



Statistical analysis using ChatGPT in medical research

Soo-Nyung Kim, MD, PhD

Department of Obstetrics and Gynecology, Konkuk University School of Medicine, Seoul, Korea

This study aimed to explore the utility of ChatGPT in streamlining statistical analyses within medical research, evaluating its capabilities in data management, exploratory data analysis (EDA), statistical test selection, and result interpretation. It also addresses the critical need for appropriate disclosures and ethical considerations when integrating artificial intelligence (AI) tools into a scientific workflow. We review the current landscape of AI adoption in medical research, focusing on the role of ChatGPT in statistical analysis. Practical examples from lecture materials demonstrate its application in generating virtual datasets, performing data cleaning, conducting EDA, and assisting in the selection of appropriate statistical tests. Furthermore, guidelines for transparently disclosing AI tool usage in scientific manuscripts in accordance with the International Committee of Medical Journal Editors recommendations are discussed. ChatGPT demonstrates considerable potential for accelerating various stages of statistical analysis, from initial data preparation to the interpretation of results. Its ability to rapidly generate virtual data for practice, assist in comprehensive data cleaning, and provide immediate insights through EDA can substantially enhance research efficiency. Although capable of suggesting statistical methods and interpreting outputs, human intervention remains crucial for verifying assumptions and ensuring calculation accuracy. ChatGPT can serve as a powerful assistant in medical statistical analyses, enabling researchers to conduct analyses more efficiently. However, its use requires careful data preprocessing, human verification of results, and transparent reporting to maintain scientific rigor and reproducibility. Adherence to ethical guidelines and journal policies regarding AI tool disclosure is paramount.

Keywords: ChatGPT; Medical statistics; Statistical analysis; Artificial intelligence

Introduction

The landscape of medical research is rapidly evolving with the integration of artificial intelligence (AI) tools, particularly large language models, such as ChatGPT. Recent data indicate a significant increase in AI adoption among physicians, from 38% in 2020 to 66% in 2024 (Fig. 1) [1,2].

Statistical analysis is a cornerstone of medical research and provides a foundation for evidence-based conclusions. However, it often involves meticulous data management, selection of appropriate statistical methods, and precise interpretation of results, which can be time-consuming and require specialized expertise.

Despite its growing adoption, the use of ChatGPT for statistical analyses in medical research presents promising opportunities and challenges. ChatGPT can efficiently perform data processing tasks, such as variable reclassification, data subsetting, and tabulation, and generate descriptive statistics. Its ability to execute inferential statistical tests has been

validated, particularly when researchers provide clear and specific prompts [3].

Studies have highlighted the limitations of ChatGPT's performance, particularly when prompts lack specificity or complex statistical reasoning is required. Furthermore, ChatGPT may omit the critical assumptions necessary for certain statistical tests or provide divergent answers to the same

Received: 2025.07.11. Accepted: 2025.07.24.

Corresponding author: Soo-Nyung Kim, MD, PhD

Department of Obstetrics and Gynecology, Konkuk University School of Medicine, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea

E-mail: dr.snkim@gmail.com

<https://orcid.org/0000-0003-4319-336x>

Articles published in Obstet Gynecol Sci are open-access, distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © 2025 Korean Society of Obstetrics and Gynecology

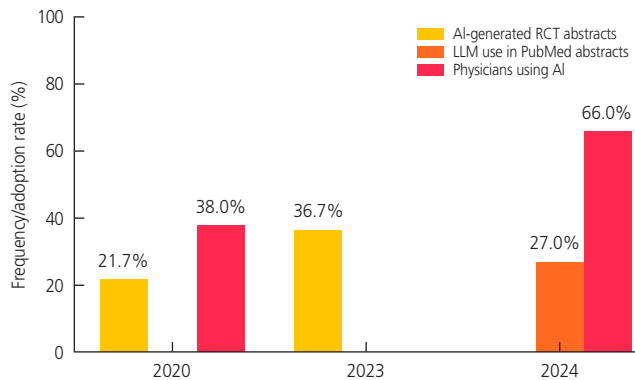


Fig. 1. Frequency of AI use in medical research. Bar graph showing the trends in AI adoption across three domains: AI-generated RCT abstracts, LLM usage in PubMed abstracts, and AI adoption by physicians. Data indicate a steep increase in physician use of AI in 2024, reaching 66%. AI, artificial intelligence; RCT, randomized clinical trial; LLM, large language model.

question, which can lead to inappropriate test selection and misinterpretation of the research findings [4,5].

This study aims to demonstrate how ChatGPT can assist researchers in various stages of statistical analysis, from data handling to the interpretation of results. Furthermore, it emphasizes the need for transparent reporting and adherence to ethical standards when incorporating AI into medical statistical workflows, following established guidelines, such as those of the International Committee of Medical Journal Editors (ICMJE) [6].

How ChatGPT can be used for statistical analysis

ChatGPT can be a powerful tool for statistical analysis, lowering the barrier to entry in data analytics and enabling non-experts to uncover patterns and insights in data [7,8].

1. Upload and initial exploration

Data files (such as CSV and Excel) or input data can be uploaded directly for analysis. ChatGPT can summarize data structure, missing values, variable types, and more.

2. Descriptive statistics and summarization

Provides basic statistics, such as mean, standard deviation, and min/max. Can compare key variables across categories (including by sex or age group).

3. Correlation and regression analysis

Analyzes relations between variables and can perform regression modeling. Automatically handles data preprocessing, such as one-hot encoding for categorical variables.

4. Data visualization

Explains visualization results or provides code for generating graphs and charts.

Practical example

1. Data management and preprocessing

Data quality is paramount for accurate statistical analyses. ChatGPT (GPT-4; OpenAI, San Francisco, CA, USA) has been explored for its ability to assist in data management and preprocessing, which often involves the handling of messy datasets.

1) Virtual data generation

ChatGPT can generate virtual medical datasets with specified field names and record counts, which are useful for performing statistical analyses or creating mock data for demonstrations. For instance, a prompt requesting 50 records with variables, such as 'SEX', 'AGE', 'HEIGHT', 'WEIGHT', 'total cholesterol (TC)', 'hypertension (HTN)', 'smoking (SMOKE)', 'EXERCISE', and 'coronary heart disease (CHD)' can result in a downloadable Excel file (Fig. 2).

2) Data cleaning

ChatGPT was utilized to clean simulated "bad data" spreadsheets. This involved. 1) Variable renaming: correcting variable names to adhere to standard conventions (including no spaces, no special characters except underscores, and lengths up to 10 characters), transforming "Tx-Dosage (mL)" to "tx_dosage_ml" and "Cured?" to "cured". 2) Handling missing values and inconsistent data: replacing specific values (including AGE=0, SEX=UNKNOWN, cured values other than "yes"/"no") with blanks to ensure data consistency. And 3) outlier detection and treatment: employing ChatGPT to identify outliers in numerical variables (including WEIGHT=150 kg was identified as a potential outlier) and to suggest methods for their treatment (removal, mean/median imputation, and transformation). Outliers were removed from the dataset for demonstration purposes.

	A	B	C	D	E	F	G	H	I
1	CHD	SEX	AGE	HEIGHT	WEIGHT	TC	HTN	SMOKE	EXERCISE
2	no	M	76	165	64	200	yes	no	none
3	no	F	73	167	64	208	no	yes	none
4	no	M	32	173	75	267	yes	yes	high
5	no	M	66	175	91	280	no	no	none
6	no	M	36	174	62	174	no	no	high
7	no	F	50	178	81	180	no	no	low
8	no	M	38	164	68	220	no	yes	high
9	no	M	68	150	54	161	no	yes	none
10	yes	M	47	165	85	300	yes	yes	low
11	no	F	33	156	53	160	no	no	high
12	no	M	54	175	79	280	no	yes	low
13	no	M	43	173	86	262	yes	no	low
14	yes	M	79	180	85	250	yes	no	none
15	no	M	38	167	65	210	no	yes	high
16	no	F	55	173	62	180	yes	no	low
17	no	M	31	160	64	170	no	no	high
18	no	F	49	160	55	151	no	no	low
19	yes	F	57	157	78	155	yes	yes	low
20	yes	F	76	175	80	203	yes	yes	none
21	no	M	36	184	90	236	yes	yes	high
22	no	F	73	166	63	278	no	yes	none
23	yes	M	37	188	85	296	yes	yes	high
24	no	F	76	158	71	200	yes	yes	none
25	no	F	64	177	77	230	yes	yes	none
26	no	F	43	156	51	175	yes	yes	low
27	no	F	46	158	54	167	no	yes	low
28	yes	F	65	157	60	172	no	yes	none
29	yes	F	79	161	70	194	no	yes	none
30	yes	F	69	175	77	287	yes	no	none

Fig. 2. Virtual medical data created by ChatGPT. A sample table of virtual records showing variables, such as sex (SEX), age (AGE), height (HEIGHT), weight (WEIGHT), cholesterol (TC), hypertension (HTN), smoking (SMOKE), exercise (EXERCISE), and coronary heart disease (CHD).

2. Exploratory data analysis (EDA)

The capabilities of ChatGPT in performing initial data exploration were assessed. This included generating summaries and visualizations of both categorical and numerical variables from a sample CHD dataset (Fig. 3).

EDA results provided insights into. 1) Basic information: number of samples and variables; presence of missing values. 2) Categorical data distribution: frequencies or percentages of variables, such as CHD, HTN, SMOKE, SEX, and EXERCISE. 3) Numerical variable distribution: descriptive statistics (mean, min and max) and visualizations (histograms) for AGE, HEIGHT, WEIGHT, and TC. And 4) correlation analysis: correlation matrices (heatmaps) were generated to visualize the relations between the numerical variables (Fig. 4).

3. Statistical analysis method selection and interpretation

The role of ChatGPT in guiding the selection of appropriate statistical tests and interpretation of results was examined based on the type of data, number of samples, and the purpose of analysis.

This included. 1) Descriptive statistics: summarizing numerical variables (mean and standard deviation for normal distributions; median and interquartile range for non-normal distributions) and categorical variables (cross-tabulation). 2) Comparative studies: suggesting tests, such as the *t*-test, analysis of variance (ANOVA), Wilcoxon rank-sum test, and chi-square test, based on data characteristics (including normal vs. non-normal distribution; paired vs. unpaired samples). 3) Relation analysis: recommending correlation or regression analyses. And 4) interpretation of statistical results: providing explanations for *P*-values and their implications—for example, for a *t*-test comparing numerical variables by CHD status.

Results

1. Efficiency in data management and preprocessing

ChatGPT proved to be a valuable assistant in initial data management tasks. Its ability to generate virtual datasets quickly provides a convenient resource for testing analytical approaches without requiring real patient data. The interactive nature of ChatGPT facilitates efficient data cleaning, including the systematic renaming of variables to conform to statistical software requirements and standardized handling of missing values and inconsistent string data. The outlier detection functionality, followed by a discussion of treatment options, highlights ChatGPT's capacity to guide researchers through critical data preparation steps. This significantly reduces the manual effort and time required for the initial phases.

2. Insights from EDA

The EDA capabilities of ChatGPT provide immediate and comprehensive summaries of categorical and numerical data distributions. The generated tables for categorical variables (including frequencies of CHD, HTN, SMOKE, SEX, and EXERCISE) and histograms for numerical variables (AGE, HEIGHT, WEIGHT, and TC) offer quick visual and quantitative insights into the characteristics of the dataset. The correlation matrix, presented as a heatmap, clearly illustrates the relations between the numerical variables (including a weak positive correlation between AGE and TC, and a relatively strong positive correlation between WEIGHT and HEIGHT). These rapid insights are crucial for understanding data patterns and informing subsequent, more complex statistical analyses.

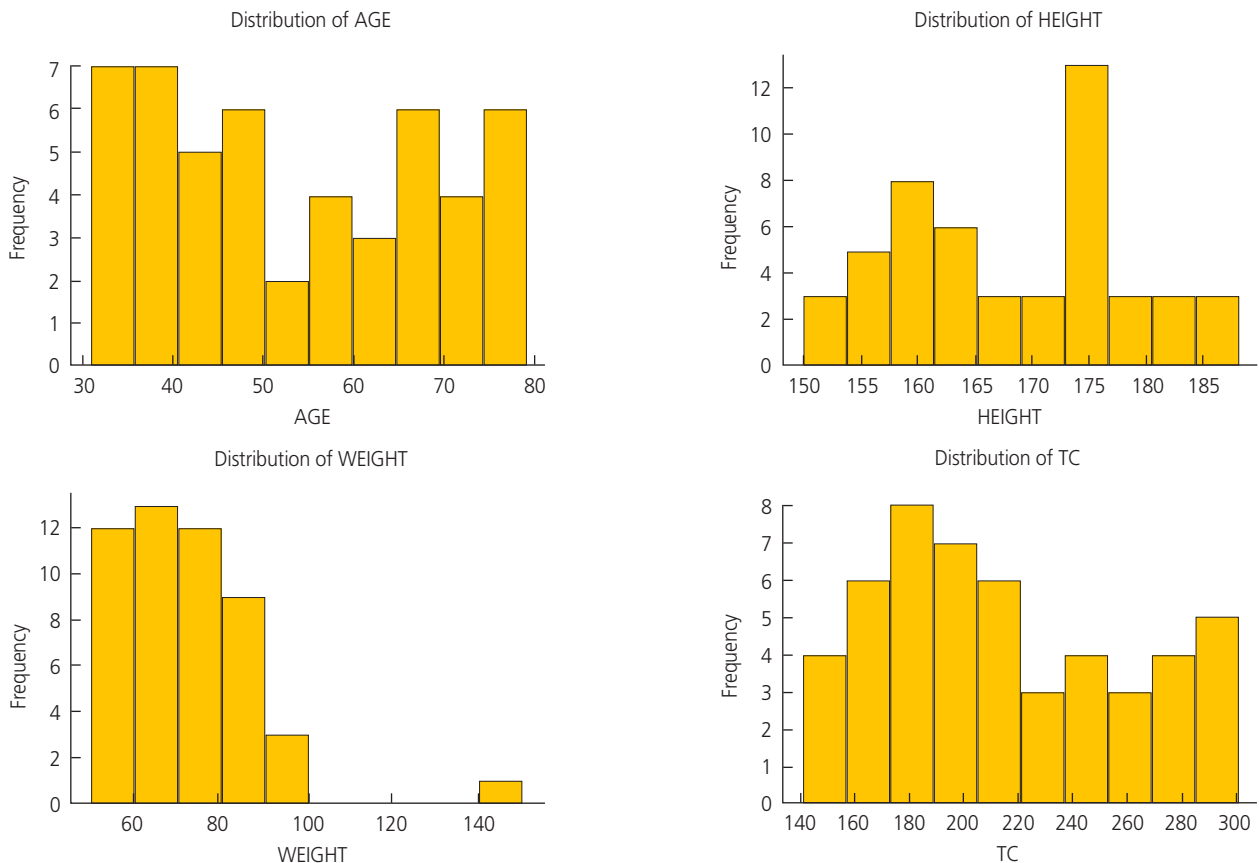


Fig. 3. Histograms of numeric variables. Distribution plots of AGE, HEIGHT, WEIGHT, and TC in the dataset. The distributions show mild skewness in WEIGHT and TC, whereas AGE and HEIGHT are more evenly spread.

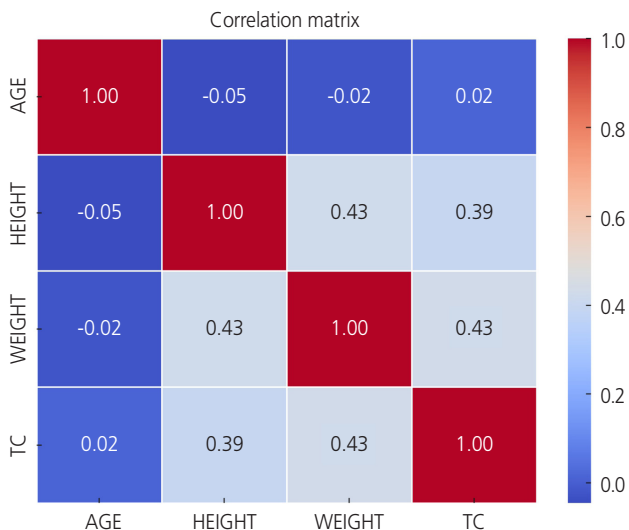


Fig. 4. Heatmap of numeric variables. Pearson correlation coefficients between numerical variables. HEIGHT and WEIGHT ($r=0.43$), as well as WEIGHT and TC ($r=0.43$), show moderate positive correlations, whereas AGE shows minimal correlation with other variables.

3. Guidance in statistical analysis and interpretation

ChatGPT demonstrates proficiency in recommending appropriate statistical tests based on the researcher's objectives and data types. For descriptive statistics, it correctly advised using mean standard deviation for normally distributed numerical variables and median (interquartile range) for non-normally distributed ones, along with cross-tabulations for categorical variables. When asked for a comparative analysis, tests, such as the t -test, ANOVA, or non-parametric alternatives, were suggested. For instance, in a t -test comparing numerical variables according to CHD status, ChatGPT accurately reported P -values and concluded that no statistically significant difference was found when $P>0.05$. The ability to interpret statistical results, including the significance of P -values, further enhances its utility as a support tool.

4. Limitations of ChatGPT

1) Inconsistent accuracy and reliability

ChatGPT's performance in statistical analysis can be inconsistent, especially with inferential statistics. Its accuracy is highly dependent on the specificity and clarity of user prompts. For example, one study found that the accuracy of inferential statistics generated by ChatGPT ranged from 32.5% with basic prompts to over 90% with advanced, detailed prompts [4]. This variability means that the results may not always be reliable unless carefully validated.

2) Lack of critical statistical reasoning

ChatGPT often fails to consider or mention the key statistical assumptions required for certain tests (including normality and homogeneity of variance). This omission can lead to inappropriate test selection and potentially invalid conclusions, as the model does not inherently check whether the data meet the prerequisites for specific analyses [9].

3) Potential for divergent or contradictory answers

The same question posed to ChatGPT multiple times may yield different answers, especially if prompts are ambiguous or lack context. Such inconsistencies can undermine confidence in the reproducibility and validity of the analysis.

4) Limited handling of complex or nuanced analyses

While ChatGPT can manage basic descriptive and inferential statistics, it struggles with complex statistical modeling, advanced multivariate analyses, and situations requiring deep domain-specific interpretation. It is not a substitute for expert statistical judgment or advanced analytical skills.

5) Dependence on prompt quality and user expertise

Effective use of ChatGPT for statistical analysis requires users to craft clear, specific, and context-rich prompts. Users with limited statistical or programming knowledge may not know how to provide such prompts, which may lead to suboptimal or incorrect results.

6) Risk of overreliance and misinterpretation

Researchers may overrely on ChatGPT's outputs without adequate verification, potentially leading to the misinterpretation of findings. The model's authoritative tone can provide a false sense of confidence in its recommendations, even when they are incorrect or incomplete.

7) Data privacy and security concerns

Using ChatGPT, particularly cloud-based versions, may raise concerns about the privacy and security of sensitive medical data, as data entered into the model can be stored or processed externally.

ICMJE conventions for AI statistical analysis in medical research

The ICMJE has established clear and evolving guidelines regarding the use of AI, including tools, such as ChatGPT, for statistical analysis and other research tasks in medical publications [6].

The ICMJE requires full transparency, human oversight, and ethical rigor when AI is used for statistical analysis in medical research. All AI involvement must be disclosed and validated and humans must retain full responsibility for the scientific integrity of the study (Table 1).

Table 1. ICMJE AI statistical analysis conventions

Area	ICMJE guidance
Disclosure	Mandatory in methods (analysis/figures) and acknowledgements (writing/editing)
Authorship	AI cannot be listed as author; only humans are accountable
Verification	Manual review of AI-generated analysis and references required
Data transparency	Data availability statements and openness to independent review
Prohibited AI uses	No data manipulation, image generation, or result fabrication by AI
Ethical standards	Uphold integrity, transparency, and proper citation of all AI-assisted content

Summary of the International Committee of Medical Journal Editors (ICMJE) guidance regarding the ethical and appropriate use of artificial intelligence (AI) in statistical analysis and scientific writing. Key areas include mandatory disclosure, authorship restrictions, verification requirements, transparency in data reporting, prohibited uses of AI (including data manipulation or fabrication), and adherence to ethical standards.

Conclusion

ChatGPT is a powerful tool for statistical analysis because it leverages AI and language models. It can provide accurate results, generate statistical code, and aid in understanding statistical concepts. While ChatGPT can suggest appropriate statistical tests and provide interpretations, it lacks the nuanced understanding of human experts regarding the underlying assumptions, data context, and clinical implications of the findings. By using ChatGPT to effectively and critically evaluate responses, users can benefit from its capabilities while supplementing its outputs with additional sources for comprehensive analysis.

In summary, while ChatGPT offers valuable support for statistical analysis in medical research, its limitations underscore the need for careful human oversight, expert validation, and a cautious approach to integrating AI-generated results into scientific studies. Therefore, ChatGPT should be viewed as an assistive tool rather than a replacement for human statisticians.

Conflict of interest

No conflict of interest.

Ethical approval

This study used virtual datasets created by ChatGPT and did not require institutional review board approval.

Patient consent

Not applicable.

Funding information

No external funding was received.

Acknowledgments

I thank the ChatGPT for support in dataset preparation.

References

1. Littrell A. New AMA report highlights physician optimism about AI in health care. *Medical Economics* 2025 Feb 19 [Epub]. <https://www.medicaleconomics.com/view/new-ama-report-highlights-physician-optimism-about-ai-in-health-care>.
2. Ehrenfeld J. Physician leadership in the new era of AI and digital health tools. *Intelligent Medicine* 2023;8:100109.
3. Miller LE, Bhattacharyya D, Miller VM, Bhattacharyya M. Recent trend in artificial intelligence-assisted biomedical publishing: a quantitative bibliometric analysis. *Cureus* 2023;15:e39224.
4. Ruta MR, Gaidici T, Irwin C, Lifshitz J. ChatGPT for univariate statistics: validation of AI-assisted data analysis in healthcare research. *J Med Internet Res* 2025;27:e63550.
5. Ignjatović A, Stevanović L. Efficacy and limitations of ChatGPT as a biostatistical problem-solving tool in medical education in Serbia: a descriptive study. *J Educ Eval Health Prof* 2023;20:28.
6. International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing and publication of scholarly work in medical journals (revised in January 2024): a Korean translation. *Ewha Med J* 2024;47:e48.
7. Huang Y, Wu R, He J, Xiang Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: a comparative analysis with SAS, SPSS, and R. *J Glob Health* 2024;14:04070.
8. Conte L, Lupo R, Lezzi P, Pedone A, Rubbi I, Lezzi A, et al. Statistical analysis and generative artificial intelligence (AI) for assessing pain experience, pain-induced disability, and quality of life in Parkinson's disease patients. *Brain Res Bull* 2024;208:110893.
9. Ordak M. ChatGPT's skills in statistical analysis using the example of allergology: do we have reason for concern? *Healthcare (Basel)* 2023;11:2554.