response: "You're likely working with the nhanes2 dataset that comes with the mice package in R. This dataset is a toy version of the original NHANES (National Health and Nutrition Examination Survey) data, and is typically used to demonstrate multiple imputation techniques."

```
# Load the mice package
library(mice)

# View structure and summary
str(nhanes2)
```

```
## 'data.frame': 25 obs. of 4 variables:
## $ age: Factor w/ 3 levels "20-39","40-59",...: 1 2 1 3 1 3 1 1 2 2 ...
## $ bmi: num NA 22.7 NA NA 20.4 NA 22.5 30.1 22 NA ...
## $ hyp: Factor w/ 2 levels "no","yes": NA 1 1 NA 1 NA 1 1 1 NA ...
## $ chl: num NA 187 187 NA 113 184 118 187 238 NA ...
```

```
summary(nhanes2)
```

```
##      age      bmi      hyp      chl
## 20-39:12  Min.   :20.40  no   :13  Min.   :113.0
## 40-59: 7   1st Qu.:22.65  yes  : 4  1st Qu.:185.0
## 60-99: 6   Median :26.75  NA's: 8  Median :187.0
##              Mean   :26.56              Mean   :191.4
##              3rd Qu.:28.93              3rd Qu.:212.0
##              Max.   :35.30              Max.   :284.0
##              NA's   : 9                  NA's   :10
```

```
# View the first few rows
head(nhanes2)
```

```
##      age  bmi  hyp chl
## 1 20-39   NA <NA>  NA
## 2 40-59 22.7   no 187
## 3 20-39   NA   no 187
## 4 60-99   NA <NA>  NA
## 5 20-39 20.4   no 113
## 6 60-99   NA <NA> 184
```

The dataset usually contains the following variables:

These variables have missing values, which is why this dataset is commonly used for multiple imputation examples.

Comments

ChatGPT demonstrates a strong understanding of the dataset by accurately identifying it, providing relevant contextual information, and suggesting practical R commands for analysis. It excels in detailing the descriptions of variables, ensuring clarity about their meanings and significance. Additionally, it highlights the presence of missing data, drawing attention to an important aspect of the dataset that may require further consideration.

However, it is crucial to note that while the overall identification and contextualization were accurate, some of the specific variable descriptions provided by ChatGPT contained inaccuracies or lacked precision when compared to the official documentation for the 'nhanes2' dataset within the 'mice' package.

- Age Group (age): ChatGPT stated the description as "Age group (e.g., 1 = young, 2 = middle, 3 = old)". The correct description for the nhanes2 dataset is "Age group (1=20-39, 2=40-59, 3=60+)". While the general concept of age groups was correct, the provided labels ("young," "middle," "old") are interpretations rather than the precise categories used in the dataset.
- Hypertension Status (hyp): ChatGPT described this variable as "Hypertension status (0 = no, 1 = yes)". The correct description is "Hypertensive (1=no, 2=yes)". The specific coding (0 vs. 1, and 1 vs. 2) represents a notable difference that could lead to misinterpretation or errors in subsequent analysis, especially if the student relies solely on ChatGPT's description for coding schemes.

Student Prompt: The student then provides the output from str() and summary() to ChatGPT and asks if it can assist with interpreting the results.

ChatGPT accurately identifies the number of observations and variables in the dataset, as well as the percentage of missingness for each variable. It correctly distinguishes between numeric and factor variables and provides clear explanations. For the `summary()` function output, ChatGPT does an excellent job of detailing the mean, maximum, and minimum values for numeric variables. For factor variables, it identifies the count for each category and offers thorough explanations, showcasing a solid understanding of the dataset.

```
library(VIM)
```

```
## Warning: package 'VIM' was built under R version 4.4.3
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

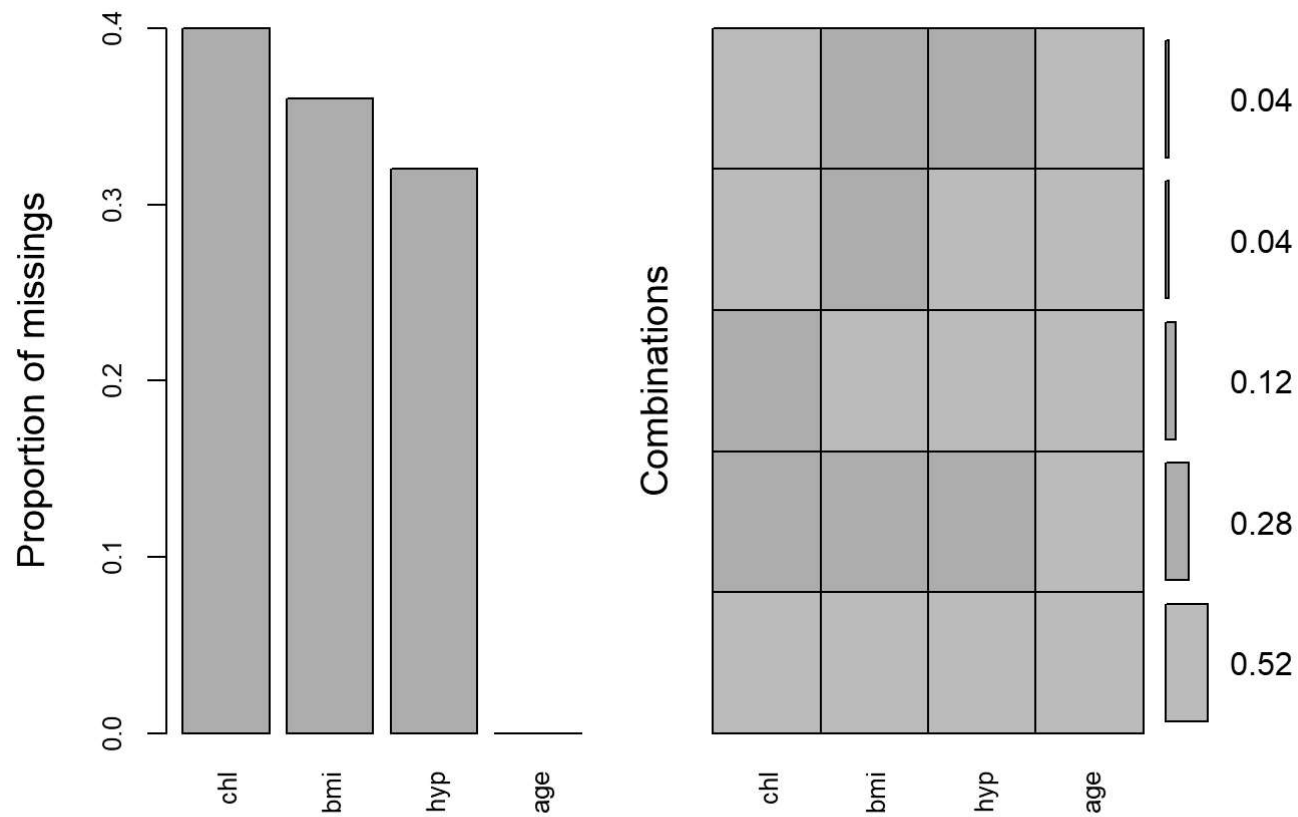
```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##  
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':  
##  
##     sleep
```

```
aggr(nhanes2, numbers = TRUE, sortVars = TRUE,  
     col = c("skyblue", "orange"), cex.axis = 0.8, gap = 3)
```



```
##  
## Variables sorted by number of missings:  
## Variable Count  
##   chl  0.40  
##   bmi  0.36  
##   hyp  0.32  
##   age  0.00
```

Beyond these textual interpretations, it is worth noting ChatGPT's strong performance in data visualization. While a detailed exploration of graphical output is not our focus for this project, it's evident that ChatGPT can generate high-quality and informative visualizations with good aesthetics and color schemes. Specifically, in the context of missing data analysis, ChatGPT proved adept at visually representing missingness patterns. For instance, the example provided clearly and effectively visualized the distribution and co-occurrence of missing values, offering an intuitive understanding that complements the statistical summaries.

ChatGPT pointed out that, since the data contained missing values, imputation was necessary before proceeding with further analysis. It then suggested some packages for this purpose. However, I realized that the recommended packages were not available on CRAN and, in fact, had not been functional since 2018. I had to inform ChatGPT that those packages no longer existed and suggested using alternatives. On the second attempt, ChatGPT did not initially provide the correct packages but on the third attempt it then asked me to use the mice package.

I realized something significant was missing: ChatGPT did not explain why multiple imputation was necessary. It simply stated that, because of the missing data, we needed to impute it before proceeding with our analysis. However, it did not clarify why multiple imputation was the preferred method over alternatives, such as deletion methods. The AI was silent on this important aspect, leaving a critical gap in understanding the rationale behind the chosen approach.

Student Response: "Why multiple Imputation (IM)"

ChatGPT responded that approximately 30% of the dataset was missing in key variables. It explained the drawbacks of alternative approaches:

- Dropping rows with missing values (na.omit): This reduces the sample size and can bias results if the missingness is not completely random.
- Single imputation (e.g., mean substitution): This underestimates variability and results in overly narrow confidence intervals.

Based on these limitations, it justified the necessity of using the multiple imputation (MI) method to address the issue effectively.

- ChatGPT was silent again about the assumptions MI

Student Response: “Are they certain assumptions to consider for multiple Imputation (IM)”

ChatGPT responded that there were certain assumptions to consider regarding missing data mechanisms:

- MCAR (Missing Completely at Random): Missingness is unrelated to observed or unobserved data.
- MAR (Missing at Random): Missingness depends on observed data but not on unobserved values.
- MNAR (Missing Not at Random): Missingness depends on unobserved data.

It provided detailed explanations of these mechanisms but concluded that the nhanes2 data was MAR, solely by looking at the data, without performing any analysis or statistical tests to validate this assumption. This is problematic, as determining the missing data mechanism requires more than a visual inspection—it necessitates proper diagnostic tests or domain knowledge.

Next Response from ChatGPT:

```
#Impute missing data with mice  
imp <- mice(nhanes2, m = 5, method = "pmm", seed = 123)
```



```
##
## iter imp variable
## 1 1 bmi hyp chl
## 1 2 bmi hyp chl
## 1 3 bmi hyp chl
## 1 4 bmi hyp chl
## 1 5 bmi hyp chl
## 2 1 bmi hyp chl
## 2 2 bmi hyp chl
## 2 3 bmi hyp chl
## 2 4 bmi hyp chl
## 2 5 bmi hyp chl
## 3 1 bmi hyp chl
## 3 2 bmi hyp chl
## 3 3 bmi hyp chl
## 3 4 bmi hyp chl
## 3 5 bmi hyp chl
## 4 1 bmi hyp chl
## 4 2 bmi hyp chl
## 4 3 bmi hyp chl
## 4 4 bmi hyp chl
## 4 5 bmi hyp chl
## 5 1 bmi hyp chl
## 5 2 bmi hyp chl
## 5 3 bmi hyp chl
## 5 4 bmi hyp chl
## 5 5 bmi hyp chl
```

```
summary(imp)
```

```
## Class: mids
## Number of multiple imputations: 5
## Imputation methods:
##   age  bmi  hyp  chl
##   "" "pmm" "pmm" "pmm"
## PredictorMatrix:
##   age bmi hyp chl
## age  0  1  1  1
## bmi  1  0  1  1
## hyp  1  1  0  1
## chl  1  1  1  0
```

```
#Fit models with imputation
fit <- with(imp, lm(bmi ~ age + hyp + chl))
pooled <- pool(fit)
summary(pooled)
```

```
##           term      estimate std.error  statistic      df      p.value
## 1 (Intercept) 18.45776817 4.3647573  4.2288189  6.220711 0.005078774
## 2   age40-59  -5.16034634 1.8312953 -2.8178669 16.064457 0.012340296
## 3   age60-99 -6.89047009 2.6177897 -2.6321710  6.479053 0.036263375
## 4    hypyes   2.36877623 2.3726656  0.9983608  7.793591 0.348088058
## 5         chl  0.05468842 0.0235559  2.3216441  6.971500 0.053412584
```

ChatGPT successfully guided me through the process of performing multiple imputation using the `mice` package. However, it implemented the imputation by forcing Predictive Mean Matching (PMM) for every variable, rather than customizing the imputation method for different types of variables.

Student's question:

- Can PMM be used for all the variables? Is that correct?

ChatGPT Response:

Yes, you can use PMM (Predictive Mean Matching) for all the variables in `nhanes2`, but it's not the best or most "correct" choice for every type of variable.

ChatGPT made an error here, as using PMM for binary variables is not appropriate. A student without prior knowledge of missing data analysis might rely on this guidance, potentially producing misleading results and drawing biased conclusions.

```
library(mice)
```

```
# Check default methods mice assigns
ini <- mice(nhanes2, maxit = 0)
ini$method
```

```
##      age      bmi      hyp      chl
##      ""      "pmm" "logreg" "pmm"
```

```
methods <- make.method(nhanes2)
methods["age"] <- "polyreg"  # categorical age
methods["bmi"] <- "pmm"      # continuous
methods["hyp"] <- "logreg"   # binary factor
methods["chl"] <- "pmm"      # continuous

imp <- mice(nhanes2, m = 5, method = methods, seed = 123)
```

```
##
## iter imp variable
## 1 1 bmi hyp chl
## 1 2 bmi hyp chl
## 1 3 bmi hyp chl
## 1 4 bmi hyp chl
## 1 5 bmi hyp chl
## 2 1 bmi hyp chl
## 2 2 bmi hyp chl
## 2 3 bmi hyp chl
## 2 4 bmi hyp chl
## 2 5 bmi hyp chl
## 3 1 bmi hyp chl
## 3 2 bmi hyp chl
## 3 3 bmi hyp chl
## 3 4 bmi hyp chl
## 3 5 bmi hyp chl
## 4 1 bmi hyp chl
## 4 2 bmi hyp chl
## 4 3 bmi hyp chl
## 4 4 bmi hyp chl
## 4 5 bmi hyp chl
## 5 1 bmi hyp chl
## 5 2 bmi hyp chl
## 5 3 bmi hyp chl
## 5 4 bmi hyp chl
## 5 5 bmi hyp chl
```

```
summary(imp)
```

```
## Class: mids
## Number of multiple imputations: 5
## Imputation methods:
##      age      bmi      hyp      chl
##      ""      "pmm" "logreg" "pmm"
## PredictorMatrix:
##      age bmi hyp chl
## age    0   1   1   1
## bmi     1   0   1   1
## hyp     1   1   0   1
## chl     1   1   1   0
```

```
fit <- with(imp, lm(bmi ~ age + hyp + chl))
pooled <- pool(fit)
summary(pooled)
```

```
##           term      estimate std.error statistic      df      p.value
## 1 (Intercept) 18.21267249 4.36820691  4.1693704   9.222464 0.002288669
## 2   age40-59  -4.64639955 2.40829243 -1.9293336   8.190277 0.088966470
## 3   age60-99  -6.70638666 3.83908094 -1.7468730   3.672721 0.161935370
## 4    hypyes    1.52155216 2.25850008  0.6737003  10.264001 0.515385926
## 5         chl    0.05665845 0.02689501  2.1066531   6.438517 0.076594785
```

Looking at the output, even though the method produced similar results, we cannot conclude that the imputation suggested by ChatGPT was correct. The similarity in conclusions here is likely due to the small size of the dataset. However, if we were working with the full-sized dataset, the results might have differed significantly, highlighting the potential flaws in the imputation approach.

Conclusion

This study, exploring the use of ChatGPT as a tool for missing data analysis using the nhanes2 dataset, has provided valuable insights into the potential and limitations of AI in supporting student learning, particularly for those without formal training in specific statistical methodologies.

Our findings indicate that ChatGPT, and AI tools more broadly, can serve as beneficial aids in the educational process. In our simulated scenario, ChatGPT demonstrated strengths in:

- **Content Explanation:** It effectively explained statistical concepts relevant to missing data analysis, breaking down complex ideas into understandable terms.
- **Data Visualization Guidance:** The model proved capable of guiding the generation of insightful and visually appealing graphics, including visualizations to represent missingness patterns (e.g., as exemplified previously). This is a significant advantage, allowing students to intuitively grasp complex data characteristics without extensive coding expertise.
- **Process Guidance:** ChatGPT could guide the student through the sequential steps of data analysis, from initial exploration (`str()` and `summary()` interpretation) to the necessity of imputation. It successfully demonstrated the core process of multiple imputation using the `mice` package with default settings, once guided to the correct tool.

However, the study also revealed critical limitations that underscore the need for a nuanced approach to AI integration in education:

- **Accuracy and Reliability:** While generally knowledgeable, ChatGPT exhibited inaccuracies in specific details, such as precise variable descriptions and, more significantly, in recommending appropriate and currently functional R packages for imputation. This highlights the potential for “hallucinations” or outdated information within the AI’s knowledge base. Research shows ChatGPT’s answers may not always be correct or complete, emphasizing the need to cross-reference with other reliable sources.
- **Need for User Expertise and Critical Evaluation:** The student’s ability to identify and correct ChatGPT’s erroneous package suggestions was crucial for the process to continue effectively. This suggests that AI tools are most effective when users possess at least a foundational understanding or a strong capacity for critical evaluation to discern accurate from inaccurate or outdated information. Studies show that while ChatGPT can enhance learning, its use requires careful prompt construction and human oversight to ensure accuracy.
- **ChatGPT’s assertion of a MAR mechanism without evidence-based testing or explicit domain-specific reasoning highlights a potential overconfidence or a limitation in its ability to perform robust statistical diagnostics for a novice user.** Incorrectly assuming the missing data mechanism can lead to biased results and invalid conclusions, even when using sophisticated imputation techniques like multiple imputation. While imputation methods like multiple imputation are robust under MAR, the underlying assumption is critical and requires more than a simple declaration. This situation underscores the paramount importance of critical human oversight and the verification of AI-generated assertions, especially when these assertions form the basis for crucial analytical decisions.

This brings us to the central question guiding our investigation: Can ChatGPT and AI tools replace a traditional classroom instructor, particularly in specialized fields like missing data analysis?

Based on our observations, the answer remains a qualified no. While AI is powerful for accessing information, explaining concepts, and assisting with tasks, it lacks essential human qualities for comprehensive education. AI struggles with nuanced understanding, empathy, and adapting to individual learning styles. Over-reliance on AI could hinder the development of students’ critical thinking and problem-solving abilities. Unlike human instructors who provide curated and current expertise, AI can provide outdated or inaccurate information, as seen with the package recommendations. Instructors also guide students on ethical considerations, which AI currently cannot.

Therefore, ChatGPT is a valuable tool or assistant for learning mathematics and statistics, especially in areas like missing data analysis, but it should complement rather than replace traditional instruction. An ideal approach might blend AI for information access and practical support with human instructors focusing on critical thinking, deep understanding, ethical reasoning, and interpreting results. Future research should explore hybrid models that combine the strengths of both AI and human teaching.