DOI: 10.4081/aiua.2025.13596

ORIGINAL PAPER

ChatGPT artificial intelligence in clinical data analysis: An example comparing standard vs fusion prostate biopsy outcomes after robotic-assisted radical prostatectomy (RaRP)

Pier Paolo Prontera [1], Francesca Romana Prusciano [1, 2], Marco Lattarulo [1], Arman Tsaturyan [3, 4], Francesco Addabbo [5], Carmine Sciorio [6], Francesco Saverio Grossi [1]

[1] Department of Urology, "S.S. Annunziata" Hospital, Taranto, Italy;
[2] Division of Urology, Hospital "Valle D'Itria", Martina Franca (TA), Italy;
[3] Department of Urology, Yerevan State Medical University after Mkhitar, Heratsi, Yerevan, Armenia;
[4] Department of Urology Erebouni Medical Center, Yerevan, Armenia;
[5] Unit of Statistics and Epidemiology, Local Health Authority of Taranto, Taranto, Italy;
[6] Department of Urology, "Alessandro Manzoni" Hospital, Lecco, Italy.

Summary *Objective: To compare statistical outputs from ChatGPT 4.0 and human experts in both comparative and correlation analyses in the evaluation of multiparametric MRI/ultrasound fusion-targeted biopsy plus random biopsy versus standard random biopsy alone, in terms of upstaging. Methods: Authors performed a retrospective evaluation on 101 patients undergoing robot-assisted radical prostatectomy (RaRP) between 2021 and 2023. Patients were divided in two groups, according to the type of prostatic biopsy received: combined fusion (MRI/US) targeted and random biopsy versus standard random biopsy. Clinical and histological data were anonymized and analyzed using logistic regression models, ANOVA, and Chi-square tests. Analysis generated by ChatGPT and by an experienced human statistician were compared. The Q-EVAL and Q-EVA tools were used to assess the quality of user-formulated questions and AI-generated answers, respectively.Results: Results revealed high concordance between statistical outputs generated by AI and expert human statistician with perfect concordance using Cohen's kappa coefficient (κ = 1.0). Logistic regression analysis demonstrated that fusion biopsy was associated with a reduced likelihood of upstaging, a consistent finding across statistical evaluations. Additionally, user interaction assessments indicated high-quality in question formulation. Conclusions: ChatGPT (version 4.0) proved reliable for statistical analysis, showing strong concordance with human statisticians (κ = 1.0) in performing logistic regression, chi-Square, and ANOVA tests. The Q-EVAL tool could reduce query errors, though ChatGPT's lack of automatic citations remains a limitation. Fusion biopsy significantly lowered upstaging risk after RaRP. In conclusion, ChatGPT is a valuable assistive tool but further research is required to optimize human-AI collaboration in clinical research.*

KEY WORDS: *Prostate cancer; ChatGPT; Artificial intelligence; Upstaging; Robot-assisted radical prostatectomy; Fusion biopsy.*

Submitted 7 January 2025; Accepted 29 March 2025

INTRODUCTION

The rapid advancement of *artificial intelligence* (AI) technologies has led to the integration of machine learning models across numerous domains of medicine, including clinical data analysis, diagnostic support, and treatment planning. Among these models, OpenAI's *Chat Generative Pre-trained Transformer* (ChatGPT) has emerged as a leading *large language model* (LLM), capable of generating human-like responses based on extensive textual data training.

Although primarily designed for natural language processing tasks, ChatGPT's potential extends beyond its original scope, raising the question of its applicability in clinical data analysis and decision-making processes (1-3). Recent advancements in AI-driven models such as ChatGPT have demonstrated potential in supporting systematic reviews and meta-analyses by automating literature screening, data extraction, and bias assessment processes, thus streamlining evidence synthesis in clinical research (4).

In clinical practice, analyzing patient data to identify prognostic and predictive factors is critical for improving therapeutic strategies and outcomes. Traditional statistical methods such as logistic regression, survival analysis, and multivariable modeling have been the gold standard for data interpretation. However, these techniques often require specialized statistical expertise and are susceptible to human error during data management, analysis, and interpretation (5).

Beyond its analytical potential, ChatGPT appears oòito be promising in automating repetitive, standardized, and compilative tasks, reducing the time and effort required in such activities. For instance, a study by Aykut Demirci in 2024 compared the performance of ChatGPT with humans in completing validated quality assessment questionnaires (6), including the DISCERN-5 and *Global Quality Scale* (GQS), which have also been used in the past by other authors for similar analyses but without the use of AI (7), for evaluating audio-video materials on YouTube. The findings revealed no statistically significant differences in median scores (IQR) for both DISCERN-5 and GQS when comparing ChatGPT to human assessments, suggesting that ChatGPT can deliver comparable results in such tasks while enhancing efficiency (8).

Postoperative management following *robot-assisted radical prostatectomy* (RaRP) presents a complex clinical challenge, particularly regarding upstaging (identification of more advanced cancer than initially assessed) and upgrading (identification of more aggressive cancer types than initially expected). These factors significantly affect postoperative management, including the need for adjuvant therapy and closer follow-up (3). The clinical management of prostate cancer has seen notable advances in biopsy techniques, specifically the comparison between standard *systematic biopsy* (SBx) and multiparametric MRI/ultrasound fusion-*targeted biopsy* (TBx). Systematic biopsy has historically been the standard method, offering spatially distributed sampling of the prostate gland. However, SBx has notable limitations, including, higher rates of false-negative, frequent under detection of aggressive tumors and over detection hof clinically insignificant cancers, potentially leading to overtreatment (9). Recent studies have highlighted a persistent risk of diagnostic misclassification in low-risk prostate cancer patients, with 10.5% experiencing pathological upstaging at radical prostatectomy, including 6.3% with *Gleason Grade Group* (GGG) 2 and 1.6% with GGG ≥ 3. Studies have consistently shown that TBx improves the detection of clinically significant cancers while reducing unnecessary diagnoses of indolent tumors (10). Research comparing these two methods indicates that TBx more accurately identifies prostate cancer's Gleason scores, leading to lower rates of undergrading and overgrading when compared to SBx alone (11). Additionally, the combination of SBx and TBx provides the highest cancer detection rates, supporting a multimodal approach to minimize the risk of misclassification. Studies highlight the significant upstaging and upgrading observed when combining these methods, which is essential for surgical planning and long-term patient management (12). The increased accuracy in detecting more aggressive lesions through TBx has been particularly evident in identifying high-risk prostate cancer types with higher Gleason scores and larger tumor volumes (13).

Given this background, the expanding role of AI in medical research and its increasing accessibility, this study aims to explore whether ChatGPT can support clinical data analysis in Urological field.

Materials and methods

A retrospective evaluation was conducted on 153 patients who underwent *robot-assisted radical prostatectomy* (RaRP) between 2021 and 2023, in a single center. 101 patients were enrolled in the study based on specific inclusion and exclusion criteria. Patients with previous negative prostate biopsies or histological diagnoses performed at external institutions were excluded. All enrolled patients underwent robot-assisted radical prostatectomy using the DaVinci Xi (Intuitive) multiport robotic system.

Techniques and histopathology

Prostate biopsy

Histological diagnosis was performed using either standard trans-perineal prostate biopsy (SBx) or MRI/US fusion combined trans-perineal biopsy (TBx+SBx); both procedures were performed under transrectal ultrasound guidance. Fifty patients underwent a combined fusion biopsy, while 51 patients received a standard random biopsy. All procedures were performed by a highly experienced operator. Fusion biopsy was indicated for patients presenting with clinically significant PIRADS lesions (PIRADS ≥ 3) and included both targeted cores (based on the number and size of *regions of interest* (ROI)) and 12 to 16 systematic cores according a prostatic template including base, mid-gland, and apex on both sides. When multiparametric MRI (mpMRI) was unavailable or no significant PIRADS lesions were detected, a standard 16-core TRUS guided trans-perineal biopsy was performed following the institutional template, sampling the base, mid-gland, apex, and transition zones bilaterally.

Staging and surgery

All patients underwent staging with contrast-enhanced total-body CT scan and total-body bone scintigraphy scan. Subsequently, they underwent *robot-assisted radical prostatectomy* (RaRP) at the same center, performed by two different surgeons. Forty-seven procedures were conducted by one surgeon and fifty-four by the other, both highly experienced in robotic prostate cancer surgery. The surgical specimens were examined by an expert pathologist.

Statistical analysis

Data collection

Clinical and laboratory data related to the sample of 101 patients enrolled in the study were collected by two urologists and recorded in a database using Microsoft Excel (version 2013).

The database comprehensively collects data on patients undergoing *robot-assisted radical prostatectomy* (RaRP). It includes demographic information such as patient age, weight, height, BMI, and waist circumference.

Preoperative clinical parameters include family history of prostate cancer (first- or second-degree relatives), preoperative PSA levels, clinical staging, biopsy Gleason score, and EAU 2024 risk classification. Intraoperative variables include the surgeon's experience, operative time (minutes), blood loss, intraoperative complications, and nerve-sparing approach. Postoperative outcomes cover hospitalization duration, final pathological staging, postoperative Gleason score, lymph node involvement, surgical margins, and recalculated EAU 2024 risk classification based on definitive histology. Functional outcomes include continence status, calculated by using a continence score based on three levels (2 - full continence, 1 - stress incontinence and 0 - complete incontinence), and by evaluating the variation in the number of pads used per day at 3, 6, 9 and 12 months, erectile function recovery (IIEF score at 3, 6, 9 and 12 months), and PSA levels at 3, 6, 9 and 12 months. The database also records perioperative complications, reintervention rates, oncological outcomes (biochemical recurrence and survival), as well as patient-reported satisfaction scores, including ratings for the hospital facility and nursing staff. Additionally, the geographical distance traveled by patients to reach the hospital is documented.

The Excel sheet was designed to use a binary coding system exclusively ("0": negative variable, "1": positive variable). Column nomenclature was simplified as much as possible. Both urologists independently reviewed the database in order to avoid potential structural and errors. To ensure patient confidentiality, all data were fully anonymized before analysis, with no personally identifiable information retained in the dataset.

The dataset included patient records with relevant clinical variables, such as biopsy type (standard = 0, fusion = 1) and upstaging outcome (no = 0, yes = 1). The dataset was sufficiently large to ensure robust regression model fitting.

Prostate volume data were collected to categorize patients into three preoperative risk groups: low, intermediate, and high. The dataset included a sufficient sample size for each group to ensure the validity of the ANOVA assumptions.

The dataset was used to conduct statistical analyses independently by: 1) an expert human statistician, 2) a user with limited statistical knowledge utilizing ChatGPT (User A), 3) a user with limited statistical knowledge utilizing ChatGPT (User B).

Evaluation criteria for questions and answers

Both users (User A and User B) independently formulated questions to the AI. Questions formulated by both users A and B and answers provided by the ChatGPT platform were subsequently evaluated using two non-validated tools specifically developed by the authors of the study: the "Q-EVAL" tool (*Quality Evaluation for Verification and Assessment of Language-based Queries*) and the "Q-EVA" tool (*Quality Evaluation of Answers*).

Both tools assess a total score (ranging from 0 to 100) for each question formulated by the user and to each answer generated by the AI. Both Q-EVAL and Q-EVA evaluates four items, each of them scored either 0 (if not met) or 25 (if met).

Q-EVAL assesses four key items related to the quality of user-formulated questions. The first criteria is minimum length, requiring questions to contain at least 10 characters. The second is specificity, emphasizing the use of action verbs to clearly define the intent of the query. The third criterion involves the absence of ambiguity, ensuring that unclear terms are avoided. Lastly, context presence is evaluated, requiring the use of complete sentences to provide clarity and context.

Q-EVA evaluates four parameters of AI-generated answers. The first criterion is the presence of citations, ensuring that sources are appropriately referenced. Internal consistency is the second aspect, requiring logically coherent responses structured with complete sentences. The third criterion focuses on the use of technical terms, verifying the inclusion of relevant keywords in the answer. Finally, the absence of opinion ensures that responses remain objective, avoiding personal judgments or subjective interpretations.

OpenAI's ChatGPT (User A and User B) analysis

Users A and B used ChatGPT (version 4.0), developed by OpenAI, under the intermediate subscription tier. This version supports advanced natural language processing, context-aware responses, and statistical query handling. Key features included extended context retention and improved reasoning capabilities, though limitations such as the lack of real-time database access and automatic source citation, were recognized.

Statistical analyses were conducted to evaluate the association between clinical variables and patient outcomes.

A logistic regression analysis was performed using the independent variable "FUSION" (biopsy type: 0 = standard, 1 = fusion) to predict the likelihood of upstaging (0 = no, 1 = yes). Additionally, a one-way ANOVA was conducted to compare prostate volumes across three preoperative risk classes, according to D'Amico Classification (14). Pairwise comparisons using independent samples t-tests were performed among the three risk classes (low *vs*. intermediate, low *vs.* high, and intermediate *vs*. high), assuming unequal variances.

All statistical analyses were performed using Python, employing libraries such as Pandas, SciPy, and scikit-learn. Data visualization and graphical outputs were generated using Matplotlib to provide clear and interpretable representations of the results.

The process of submitting data to ChatGPT for analysis follows a structured yet intuitive workflow. Users can upload data by either clicking the "+" button in the chat interface or simply dragging and dropping the file into the conversation. In this case, the file was uploaded by dragging and dropping it directly onto the search bar and was an Excel spreadsheet (.xlsx), but other commonly used formats such as CSV, JSON, and TXT are also supported.

To initiate an analysis, users provide instructions in natural language, in this case in Italian language, specifying the tasks they need, such as data cleaning, statistical analysis, or visualization. For example, a user typed, "*Analyze the patient dataset and provide a summary of the main variables*."

Context and specific objectives can also be outlined, making it easier for ChatGPT to tailor the analysis. A more detailed instruction was, "*This dataset contains information on patients undergoing prostatectomy. Please summarize key variables such as age, BMI, and PSA levels*."

Once instructions are provided, ChatGPT first cleans the data by handling missing values, renaming columns if necessary, and ensuring proper formatting. It then performs the requested analysis by interpreting the instructions and converting them into Python scripts, which are executed in a built-in Python environment. The results are then returned in user-friendly formats, such as tables, figures, and summaries.

For example, when a user asks, "*Show me the average BMI of patients in the dataset*", ChatGPT will read the Excel file, extract the relevant column, calculate the mean BMI using Python, and display the result either as a numerical output or as a graphical representation. The final results are presented in natural language, ensuring accessibility even for users without technical expertise. An example output could be: "*The average BMI of patients in the dataset is 27.5*", accompanied by a chart showing the BMI distribution.

This process makes data analysis efficient and accessible, allowing users to gain insights without needing programming knowledge while still leveraging the power of Python for advanced computations.

Experienced human statistician analysis

Statistical analyses were performed using the Chi-Squared test and Logistic Regression to evaluate the association between type of biopsy and upstaging.

The statistical computations were carried out using the online platform "Statistics Kingdom" (*https://www.statskingdom.com/ 180Anova1way.html*). The platform facilitated the implementation of the ANOVA test, the validation of assumptions using the Shapiro-Wilk test, and the assessment of homogeneity of variances using Levene's test. Effect size calculations, including the eta-squared (η^2) statistic, were also generated. Additionally, one-way analysis of variance (ANOVA) was conducted using the F distribution to determine whether significant differences existed between prostate volumes emerged among the evaluated risk classes, according to D'Amico classification (14). Post hoc comparisons were performed using Tukey's *Honest Significant Difference* (HSD) test to identify pairwise differences between group means.

Comparison of statistical analyses

To comprehensively evaluate the agreement between ChatGPT-driven analyses and those performed by a human statistician, a dual-approach methodology was applied. First, standard statistical metrics such as F-statistics, regression coefficients, p-values, and model fit indicators were computed for both ANOVA and logistic regression analyses. Absolute and percentage differences in these key parameters were calculated.

Secondly, to quantitatively assess the agreement between ChatGPT-generated results and human expert analyses, Cohen's kappa statistics were employed. This metric measures inter-rater reliability and accounts for agreement occurring by chance. The comparison focused on two key clinical outcomes: biopsy type classification and upstaging detection. For biopsy type, the variable 'FUSION' from the dataset was compared with AI-predicted biopsy classifications. For upstaging, preoperative and postoperative risk classifications were used to determine whether upstaging occurred, and these results were cross-referenced with AI-generated outputs. Cohen's kappa coefficients were computed for both variables, yielding values of 1.0 for both biopsy type and upstaging detection. Graphical representations were generated to highlight the results.

A p-value less than or equal to 0.05 was considered statistically significant.

Results

This retrospective study included 101 patients who underwent *robot-assisted radical prostatectomy* (RaRP) between 2021 and 2023, in a single center. None of the patients undergoing RaRP developed significant perioperative complications. The cohort had a mean age of 68.6 years, with a mean preoperative PSA level of 9.23 ng/mL and an average prostate volume of 54.4 ml. The distribution of patients according to preoperative risk class showed that 24.757% were classified as low risk, 48.51% as intermediate risk, and 26.73% as high risk. Regarding preoperative clinical stages, the most frequent stages were T2b (40.59%), T2a (24.75), T1c (17.82%), T2c (10.89%) and T3b (5.94) (Table 1).

The evaluation process revealed that both Users, A and B, consistently formulated well-defined questions, receiving perfect Q-EVAL scores of 100 points across all queries (Table 2).

In addition to statistical accuracy, ChatGPT's response quality was evaluated using Q-EVA scoring criteria. Despite the absence of automatically generated citations, ChatGPT's responses demonstrated high internal consistency, appropriate use of technical terminology, and objectivity, yielding a consistent Q-EVA score of 75 points (Table 3). This underscores its reliability as a research tool when complemented by human validation. However, the responses provided by ChatGPT received a consistent Q-EVA score of 75 points due to the lack of citations, despite excelling in internal consistency, technical accuracy, and objectivity. The combined use of these evaluation tools highlighted the effectiveness of ChatGPT as a support system for clinical data analysis.

The relationship between biopsy type and upstaging was assessed using a Logistic Regression and Chi-Square test, both indicated a statistically significant association (p-value < 0.05). Chi-Square tests shows the following findings: Chi-Square value 4.48 and p-value 0.034 for User

Clinical and pathological characteristics of enrolled patients undergoing RaRP

	Fusion-combined biopsy	Random biopsy	Tot.
Mean age, y (range)	67.7 (55-76)	69.5 (54-75)	68.6 (54-76)
Mean pre-operative total PSA, ng/ml (range)	8.46 (3.9-13.5)	10 (3.54-31.5)	9.23 (3.54-31.5)
Mean prostate volume, ml (range)	48.3 (20-142)	54.4 (25-120)	51.4 (20-142)
Pre op. staging, n° (%)			
T1c	12 (24)	6 (11.76)	18 (17.82)
T2a	15 (30)	10 (19.6)	25 (24.75)
T2b	16 (32)	25 (49)	41 (40.59)
T2c	4 (8)	7 (13.72)	11 (10.89)
T3b	3 (6)	3 (5.88)	6 (5.94)
Pre op. Risk classification (acc. D'Amico)			
Low, n° (%)	15 (30)	10 (19.6)	25 (24.75)
Intermediate n° (%)	22 (44)	27 (52.94)	49 (48.51)
High n° (%)	13 (26)	14 (27.45)	27 (26.73)

Table 1.
Patient demographics, pre-operative clinical features, and risk classifications according to D'Amico criteria are summarized. Mean age, total pre-operative PSA levels, and prostate volume are presented with ranges. Pre-operative staging is reported in absolute numbers and percentages. Risk classification includes low, intermediate, and high-risk categories based on clinical assessment prior to surgery. The distribution of these variables highlights differences between patients undergoing fusion biopsy and those managed through standard random biopsy.

Table 2.
Quality assessment of user-formulated questions using the Q-EVAL tools.

		User A	User B
Question 1:	Length (at least 10 characters)	25	25
Assess of statistically significant differences in upstaging between patients undergoing	Specificity (use of action verbs)	25	25
standard biopsy and those undergoing fusion biopsy	Un-ambiguous (absence of vague or unclear terms)	25	25
	Clarity and context	25	25
Question 2:	Length (at least 10 characters)	25	25
Correlation between prostate volume and preoperative risk categories	Specificity (use of action verbs)	25	25
	Un-ambiguous (absence of vague or unclear terms)	25	25
	clarity and context	25	25

Table 3.
Quality assessment of ai-generated answers using the Q-EVA tools.

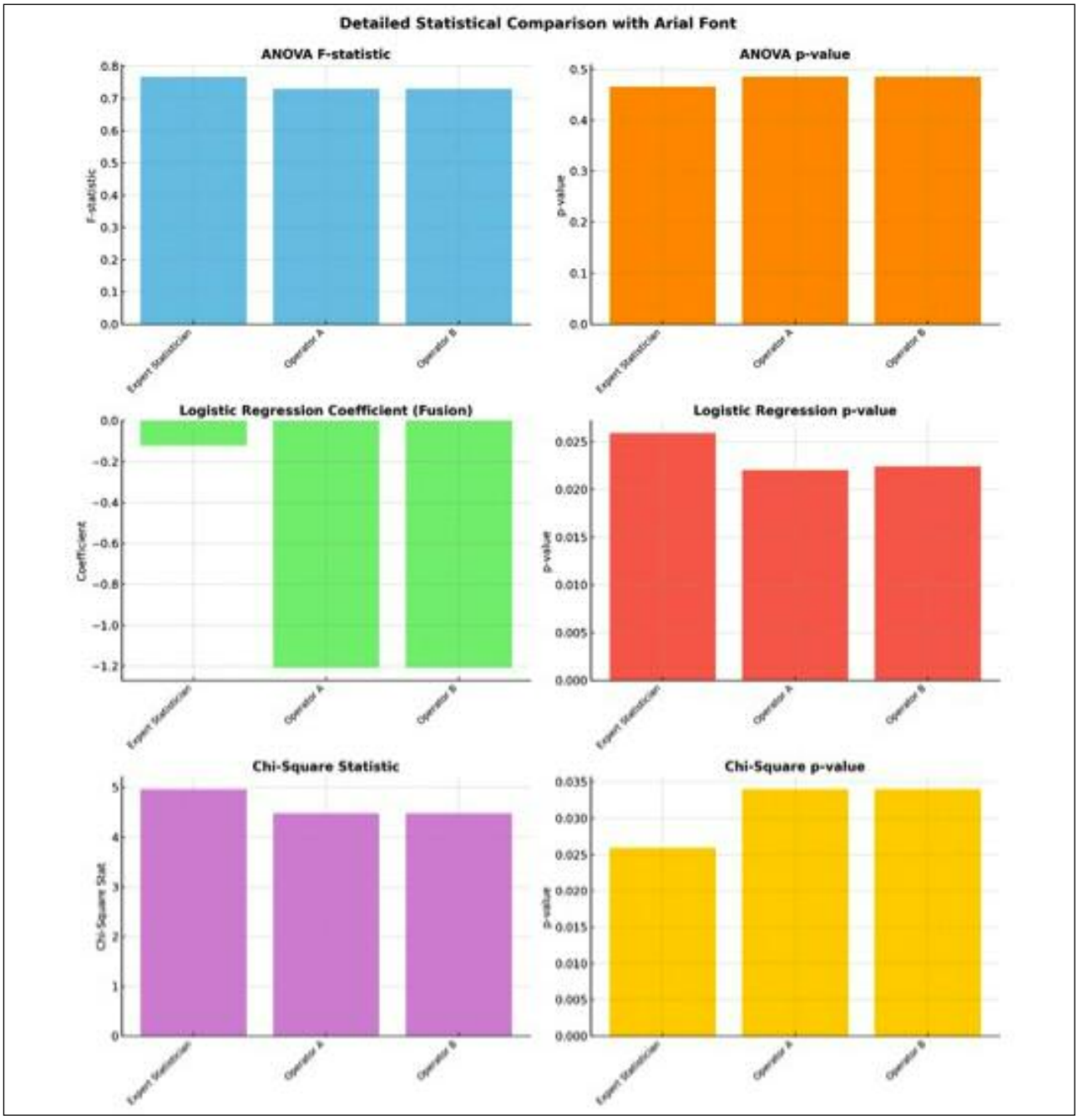
		User A	User B
Question 1:	Presence of citations	0	0
Assess of statistically significant differences in upstaging between patients undergoing	Internal consistency (logical coherence)	25	25
standard biopsy and those undergoing fusion biopsy	Use of technical terms	25	25
	Absence of opinion	25	25
Answer 2:	Presence of citations	0	0
Correlation between prostate volume and preoperative risk categories	Internal consistency (logical coherence)	25	25
	Use of technical terms	25	25
	Absence of opinion	25	25

A; Chi-Square 4.48 and p-value 0.034 for Users B. Logistic regression analysis confirmed this finding: intercept -0.783 (p-value 0.009) and Fusion biopsy coefficient -1.210 (p-value 0.022) for User A; Intercept -0.78 (p-value 0.0095) and Fusion biopsy coefficient -1.21 (p-value 0.0224) for User B. Both users obtained consistent results indicating a significant relationship between biopsy type and the likelihood of upstaging, with fusion biopsy being associated with a lower probability of upstaging compared to standard biopsy. The one-way ANOVA comparing prostate volumes across preoperative risk classes (low, intermediate, high) revealed no statistically significant differences (F-value = 0.73 and p-value = 0.485 for User A; F-value 0.7294 anche p-value 0.4848 for User B). Post-hoc Tukey HSD comparisons confirmed the lack of significance, with all pairwise comparisons yielding p-values greater than 0.05 for both analysis, with no statistically significant differences in prostate volumes among the evaluated risk classes.

Additionally, an experienced human statistician performed statistical analysis, using the same dataset. To evaluate the association between type of biopsy and upstaging, Statistical analyses were conducted using the Chi-Squared test and Logistic Regression. Chi-Square test shows the following results: Chi-Square Stat 4.9629, p-value 0.0259, freedom degrease 1, Phi 0.22167 and Cramer's V 0.22167. From Logistic Regression emerged a correlation coefficient -0.121. Overall, both statistical tests reinforce the presence of a meaningful relationship, with the Chi-Square test highlighting statistical significance and the logistic regression supporting the clinical relevance of the association. One-way analysis of variance (ANOVA) was conducted comparing prostate volumes across preoperative risk classes (low, intermediate, high) revealed no statistically significant differences (F-statistic 0.7673, p-value 0.4656, Effect Size 0.088).

A detailed comparison was conducted between the results obtained by a human statistician and those generated by Operators A and B, independently, using ChatGPT 4.0. A dual-approach methodology was applied to compare statistical results from an expert statistician, Operator A, and Operator B. Standard statistical metrics were computed for ANOVA, logistic regression, and Chi-Square analyses (Figure 1). For ANOVA Analysis, F-statistic values were 0.7673 (expert), 0.73 (Operator A), and 0.73 (Operator B), with p-values of 0.4656, 0.485, and 0.485, respectively, indicating no statistically significant differences in prostate volume across risk classes. For Logistic Regression, Regression coefficients for biopsy type ("Fusion") were -0.121 (expert) versus -1.210 (Operators A and B), with p-values of 0.0259 (expert) and 0.022/0.0224 (Operators A/B). These differences suggest methodological variations in model specification. For Chi-Square Test, Chi-Square statistics were 4.9629 (expert) versus 4.48 (Operators A/B), with p-values of 0.0259 (expert) and 0.034 (Operators A/B), confirming similar statistical significance despite slight numerical discrepancies (Table 4). Agreement was assessed using Cohen's Kappa, yielding perfect scores of 1.0 for both biopsy type classification and upstaging detection, indicating full concordance between human and AI-driven analyses (Figure 2). These results suggest high reliability and comparable accuracy across all tested methods. These comparisons underscore the high concordance between AI-supported analyses and expert-driven statistical evaluations. Despite different methodological

Figure 1
Detailed Statistical Comparison of ANOVA, Logistic Regression, and Chi-Square Tests: Expert Statistician vs. Operator A vs. Operator B: The chart presents a comparative analysis of key statistical metrics obtained from analyses conducted by the expert statistician, Operator A, and Operator B. Metrics include F-statistics and p-values for ANOVA, regression coefficients and p-values for logistic regression, and Chi-Square statistics with corresponding p-values. The comparison highlights areas of agreement and methodological differences across the three analytical approaches.



Metric	Expert statician	User A	User B
ANOVA F-statistic	0.7673	0.73	0.73
ANOVA p-value	0.4656	0.485	0.485
Logistic regression coefficient (Fusion)	-0.121	-1.21	-1.21
Logistic regression p-value (Fusion)	0.0259	0.022	0.0224
Chi-Square statistic	4.9629	4.48	4.48
Chi-Square p-value	0.0259	0.034	0.034

Table 4.
Detailed Statistical Comparison of ANOVA, Logistic Regression, and Chi-Square Tests: Expert Statistician vs. User A vs. User B: The table presents a detailed comparison of key statistical metrics, including F-statistics, p-values, regression coefficients, and Chi-Square statistics. Results from the expert statistician, Operator A, and Operator B are presented, highlighting areas of agreement and statistical differences observed during the analyses.

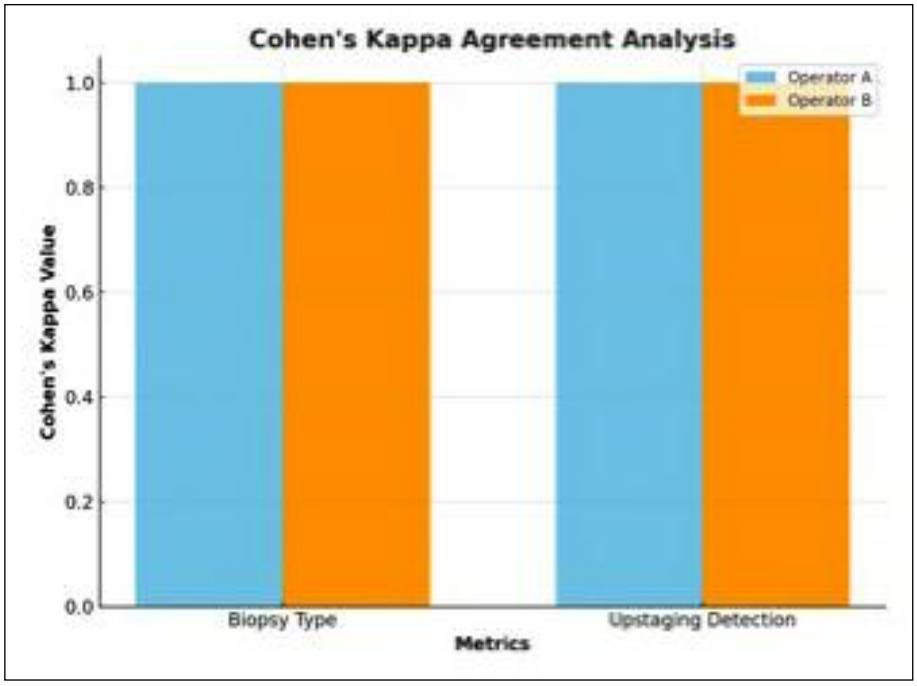


Figure 2.
Cohen's Kappa Agreement Analysis for Biopsy Type and Upstaging Detection: The bar chart displays Cohen's Kappa scores for biopsy type classification and upstaging detection, comparing the results obtained by Operator A and Operator B. Both operators achieved perfect agreement (Cohen's Kappa = 1.0) across all evaluated metrics, indicating complete concordance with the expert statistician's analysis.

approaches, ChatGPT delivered reliable statistical results consistent with the expert's interpretations.

Discussion

The use of artificial intelligence platforms, such as the ChatGPT 4.0 tool, in the clinical field is a highly relevant topic and the application of such tools in clinical research represents a field that is not yet fully explored. Nowadays, a comprehensive review of the updated literature revealed only one published study evaluating the effectiveness of ChatGPT in performing clinical statistical analyses, compared to a human statistician (5). This finding highlights a significant research gap and underscores the novelty of investigating ChatGPT's potential as a reliable analytical tool in clinical research settings. The integration of *artificial intelligence* (AI) into clinical practice has become increasingly relevant, as demonstrated by numerous studies exploring AI-driven tools, such as ChatGPT. In this context, our study provides compelling evidence of ChatGPT's capability to perform accurate and consistent statistical analyses in clinical research, particularly in evaluating the impact of biopsy methods on prostate cancer upstaging. This aligns with findings from the MRI-Targeted Biopsy Study, which demonstrated how advanced imaging-guided biopsies improved diagnostic precision and incorporating AI-driven analyses into such protocols could further streamline clinical decision-making processes (9). Our findings resonate with broader research on AI applications in healthcare, highlighting both the strengths and limitations of such technologies. For instance, other studies found that ChatGPT could assist in generating diagnostic insights in urology, providing context-aware responses based on clinical scenarios (8). This demonstrates how AI-driven models can augment diagnostic accuracy when integrated with standard medical protocols (3). Moreover, ChatGPT has been employed in medical research tasks such as systematic reviews, significantly reducing the time required for literature screening and data extraction, as shown in cardiology-focused AI research tutorials (13). Several studies emphasize the advantages of using ChatGPT in clinical data analysis. showing that ChatGPT-4 could achieve analytical efficiency and user-friendliness comparable to traditional statistical software like SAS, SPSS, and R. (5). This evidence aligns with our results, showing a strong agreement between ChatGPT and human-driven analyses in Logistic Regression, ANOVA, and Chi-Square tests. Its ability to generate accurate statistical outputs, without requiring complex coding, underscores its potential as a valuable analytical tool in clinical research, even for users with limited statistical expertise. In the clinical domain, ChatGPT has also turned out to be a decision-support tool. It has been noted that while ChatGPT is not a substitute for clinical judgment, it can enhance decision-making by providing up-to-date medical research and clinical guidelines (5). Similarly, ChatGPT's effectiveness in producing radiology reports with clarity and precision was highlighted in a review of its clinical application. This endorses our findings, where ChatGPT consistently produced statistical summaries comparable to those generated by experienced statisticians. However, our study also reveals inherent limitations. ChatGPT's lack of automatic citation generation was a notable drawback, reflected in lower evaluation scores in our study (Q-EVA). This issue has been widely discussed, particularly regarding the risk of academic misconduct and the spread of misinformation. The need for human oversight in interpreting AI-generated results is thus critical, especially in high-stakes clinical research and decision-making environments. A 2023 review highlights similar concerns, stressing how, while AI tools like ChatGPT can automate complex data analyses, transparent reporting and human verification remain essential to prevent errors and maintain clinical

integrity (15). Similarly, it is crucial to standardize the formulation of queries submitted to the AI platform, in order to minimize result variability caused by potential system misinterpretations. The Q-EVAL tool was designed for this purpose, although further studies are required to evaluate its effectiveness and explore its potential implementation. From a clinical perspective, our study established the superiority of Fusion combined (MRI/US) target and random biopsy over random biopsy alone in reducing post-operative upstaging for prostate cancer, in accordance with findings reported in the literature (16). The protective effect of fusion biopsy in reducing upstaging risk, confirmed through logistic regression, aligns with these clinical outcomes. This highlights the potential of combining AI-driven analytics with advanced diagnostic techniques for improved patient management. Looking ahead, the integration of ChatGPT into clinical workflows could enhance productivity, reduce human error, and lower the operational threshold for performing complex analyses. Additionally, research about prostate cancer diagnostics suggests that combining AI-assisted analyses with MRI-TRUS fusion techniques could further enhance the precision of tumor localization and reduce unnecessary biopsies (10). As A.I. technologies continue to evolve, addressing issues such as source transparency, data privacy, and interpretability will be essential. Future research should explore hybrid models that combine A.I. capabilities with human expertise, promoting data-driven decision-making in clinical practice, refine and standardize querying methods for interrogation of A.I. systems in the medical-scientific field. In this evolving landscape, ChatGPT and similar A.I. tools hold significant promise as supportive technologies in medical research and clinical care. The integration of ChatGPT as an analytical tool demonstrated significant potential for supporting clinical research by providing accurate statistical analyses comparable to those of expert statisticians. These results suggest that AI-powered tools could streamline future clinical research workflows, enhancing efficiency without compromising analytical accuracy.

Conclusions

This study demonstrates how ChatGPT (version 4.0) is a reliable and consistent tool for statistical analysis in clinical research, showing high concordance with human statisticians across both correlation analyses (e.g., logistic regression) and comparative tests (e.g., Chi-Square, ANOVA). Statistical agreement was confirmed by p-values < 0.05, with logistic regression for biopsy type yielding Chi-Square tests. Cohen's kappa coefficients for biopsy type and upstaging reached 1.0, indicating perfect alignment between AI-driven and human analyses. Q-EVAL tool can reduce inaccuracies and errors related to poorly formulated user queries, although further studies are needed to explore this hypothesis. Application of Q-EVA tools highlighted ChatGPT's capacity for logical and technically sound responses, though its inability to generate automatic source citations underscores a key limitation requiring future improvement. Regarding the secondary objective, fusion biopsy significantly reduced the risk of upstaging after RaRP. While ChatGPT shows strong analytical potential, it should currently be viewed as an assistive tool complementing human expertise rather than replacing it. Future research should explore AI's analytical boundaries and the dynamics of human-AI collaboration to enhance the integration of AI-powered tools into clinical research workflows.

References

1. Qin S, Chislett B, Ischia J, et al. ChatGPT and generative AI in urology and surgery - a narrative review. BJUICompass. 2024; 5:813-21.

2. Mu Y, He D. The Potential Applications and Challenges of ChatGPT in the Medical Field. Int J Gen Med. 2024; 17:817-826.

3. Lazaros T, Konstantinos K, Georgios F, et al. ChatGPT in clinical medicine, urology and academia: a review. Arch. Esp. Urol. 2024; 77: 708-717.

4. Teperikidis L, Boulmpou A, Papadopoulos C, et al. Using ChatGPT to perform a systematic review: a tutorial. Minerva Cardiology and Andrology. 2024; 72:547-67.

5. Huang Y, Wu R, He J, et al. Evaluating ChatGPT-4.0's data ana-

Declarations

Ethical approval: *This study was approved by the Local Ethics Committee of Bari (BA), IRCCS Oncological Institute "Gabriella Serio" (Protocol number: 2112/CEL - Study "PrOPT").*

Availability of data and material: *The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.*

Competing interests: *The authors declare no competing interests.*

Funding: *This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.*

Authors' contributions (according to http://www.icmje.org/#author): *1: Pier Paolo Prontera - author corresponding Substantial contributions to the conception, design of the work, acquisition, analysis and interpretation of data for the work. Drafting the work, reviewing it critically for important intellectual content. Final approval of the version to be published. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved; 2: Francesca Romana Prusciano, Francesco Saverio Grossi: Substantial contributions to the conception, design of the work, acquisition, analysis and interpretation of data for the work. Drafting the work, reviewing it critically for important intellectual content. Final approval of the version to be published. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved; 3: Marco Lattarulo, Arman Tsaturyan; Francesco Addabbo, Carmine Sciorio: Substantial contributions to the interpretation of data for the work. Reviewing it critically for important intellectual content. Final approval of the version to be published. Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.*

Acknowledgments: *Not applicable.*

lytic proficiency in epidemiological studies: a comparative analysis with SAS, SPSS, and R. J Glob Health 2024; 14:04070.

6. Aykut D. A Comparison of ChatGPT and human questionnaire evaluations of the urological cancer videos most watched on YouTube. Clinical Genitourinary 2024; 22:102145.

7. Prontera PP, Prusciano FR, Lattarulo M, et al. Quality of bladder cancer treatment information on YouTube: may the user's profile affect the quality of results? Arch Ital Urol Androl 2024; 96:12179.

8. Braga Martinelli AVN, Nunes NC, Santos EN, et al. Use of ChatGPT in urology and its relevance in clinical practice: is it useful? Int Braz J Urol 2024; 50:192-198.

9. Kasivisvanathan V, Rannikko AS, Borghi M, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. N Engl J Med 2018; 378:1767-77.

10. Baco E, Ukimura O, Rud E, et al. Magnetic resonance imaging-transectal ultrasound image-fusion biopsies accurately characterize the index tumor: correlation with step-sectioned radical prostatectomy specimens in 135 patients. Eur Urol 2014; 67:787-794.

11. Porpiglia F, De Luca S, Passera R, et al. Multiparametric-magnetic resonance/ultrasound fusion targeted prostate biopsy improves agreement between biopsy and radical prostatectomy gleason score. Anticancer Res. 2016; 36:4833-9.

12. Borkowetz A, Platzek I, Toma M, et al. Direct comparison of multiparametric magnetic resonance imaging (MRI) results with final histopathology in patients with proven prostate cancer in MRI/ultrasonography-fusion biopsy. BJU Int. 2016; 118:213-20.

13. Lanz C, Cornud F, Beuvon F, et al. Gleason score determination with transrectal ultrasound-magnetic resonance imaging fusion guided prostate biopsies are we gaining in accuracy? J Urol. 2016; 195:88-93.

14. D'Amico AV, Whittington R, Malkowicz SB, et al. Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer. AMA. 1998; 280:969-74.

15. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front. Artif. Intell. 2023; 6:1169595.

16. Flammia RS, Hoeh B, Hohenhorst L, et al. Adverse upgrading and/or upstaging in contemporary low-risk prostate cancer patients. Int Urol Nephrol. 2022; 54:2521-2528.

Correspondence

Pier Paolo Prontera, MD (Corresponding Author)
pierpaolo.prontera@asl.taranto.it

Lattarulo Marco, MD
marco.lattarulo@asl.taranto.it

Francesco Saverio Grossi, MD, PhD
francescos.grossi@asl.taranto.it

Department of Urology, "S.S. Annunziata" Hospital,
Via Bruno Francesco 1 - 74010 Taranto (Italy)

Francesca Romana Prusciano, MD
francescaprusciano@gmail.com

Division of Urology, Hospital "Valle D'Itria", Martina Franca (TA), Italy

Arman Tsaturyan, MD, PhD
tsaturyanarman@yahoo.com

Department of Urology, Yerevan State Medical University after Mkhitar
and Department of Urology Erebouni Medical Center, Heratsi, Yerevan, Armenia

Francesco Addabbo, MD, PhD
francesco.addabbo@asl.taranto.it

Unit of Statistics and Epidemiology, Local Health Authority of Taranto,
Taranto (Italy)

Carmine Sciorio, MD
carmine.sciorio@gmail.com

Department of Urology, "Alessandro Manzoni" Hospital of Lecco, 23900 (Italy)