JACC: ADVANCES VOL. 5, NO. 7, 2026
© 2026 THE AUTHORS. PUBLISHED BY ELSEVIER ON BEHALF OF THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION. THIS IS AN OPEN ACCESS ARTICLE UNDER THE CC BY-NC-ND LICENSE (http://creativecommons.org/licenses/by-nc-nd/4.0/).

ORIGINAL RESEARCH

EMERGING TECHNOLOGIES AND INNOVATIONS

From Theory to Practice

GPT-Supported Data Extraction in Observational Studies on Transcatheter Aortic Valve Replacement

Gloria Brigiari, MD,[a,b,c] Roberta Dotto, MA,[d] Carlo Cernetti, MD,[d] Dario Gregori, PhD,[a,b,*] Giulia Lorenzoni, PhD[a,b,*]

ABSTRACT

BACKGROUND Manual record abstraction is the standard method for data collection in observational studies but is labor-intensive, error-prone, and difficult to scale, particularly when information is embedded in unstructured electronic health records. Large language models (LLMs) may streamline this process, yet evidence from real-world clinical research remains limited.

OBJECTIVES This study evaluated the performance of a generative pre-trained transformer (GPT)-based LLM in extracting sociodemographic, procedural, and outcome variables from free-text electronic health records of patients undergoing transcatheter aortic valve replacement.

METHODS We conducted a retrospective analysis of medical and nursing records for all transcatheter aortic valve replacement procedures performed at Ca' Foncello Hospital (Treviso, Italy) between January and June 2024. Manual abstraction by 2 reviewers served as the reference standard. Accuracy, sensitivity, and specificity with 95% CIs were calculated. Agreement for continuous variables was assessed using Bland-Altman analyses.

RESULTS A total of 108 cases were included. GPT achieved accuracy ranging from 0.657 (valve brand) to 1.00 (gender, Barthel index, procedure timings, and several intraoperative complications). Sensitivity reached 1.00 for rare but clinically important events, including intraoperative neurological complications, whereas specificity exceeded 0.90 for most variables. For vital parameters, Bland-Altman analyses demonstrated minimal bias and narrow limits of agreement.

CONCLUSIONS GPT-based data extraction showed high accuracy across a broad range of variables, particularly continuous repeated measurements and rare intraoperative outcomes. Performance was lower for some infrequent postoperative events, reflecting sparse true positives. These findings support the feasibility of integrating LLM-assisted extraction into observational research workflows, with further validation needed in larger or multicenter cohorts. (JACC Adv. 2026;5:102871) © 2026 The Authors. Published by Elsevier on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

From the [a]Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Padova, Italy; [b]BIOSTAT-X Biostatistics & AI for Biomedical Discovery, Pediatric Research Institute (IRP) "Città della Speranza", Padova, Italy; [c]PhD Program in Translation Specialistic Medicine "G.B. Morgagni", Curriculum "Biostatistics and Clinical Epidemiology", University of Padova, Padova, Italy; and the [d]Division of Cardiology, Neuro-Cardio-Vascular Department, Ca' Foncello Hospital, Treviso, Italy. *These are co-last authors

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

Manuscript received December 17, 2025; revised manuscript received April 13, 2026, accepted May 11, 2026.

ISSN 2772-963X https://doi.org/10.1016/j.jacadv.2026.102871

ABBREVIATIONS AND ACRONYMS

GPT = generative pretrained transformer

LLM = large language model

MRA = manual record abstraction

NA = not available / missing value category

TAVR = transcatheter aortic valve replacement

Observational studies account for a substantial portion of published clinical research, approximately 75% of the studies in PubMed, with cross-sectional and cohort studies showing an increasing trend in recent years.[1] Furthermore, although randomized controlled trials are at the top of the evidence pyramid in clinical epidemiological research, there has been growing recognition of the need for real-world evidence to bridge the gap between clinical trials and everyday clinical practice.[2]

One of the primary reasons for the high proportion of observational compared to experimental studies is the relative ease of conducting them as they require far less time and fewer resources than randomized controlled trials, for example, data from more than seven hundred phase 3 trials indicate that the median cost is 21.4 million dollars.[3] Consequently, 1 of the most appealing aspects of observational studies—beyond their faster regulatory processes—is that they can often be conducted at very low cost, especially in retrospective designs that do not require active longitudinal patient follow-up.

Although they are less expensive, it is important not to underestimate the burden of observational studies, even in the retrospective setting. Certainly, their costs are markedly lower than those of clinical trials and prospective observational studies, but they are not zero—especially once the work required for data collection is accounted for. The increasing availability of electronic medical records has helped speed up data collection in retrospective observational studies. Unfortunately, these systems are not yet universally available, and even when they are, they are primarily designed for clinical care rather than specific research purposes. In addition, restrictions may exist on their use for purposes other than clinical care, making it challenging to directly use data gathered during clinical management for research. Finally, multicenter observational studies often require centralized data collection systems that do not easily align with data already collected for clinical management, contributing to the persistent use of manual record abstraction (MRA).

Often, this manual collection is performed by health care personnel in training who are not explicitly trained in data collection or deeply familiar with the topic under study. As a result, manual data collection has frequently been shown to be prone to errors,[4] compromising data quality—especially because, in most single-center retrospective observational studies, there are no formal systems to monitor data quality.

The recent advent of large language models (LLMs) may transform data collection through automatic extraction from clinical records, with promising results.[5,6] Such systems could partially replace human involvement in data collection, allowing researchers to set up databases with automatically extracted clinical data for research purposes.

The present study aimed to assess the ability of an LLM to extract and organize a database for clinical research in interventional cardiology, specifically for transcatheter aortic valve replacement (TAVR), a highly active area of research in recent years.

MATERIALS AND METHODS

This retrospective study analyzed anonymous electronic health records, including medical and nursing diaries, of patients who underwent TAVR in the Cardiology Department of the ULSS 2 Marca Trevigiana Ca' Foncello Hospital, Treviso, Veneto, Italy, and identified from procedural logs by the clinical team. The analyses included all TAVR cases from 1 January to 30 June 2024. This study was conducted using retrospectively collected, fully anonymized data. According to applicable national regulations and institutional policies, formal approval by an institutional review board or ethics committee was not required for studies based exclusively on anonymized data.

GOLD STANDARD: MANUAL CLASSIFICATION. Two field experts manually classified the information contained in the medical and nursing diaries, which were documented in an unstructured manner according to the following categories.

1) Sociodemographic data (gender and social context);
2) Admission data (Barthel index and allergies);
3) Hospital admission information (date of admission and discharge);
4) Vital parameters (date and time of each data collection; body temperature; daily diuresis; blood pressure; heart rate; oxygen saturation).
5) Procedural characteristics (procedure date and time; vascular access; valve brand; valve size, device used to close the vascular site, intraoperative complications)
6) Postoperative data (postoperative complications and Barthel index at discharge)

All the information pertained to a single point during the hospital stay except for the vital parameters. Vital parameters recorded throughout the hospital stay were collected, including all available time points. As measurements were repeatedly recorded

at regular intervals, all data points were considered for the analysis to capture the complete temporal profile of the patient's clinical status.

The data were extracted using a REDCap data collection form. A third independent reviewer solved any disagreement. The manual extraction served as the gold standard to evaluate the performance of the LLM model.

GENERATIVE PRETRAINED TRANSFORMER-BASED CLASSIFICATION. The LLM chosen for the task was generative pretrained transformer (GPT)-5, accessed via the OpenAI API. GPT was chosen because it is widely used in literature for classification tasks within clinical data.

All records submitted to the model were anonymous and thoroughly reviewed to prevent any risk of data disclosure. They were provided in the Italian language.

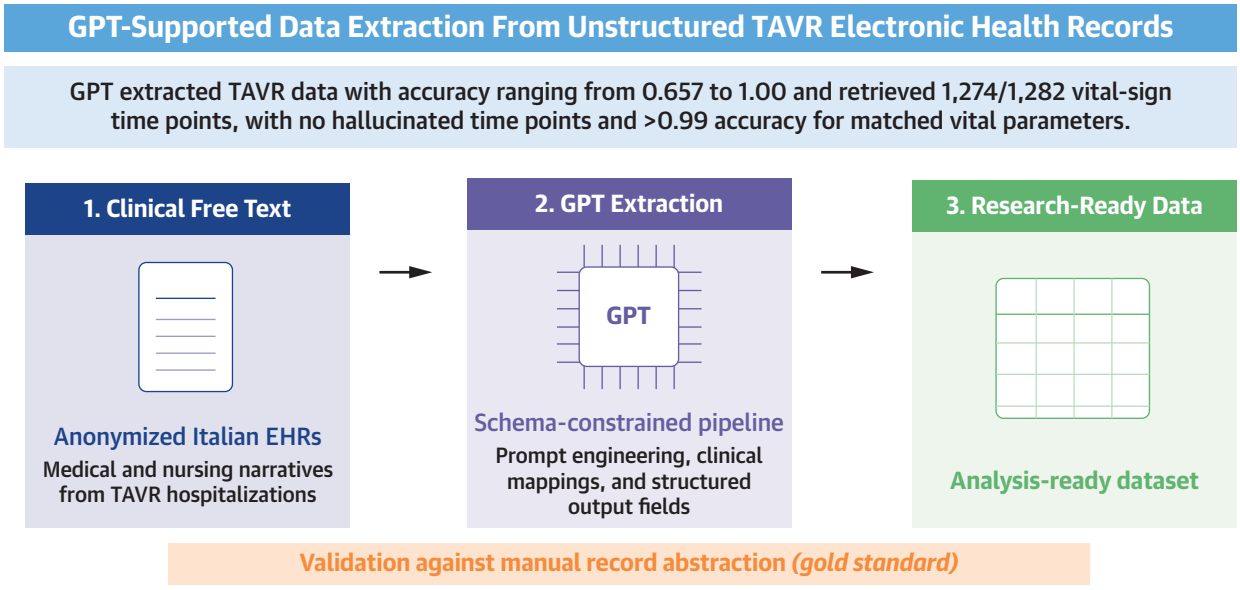
The single, comprehensive prompt was generated according to OpenAI guidelines, and it underwent an iterative generation process. GPT was asked to extract and classify specific information according to a predefined list of options (eg, valve brand).

The model was also tasked with performing more complex operations, such as calculating the procedure duration. This information was not explicitly provided in the electronic health records; therefore, the model had to derive it from the relevant dates and times available in the data source. The complete prompt and type specifications provided to GPT are presented in the Supplemental Appendix. It is important to note that the system under evaluation is not the language model alone, but rather an integrated extraction pipeline combining the GPT model with structured prompt engineering, predefined clinical mappings (eg, valve brand families, access-site terminology, closure-device variants), and a strictly constrained output schema. Accordingly, the performance reported in this study reflects the behavior of this pipeline as a whole. Each record was processed once, and all model calls were performed using OpenAI's default inference parameters. In particular, the temperature was set to 1.0, top-p to 1.0, whereas frequency and presence penalties were left at their default value of 0. The choice to retain the default temperature of 1.0 was intentional, as it reflects the out-of-the-box configuration that a clinical researcher would most likely adopt. Although a lower temperature would reduce stochastic variability and may be preferable for deterministic extraction tasks, the constrained output schema limits the space of valid responses, partially mitigating the effect of sampling randomness. Model-based data processing was performed within GPT Teams in a secure, isolated environment designed to prevent unauthorized access and ensure that no data are used for subsequent model training. The overall extraction and evaluation workflow is summarized in the **Central Illustration**.

STATISTICAL ANALYSIS. Descriptive statistics were presented, summarizing continuous variables with median (I-III quartiles) and categorical data as absolute number (percentages).

Both for categorical and continuous variables, the model's performance was assessed using accuracy with 95% CIs as the primary metric, defined as the proportion of GPT-extracted values that exactly matched the corresponding gold standard value. All categorical variables, including binary ones, included "not available/missing value category (NA)" as a valid response. The model was instructed to output "NA" when information was missing or not inferable from the record, distinguishing this from a definitive "No". Whenever missing data were present in either the gold standard extraction or the GPT-based extraction, these missing values were made explicit by defining an additional response category of the variable ("NA"). Performance metrics therefore accounted for missingness as a legitimate class rather than excluding these observations, as distinguishing absent information from a confirmed negative finding is clinically meaningful, and ignoring NA would introduce systematic misclassification, albeit at the cost of potentially optimistic accuracy and specificity estimates for variables with nonzero missingness rates in the gold standard. Sensitivity and specificity were estimated using a one-vs-rest formulation for each response level, with 95% CIs derived from exact binomial (Clopper-Pearson) intervals; for binary variables these were computed in the standard fashion, whereas for variables with 3 or more response levels a one-vs-rest formulation was applied, computing sensitivity and specificity for each level separately against the remaining categories. For continuous variables, the agreement between model-derived values and the gold standard was also assessed using Bland-Altman plots, which allowed visual evaluation of model performance, as well as the magnitude and direction of misclassification. For vital parameters, GPT-derived measurements were matched to gold standard values by patient identifier and datetime stamp, and then accuracy was computed. Analyses were performed in R (version 4.5.1; R Foundation for Statistical Computing, Vienna, Austria), with the caret,[7] binom, blandr,[8] and ellmer[9] packages.

CENTRAL ILLUSTRATION **Workflow of GPT-Based Data Extraction and Evaluation in TAVR EHRs**



Unstructured EHRs are processed through a structured pipeline combining prompt engineering, clinical mappings, and a constrained output schema to extract clinical variables. Extracted data are compared with manual record abstraction to assess performance and agreement. EHR = electronic health records; GPT = generative pre-trained transformer; TAVR = transcatheter aortic valve repalcement.

RESULTS

The study included 108 cases of TAVR during the study period (Table 1). The male gender was the most prevalent (57%) (Table 1). The median Barthel index was similar at admission (90) and discharge (95), with a median hospital stay of 5 days. Concerning the procedural characteristics, the median duration of the procedure was 90 minutes, and 22% of the patients suffered from intraoperative complications. Postoperatively, complications occurred in 43% of patients, with hypo arrhythmias being the most frequent event (18%).

GPT MODEL PERFORMANCE. The performance of the GPT model in data extraction and classification was very good (Table 2). The accuracy values ranged from a minimum of 0.657 (for valve brand) to a maximum of 1 (indicating perfect classification performance as per Barthel index at admission, gender, procedure start and stop times, and several intraoperative complication items). GPT showed excellent sensitivity for rare but clinically relevant outcomes, such as intraoperative neurological complications (1.00), correctly identifying all true cases despite their low prevalence. Similarly, specificity consistently exceeded 0.90 for the majority of variables, indicating a low rate of false positives. Conversely, a number of variables with extremely low prevalence (eg, intraoperative dissection, rare closure devices, and selected postoperative events) showed sensitivity values close to zero. These estimates should be interpreted with caution, as they reflect the scarcity of true positive cases in the data set rather than systematic misclassification by the model. Of 1,282 vital sign measurement time points

TABLE 1 Sample Characteristics (Gold Standard) (N = 108)

Gender	
NA	0 (0%)
F	46 (43%)
M	62 (57%)
Barthel index at admission	90 (80, 95)
Unknown	4
Barthel index at discharge	95 (85, 95)
Unknown	14
Family status	
NA	9 (8.3%)
Live alone	31 (29%)
Live with others	68 (63%)
Allergies	
NA	1 (0.9%)
No	83 (77%)
Yes	24 (22%)
Procedure duration	90 (75, 100)
Unknown	1
Access site	
NA	0 (0%)
Left femoral artery	10 (9.3%)
Left subclavian artery	1 (0.9%)
Right femoral artery	96 (89%)
Right subclavian artery	1 (0.9%)
Valve brand	
NA	0 (0%)
Edwards Sapien	46 (43%)
Medtronic	62 (57%)
Valve size	
20	1 (0.9%)
23	20 (19%)
26	31 (29%)
29	39 (36%)
34	16 (15%)
Unknown	1
Angioseal closure device	
NA	0 (0%)
No	12 (11%)
Yes	96 (89%)
Proglide closure device	
NA	0 (0%)
No	9 (8.3%)
Yes	99 (92%)
Other closure device	
NA	0 (0%)
No	104 (96%)
Yes	4 (3.7%)
Intraoperative complications	
NA	0 (0%)
No	84 (78%)
Yes	24 (22%)
Intraoperative hemodynamic instability	
NA	0 (0%)
No	103 (95%)
Yes	5 (4.6%)

Continued in the next column

TABLE 1 Continued

Intraoperative vascular complications	
NA	0 (0%)
No	96 (89%)
Yes	12 (11%)
Intraoperative dissection	
NA	0 (0%)
No	105 (97%)
Yes	3 (2.8%)
Intraoperative hypo arrhythmias	
NA	0 (0%)
No	100 (93%)
Yes	8 (7.4%)
Intraoperative neurological complications	
NA	0 (0%)
No	107 (99%)
Yes	1 (0.9%)
Other intraoperative complications	
NA	0 (0%)
No	105 (97%)
Yes	3 (2.8%)
Postoperative complications	
NA	0 (0%)
No	62 (57%)
Yes	46 (43%)
Postoperative hypo arrhythmias	
NA	0 (0%)
No	89 (82%)
Yes	19 (18%)
Postoperative stroke	
NA	0 (0%)
No	105 (97%)
Yes	3 (2.8%)
Postoperative vascular complications	
NA	0 (0%)
No	96 (89%)
Yes	12 (11%)
Postoperative hemodynamic instability	
NA	0 (0%)
No	100 (93%)
Yes	8 (7.4%)
Postoperative death	
NA	0 (0%)
No	108 (100%)
Yes	0 (0%)
Postoperative pacemaker implantation	
NA	0 (0%)
No	108 (100%)
Yes	0 (0%)
Other postoperative complication	
NA	0 (0%)
No	85 (79%)
Yes	23 (21%)

Continued on the next page

TABLE 1 Continued

Vital parameters	
Diuresis (mL), n = 490	800 (500, 1,100)
Temperature, n = 583	36 (36, 36)
Blood pressure (systolic), n = 968	129 (117, 141)
Blood pressure (diastolic), n = 967	65 (58, 74)
Heart rate, n = 924	67 (60, 76)
Oxygen saturation, n = 901	96 (95, 98)

Data are median (I, III quartiles) for continuous variables and absolute numbers (percentages) for categorical variables.
NA = not available / missing value category.

in the gold standard, GPT identified 1,274, with no hallucinated time points. Eight time points present in the gold standard were not retrieved. For matched vital parameters, accuracy exceeded 0.99 across all measurements, and Bland-Altman plots demonstrated close agreement between GPT-derived values and manual extraction, with minimal bias and narrow limits of agreement (**Figure 1**). No systematic overestimation or underestimation was observed.

DISCUSSION

In recent years, interest in LLMs in clinical research has grown rapidly, particularly in extracting information from free texts that would otherwise remain unused for research purposes. Cardiology is one of the clinical fields in which these applications are being actively explored.[10] For example, a recent study reported that ChatGPT achieved 90.6% accuracy, 100% recall, and an F1 score of 95% in identifying symptoms and signs of heart failure from electronic medical records, clearly outperforming traditional machine learning and natural language processing approaches.[11] Going more specifically into the field of interventional cardiology, a recent study evaluated ChatGPT's performance at extracting pulmonary perfusion and cardiac output indices from catheterization reports, achieving an encouraging 3% error rate, consistent with the present results.[6] Nevertheless, data extraction is not the only application currently being explored:[12] LLMs have also been applied to support clinical decision-making, such as in heart team discussions for aortic stenosis treatment,[13] to assist in the interpretation of coronary angiography images,[14] and to deliver patient education.[15]

Overall, the results obtained from data extraction studies, including the present one, are encouraging. In particular, Bland-Altman analyses demonstrated excellent agreement between GPT-derived values and the gold standard for all continuous variables, with negligible bias and narrow limits of agreement. This finding suggests that LLMs may be especially reliable in handling repeated measurements, an area in which manual extraction is time-consuming and error prone.

However, so far, automatic data extraction has often been presented as an interesting exercise, demonstrating good performance but with limited practical application, usually concluding with observations on the need for human supervision. The key question now is: can LLMs be integrated into everyday clinical research activities? Some studies comparing manual and automated review of medical records have demonstrated excellent performance with LLMs, but also highlight limitations for more complex tasks, where human supervision still seems necessary.[16]

It is interesting to note that nearly all items in the study show accuracy above 90%. A recent systematic review reported an overall error rate of 6.57% for medical record extraction (MRA),[17] a figure consistent with the performance observed with GPT. The literature suggests that a substantial reduction in MRA errors requires implementing systematic quality control methods and the training of data collectors.[18,19] However, these procedures are resource-intensive, and in an observational setting, they are generally only feasible in sponsored, usually multicenter, studies. This brings us a crucial question: Are we ready to effectively implement these systems in clinical research?

The usefulness of LLMs for data extraction is becoming increasingly apparent. However, it is important to note that comparing results across different studies remains difficult, given the rapid pace of model evolution and the variability of both data sources and target information. The GPT model remains the most widely used and produces the best overall results, which explains our decision to focus on this model for the present work. The results of the

TABLE 2 Performance Metrics

	Sensitivity (CI 95%)	Specificity (CI 95%)	Accuracy (CI 95%)
Sociodemographic and admission data			
Sex			
F	1 (0.92-1)	0.969 (0.892-0.996)	0.981 (0.935-0.998)
M	1 (0.942-1)	1 (0.923-1)	
NA	0 (0-0.842)	1 (0.966-1)	
Family status			
Live alone	0.966 (0.822-0.999)	0.975 (0.912-0.997)	0.954 (0.895-0.985)
Live with others	0.957 (0.88-0.991)	0.947 (0.823-0.994)	
NA	0.889 (0.518-0.997)	0.99 (0.945-1)	
Allergies			
Yes	0.92 (0.74-0.99)	0.964 (0.898-0.992)	0.944 (0.883-0.979)
No	0.963 (0.896-0.992)	0.889 (0.708-0.976)	
NA	0.5 (0.013-0.987)	1 (0.966-1)	
Barthel Index (at admission)			1 (0.965-1)
Procedural characteristics			
Brand valve			
Edwards Sapien	0.957 (0.855-0.995)	0.984 (0.912-1)	0.657 (0.56-0.746)
Medtronic	1 (0.868-1)	0.561 (0.447-0.67)	
Other	0 (0-0.103)	1 (0.951-1)	
NA	0 (0-0.975)	1 (0.966-1)	
Valve size			
20	1 (0.025-1)	1 (0.966-1)	1 (0.966-1)
23	1 (0.832-1)	1 (0.958-1)	
26	1 (0.888-1)	1 (0.953-1)	
29	1 (0.91-1)	1 (0.947-1)	
34	1 (0.794-1)	1 (0.96-1)	
Angioseal closure device, Yes vs no	0.99 (0.944-1)	0.6 (0.262-0.878)	0.963 (0.908-0.99)
Proglide closure device, Yes vs no	0.99 (0.945-1)	0.556 (0.212-0.863)	0.954 (0.895-0.985)
Other closure device, Yes vs no	0 (0-0.975)	0.72 (0.625-0.802)	0.713 (0.618-0.796)
Access site			
Left femoral artery	0.8 (0.444-0.975)	0.98 (0.928-0.998)	0.944 (0.883-0.979)
Left subclavian artery	1 (0.025-1)	1 (0.966-1)	
Right femoral artery	0.969 (0.911-0.994)	0.75 (0.428-0.945)	
Right subclavian artery	0 (0-0.975)	0.991 (0.949-1)	
Procedure start time, Continuous			0.981 (0.935-0.998)
Procedure stop time, Continuous			0.962 (0.906-0.99)
Intraoperative complications, Yes vs no	0.714 (0.513-0.868)	0.975 (0.913-0.997)	0.907 (0.836-0.955)
Intraoperative dissection, Yes vs no	0 (0-0.842)	0.991 (0.949-1)	0.972 (0.921-0.994)
Intraoperative hemodynamic instability, Yes vs no	0.4 (0.053-0.853)	0.961 (0.904-0.989)	0.935 (0.871-0.974)
Intraoperative hypo arrhythmias, Yes vs no	0.667 (0.299-0.925)	1 (0.963-1)	0.972 (0.921-0.994)
Intraoperative neurological complications, Yes vs no	1 (0.025-1)	1 (0.966-1)	1 (0.966-1)
Intraoperative vascular complications, Yes vs no	0.833 (0.516-0.979)	0.99 (0.943-1)	0.972 (0.921-0.994)
Postoperative data			
Postoperative complications			
Yes	0.69 (0.529-0.824)	0.788 (0.67-0.879)	0.731 (0.638-0.812)
No	0.781 (0.66-0.875)	0.659 (0.501-0.795)	
NA	0 (0-0.842)	1 (0.966-1)	
Postoperative hemodynamic instability, Yes vs no	0.444 (0.137-0.788)	0.949 (0.886-0.983)	0.907 (0.836-0.955)
Postoperative hypo arrhythmias, Yes vs no	0.714 (0.419-0.916)	0.936 (0.866-0.976)	0.907 (0.836-0.955)
Postoperative stroke, Yes vs no	1 (0.025-1)	0.981 (0.934-0.998)	0.981 (0.935-0.998)
Postoperative pacemaker implantation, Yes vs no		0.981 (0.935-0.998)	0.981 (0.935-0.998)
Postoperative vascular complications, Yes vs no	0.636 (0.308-0.891)	0.907 (0.831-0.957)	0.88 (0.803-0.934)
Barthel Index (at discharge), Continuous			1

Continued on the next page

TABLE 2 Continued

	Sensitivity (CI 95%)	Specificity (CI 95%)	Accuracy (CI 95%)
Vital parameters			
Diuresis, Continuous			0.998 (0.988-1)
Heart rate, Continuous			0.996 (0.989-0.999)
Blood pressure (systolic), Continuous			0.992 (0.984-0.996)
Blood pressure (diastolic), Continuous			0.995 (0.988-0.998)
Oxygen saturation, Continuous			0.997 (0.99-0.999)
Temperature, Continuous			0.998 (0.99-1)

The table reports the accuracy together with the 95% CI. For categorical variables, also the specificity and sensitivity are presented.

present work should be interpreted as reflecting the performance of the full extraction pipeline—including prompt design, clinical mappings, and the constrained output schema—rather than the GPT model in isolation. Any change in these components (eg, different prompt structure, updated mappings, or alternative schema enforcement) could influence extraction accuracy.

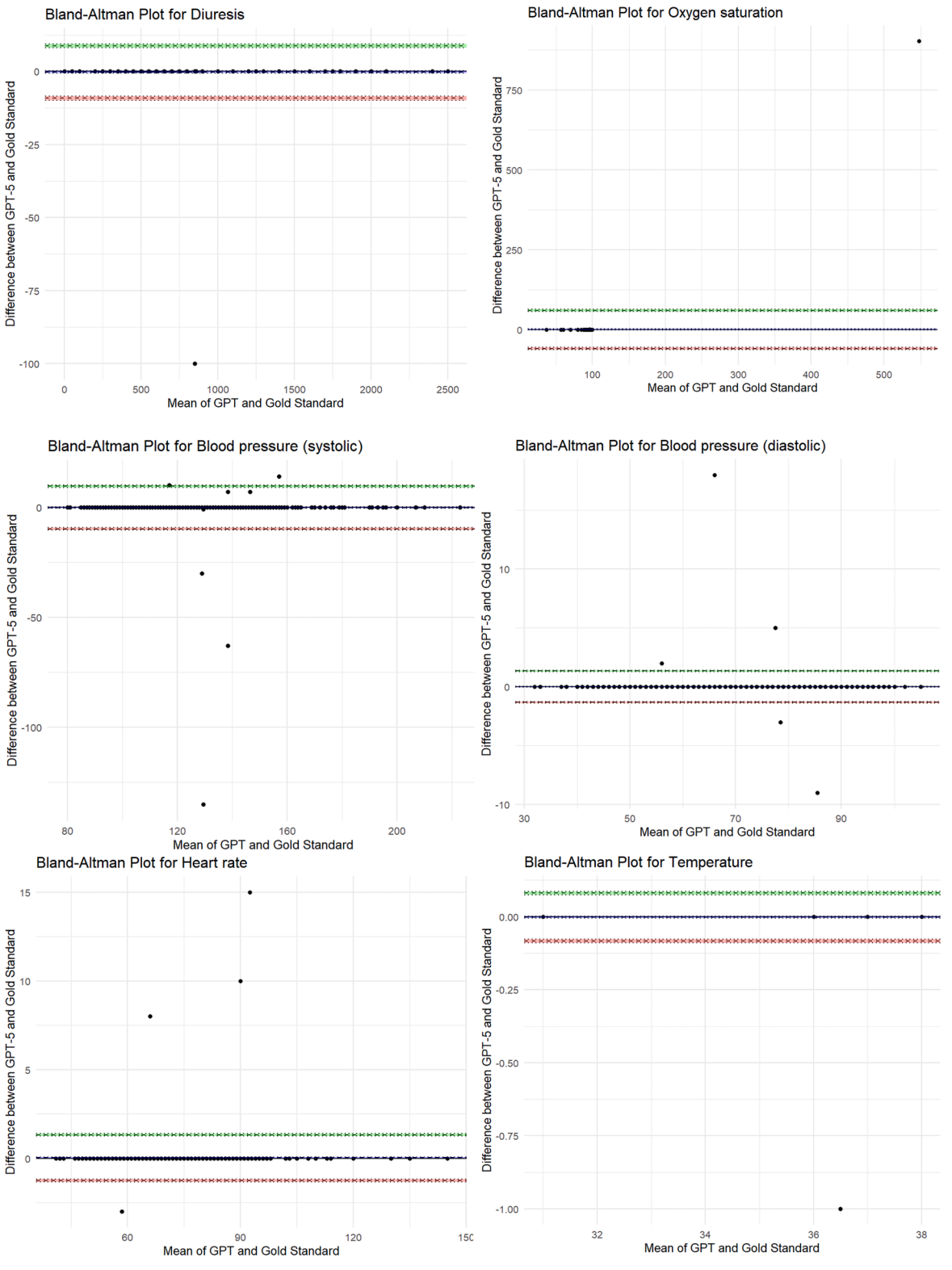
It would be helpful to explore further the possibility of actively integrating GPT into routine data collection in observational studies to assess its potential to reduce data-collection burden.

STUDY LIMITATIONS. Several limitations should be considered when interpreting the findings of this study. Because the analysis was conducted in a single center, model performance reflects the specific documentation style and workflow of that institution. Importantly, the extraction prompt was designed for Italian language documentation and was tailored to the specific narrative conventions, abbreviation systems, and workflow of a single institution. Transferring this pipeline to another hospital—even one operating in the same language—would likely require substantial prompt re-engineering, as documentation styles, clinical terminology, and data completeness can vary considerably. The use of a different LLM, or even a newer version of the same model, could also yield different extraction behavior. Therefore, the present results should not be generalized to other settings without dedicated validation, and broader adoption will require multicenter studies, ideally spanning different languages, documentation systems, and LLM platforms. Caution is also warranted when interpreting performance metrics for rare complications. Very low or zero sensitivity values for some variables primarily reflect the extremely small number of true positive cases rather than a systematic failure of the model to detect them. In the presence of such sparse outcomes, sensitivity estimates become inherently unstable, and larger samples or multicenter cohorts will be required to obtain more reliable assessments of model performance. In addition, each record was processed in a single run, and no assessment of output stability across repeated extractions was performed. Although the constrained output schema limits the range of possible responses, variability inherent to a temperature of 1.0 may lead to different outputs on repeated runs, particularly for ambiguous clinical text. Future studies should evaluate test-retest reliability and explore lower temperature settings to enhance reproducibility.

CONCLUSIONS

In this single-center retrospective cohort of patients undergoing TAVR, we showed that a GPT-based model can automatically extract a broad range of sociodemographic, procedural and outcome variables from unstructured clinical documentation with high accuracy. Overall accuracy ranged from approximately 0.66 to 1.00 across variables, with excellent performance for continuous repeated measures and several rare but clinically relevant events. At the same time, model performance was less robust for some complex postoperative outcomes, indicating that human oversight remains necessary, especially for nuanced or ambiguously documented events. These findings highlight that GPT-based systems should currently be viewed as tools to augment, rather than replace, MRA.

FIGURE 1 **Bland-Altman plots Showing Agreement Between GPT-Derived Values and Manual Extraction for Continuous Clinical Variables**



Each plot displays the mean of the GPT and gold-standard measurements on the x-axis and their difference on the y-axis. The central solid line represents the mean bias, whereas the dashed lines indicate the 95% limits of agreement. GPT = generative pre-trained transformer.

Taken together, our results support the feasibility of integrating GPT-assisted data extraction into observational research workflows in interventional cardiology. If prospectively validated in larger and multicenter cohorts, such systems could substantially reduce the burden and cost of manual data collection, facilitate higher-quality real-world evidence generation and ultimately accelerate clinical research in TAVR and other cardiovascular interventions.

FUNDING SUPPORT AND AUTHORS DISCLOSURES

This work was supported in part by the high-performance computing infrastructure developed under the project "CONVECS", funded by the PR Veneto FESR 2021 to 2027 program, Priority 1 - Specific Objective 1.1 - Action 1.1.2. The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Dario Gregori, Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Via Loredan, 18, Padova 35131, Italy. E-mail: dario.gregori@unipd.it.

PERSPECTIVES

COMPETENCY IN SYSTEMS-BASED PRACTICE Manual record abstraction remains the standard method for data collection in observational studies, but it is labor-intensive, error-prone, and difficult to scale, particularly when information is embedded in unstructured electronic health record notes. In this study, a GPT-based large language model automatically extracted sociodemographic, procedural, and outcome variables from free-text electronic health records of patients undergoing TAVR with high accuracy and agreement with manual abstraction, supporting its potential role in augmenting observational research workflows.

TRANSLATIONAL OUTLOOK Prospective validation in larger and multicenter cohorts is needed to confirm the generalizability and reliability of GPT-assisted data extraction, particularly for rare events and complex postoperative complications. Future studies should also evaluate its impact on time, cost, data quality, and integration into routine clinical research workflows.

REFERENCES

1. Zhao X, Jiang H, Yin J, et al. Changing trends in clinical research literature on PubMed database from 1991 to 2020. *Eur J Med Res*. 2022;27(1):95.

2. Wilson BE, Booth CM. Real-world data: bridging the gap between clinical trials and practice. *EClinicalMedicine*. 2024;78:102915.

3. Martin L, Hutchens M, Hawkins C, Radnov A. How much do clinical trials cost? *Nat Rev Drug Discov*. 2017;16(6):381-382.

4. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *BMC Med Res Methodol*. 2017;17(1):152.

5. Yu S, Le A, Feld E, et al. A natural language processing-assisted extraction system for Gleason scores: development and usability study. *JMIR Cancer*. 2021;7(3):e27970.

6. Barak-Corren Y, Gupta M, Tang J, et al. From text to data: automatically extracting data from catheterization reports using generative artificial intelligence. *J Soc Cardiovasc Angiogr Interv*. 2025;4(3):102242.

7. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw*. 2008;28(5):1-26.

8. Datta D. blandr: A Bland-Altman Method Comparison package for R [Internet]. 2017. Disponibile su. 2017. Accessed December 15, 2025. https://github.com/deepankardatta/blandr

9. Wickham H, Cheng J, Jacobs A, Aden-Buie G, Schloerke B. Ellmer: chat with large language models [Internet]. 2025. Disponibile su. 2025. Accessed December 15, 2025. https://CRAN.R-project.org/package=ellmer

10. Sharma A, Medapalli T, Alexandrou M, Brilakis E, Prasad A, Alexandrou M. Exploring the role of ChatGPT in cardiology: a systematic review of the current literature. *Cureus*. 2024;16(4):e58936.

11. Workman TE, Ahmed A, Sheriff HM, et al. ChatGPT-4 extraction of heart failure symptoms and signs from electronic health records. *Prog Cardiovasc Dis*. 2024;87:44-49.

12. Sule AA, Poh KK, Srinivasan DK. Role of ChatGPT in interventional cardiology. *Singapore Med J*. 2025. https://doi.org/10.4103/singaporemedj.SMJ-2024-139

13. Salihu A, Meier D, Noirclerc N, et al. A study of ChatGPT in facilitating heart team decisions on severe aortic stenosis. *EuroIntervention*. 2024;20(8):e496.

14. Hoppe JM, Kellnar A, Esser D, Diegruber K, Stremmel C. Artificial intelligence in coronary angiography: benchmarking the diagnostic accuracy of ChatGPT-4o against interventional cardiologists. *Open Heart*. 2025;12(2):e003316.

15. Lorenzoni G, Zanotto C, Sordo A, et al. Large language models to develop evidence-based strategies for primary and secondary cardiovascular prevention. *Eur Heart J Digit Health*. 2025;6(5):1069-1075.

16. Ge J, Li M, Delk MB, Lai JC. A comparison of a large language model vs manual chart review for the extraction of data elements from the electronic health record. *Gastroenterology*. 2024;166(4):707-709.

17. Garza MY, Williams T, Ounpraseuth S, et al. Error rates of data processing methods in clinical research: a systematic review and meta-analysis of manuscripts identified through PubMed. *Int J Med Inf*. 2024;195:105749.

18. Garza MY, Williams TB, Ounpraseuth S, et al. Comparing Medical Record Abstraction (MRA) error rates in an observational study to pooled rates identified in the data quality literature. *BMC Med Res Methodol*. 2024;24(1):304.

19. Garza MY, Williams T, Myneni S, et al. Measuring and controlling medical record abstraction (MRA) error rates in an observational study. *BMC Med Res Methodol*. 2022;22(1):227.

KEY WORDS automated data extraction, electronic health records (EHRs), large language models (LLMs), manual record abstraction (MRA), observational studies, transcatheter aortic valve replacement (TAVR)

APPENDIX For a supplemental table, please see the online version of this paper.