PAPER - RE14

Leveraging Large Language Models (LLMs) for Missing Data Imputation and Interpretation of Real-World Evidence Outputs.

Rahul Somavanshi[1] and Spencer Langerman[2]

[1]Navitas Life Science, Bangalore, India,

[2]Flatiron Health, New York, USA

Abstract

Real-world data (RWD) frequently contain substantial missingness, which can compromise the validity and interpretability of real-world evidence (RWE) analyses. Traditional statistical imputation methods perform well under restrictive assumptions but struggle with complex, nonlinear, and context-dependent clinical data. In this study, we evaluated the performance of a Large Language Model (LLM; Claude Sonnet 4.5) for imputing missing clinical variables and interpreting RWE outputs using the MIMIC-IV intensive care database. LLM-based imputation was compared with Multiple Imputation by Chained Equations (MICE) and MissForest across varying levels of missingness and sample sizes. For continuous variables, the LLM demonstrated accuracy comparable to MissForest and greater robustness than MICE, particularly at higher missingness levels, while categorical imputation performance varied across methods without a consistent best performer. Bias remained low across methods, though variability increased with extreme missingness. Few-shot prompting provided minimal benefit over zero-shot prompting, indicating stable model behavior without extensive prompt engineering. Additionally, the LLM generated accurate and clinically appropriate summaries of baseline tables and survival analyses, as validated by expert review. These findings suggest that LLMs can complement existing statistical approaches by improving both data completeness and interpretability in RWD-based clinical research.

Introduction

Real-world data (RWD), including electronic health records (EHRs), claims, registries, and patient-generated data, frequently exhibit substantial missingness due to irregular measurement, incomplete documentation, variable clinical workflows, and heterogeneous data capture systems. Unlike randomized controlled trials (RCTs), which operate under standardized data collection procedures, RWD originates from routine clinical practice where information is recorded primarily for patient care rather than research. As a result, key variables such as laboratory measurements, vital signs, comorbidities, medication records, and outcomes may be sporadically observed or systematically absent (1). Missing data can influence downstream analyses in RWD-based clinical studies, as inaccurate imputation may lead to misclassification of disease severity, risk scores, and prognostic indicators, ultimately biasing estimates of hospitalization, survival, and treatment effectiveness.

Conventional statistical imputation approaches, such as mean/median substitution, Multiple Imputation by Chained Equations (MICE), and Random Forest–based methods like MissForest, provide well-established frameworks for estimating missing values under explicit assumptions. These methods perform well when relationships among variables are relatively stable and low-dimensional but struggle when dependencies are complex, nonlinear, or context-dependent. Moreover, they cannot leverage unstructured clinical context—such as notes, temporal trajectories, or treatment pathways—that may implicitly inform missing values. These limitations motivate the exploration of language-model–based approaches that integrate structured and unstructured signals. Large Language Models (LLMs) offer a fundamentally different advantage for missing data imputation: they perform contextual inference by integrating variable distributions, clinical semantics, and patient-level patterns learned from large medical corpora (2, 3).

Accurate clinical interpretation of statistical outputs is critical for translating analytic findings into patient care decisions. Clinicians depend on summary tables, risk models, survival curves, and treatment effect estimates to guide therapy choices and stratify risk, yet these outputs are often misunderstood when statistical assumptions, effect sizes, or confidence intervals are misinterpreted. Large Language Models can act as interpretive assistants by summarizing complex results, clarifying prognostic implications, and highlighting clinically relevant patterns. By reducing cognitive burden and providing consistent, context-aware explanations, LLMs may help bridge the gap between advanced analytics and real-world decision-making, enhancing transparency and the practical usefulness of RWD-based insights.

In this study, we used a Large Language Model (Claude Sonnet 4.5) to perform imputation of 7 continuous and 2 categorical variables from a real-world dataset (MIMIC IV) of ICU admitted patients with cardiac arrest. We compared these imputed results with those from two statistical methods - MICE and missForest. Across continuous variables, LLM-based imputation and MissForest demonstrated comparable performance, with both methods maintaining stable MAE and NRMSE across missingness levels. In contrast, MICE consistently showed higher error and was more sensitive to increases in missingness, particularly beyond moderate levels. For categorical variables, imputation accuracy varied across methods but showed no consistently superior model, and accuracy remained relatively stable across missingness levels. Bias remained low overall, though the spread of bias increased substantially with higher missingness, indicating reduced stability across all methods at extreme data loss. Sample size had minimal influence on continuous-variable NRMSE, and MICE remained the least robust method irrespective of sample size. Categorical accuracy was likewise insensitive to sample size, with no method exhibiting persistent advantage. Importantly, the LLM demonstrated an ability to capture nonlinear relationships between variables, performing similarly to tree-based models like MissForest in settings where such structure is present. Few-shot and zero-shot prompting yielded nearly identical imputation performance, suggesting limited reliance on handcrafted examples. Overall, LLMs and MissForest emerged as the most reliable and stable imputation methods, while MICE performed weakest across nearly all conditions.

We prompted the LLM to interpret and generate clinical summaries of a range of RWE outputs including baseline tables and Kaplan–Meier curves. The interpretations and summaries were qualitatively evaluated by a clinician and statistician with a custom developed rubric. Overall, the summaries were reasonably accurate and free of hallucinations, but they could be further improved with engineering prompts and few-shot training.

Methods

This study utilized data extracted from the MIMIC-IV (Medical Information Mart for Intensive Care IV) database (4). The MIMIC-IV database is an open and publicly available database that contains high-quality data between 2008 and 2019 constructed by Institutional Review Boards of the Massachusetts Institute of Technology (MIT, Cambridge, MA, America) and Beth Israel Deaconess Medical Center. Data was accessed using a custom script written in Google Big query.

Study Patients

Patients with a diagnosis of cardiovascular arrest (CA), defined as ICD-9 codes of 4275 or ICD-10 codes of I46, I462, I468 and I469, and greater than18 years old at the time of ICU admission were included in the study. The data set was filtered to include patients with complete records across several categories of variables. Physiological variables, which tracked immediate patient status, included mean blood pressure (mmHg), heart rate (beats/min), respiratory rate (breaths/min), temperature (°C), and oxygen saturation (%). Comorbidities included hypertension, congestive heart failure, myocardial infarction, diabetes mellitus, and chronic obstructive pulmonary disease. Laboratory values included lactate (mmol/L), sodium (mmol/L), chloride (mmol/L), bicarbonate (mmol/L), creatinine (mg/dL), glucose (mg/dL), hemoglobin (g/dL), and platelet count (cells/µL). Finally, the outcome variable tracked whether the patient experienced in-hospital death. Along with these, the admission and discharge dates/times for each patient were also extracted. This resulted in a dataset with 1200 patients.

Data missingness and imputation- To simulate missingness, values within each data column/variable were randomly masked such that the total number of missing values per column was 10%, 30%, 50%, 70%, and 90%. Similarly, to vary the dataset size, patient records were randomly selected to create 5 datasets with 60, 100, 300, 600 and 1200 records. Eight variables selected for imputation include respiratory rate, heart rate, systolic blood pressure, diastolic blood pressure, hypertension, congestive heart failure, lactate, sodium, and oxygen saturation.
To impute missing values, R packages mice (5) and missForest (6) were used with their default parameter values. Anthropic's LLM model - Claude 4.5 (Sonnet) was used to impute missing values. The paws R package was used to access the LLM model via AWS bedrock.

RWE output summary evaluation metric -

LLM-generated descriptive summaries of RWE baseline tables were evaluated using a structured scoring rubric assessing six domains: factual accuracy, completeness, interpretation quality, clarity of communication, clinical appropriateness, and absence of hallucination. Each domain was scored from 0–5, yielding a total score (0–30) representing overall summary quality.

1. Factual accuracy - Assesses whether the LLM correctly interprets the numerical and categorical information in the summary table.
2. Coverage and completeness - Evaluates whether the summary includes all major categories and clinically important variables.
3. Interpretation quality - Measures the *clinical reasoning* quality and ability to highlight meaningful patterns.
4. Clarity and communication - Evaluates writing quality, organization, and readability.
5. Clinical appropriateness - Measures whether statements are consistent with accepted clinical knowledge and avoid over-interpretation.

6. Absence of hallucinations - Did the LLM introduce variables, columns, or interpretations not present in the table?

Similarly, a scoring rubric was constructed to evaluate KM plots.

1. Factual accuracy - Evaluates whether the LLM accurately explains what is visible in the Kaplan–Meier plot or survival function.
2. Statistical accuracy (HR, p-values, medians) - Assesses whether the LLM correctly interprets numeric survival estimates (if provided).
3. Understanding of censoring and risk tables - Measures ability to correctly explain censoring patterns and risk table.
4. Clinical interpretation and relevance - Evaluates whether the LLM appropriately contextualizes observed survival patterns.
5. Completeness and coverage of key survival aspects - Assesses whether the LLM addresses the full scope of survival analysis elements.
6. Absence of hallucinations - Evaluates whether the LLM sticks to what is shown and avoids inventing numbers or variables.

Results

LLM and MissForest show similar imputation performance for continuous variables, with MICE more sensitive to missingness

To evaluate the imputation performance of MICE, MissForest, and Claude 4.5 (Sonnet), we selected eight variables from the database including diastolic BP, heart rate, lactate, respiratory rate, sodium, peripheral oxygen saturation, and systolic BP. Values for each variable with the dataset were randomly masked such that each variable has a specified number of values (percentage) masked for imputation.

Across both the mean absolute error (MAE) trends (Figure 1) and normalized root mean squared error (NRMSE) distributions (Figure 2), all three imputation methods show increasing error as missingness rises, which is expected as more information is removed from the data. The MAE curves for Claude 4.5 (Sonnet) and MissForest often follow similar trajectories across variables, suggesting broadly comparable performance as missingness increases. In contrast, MICE tends to show larger increases in MAE at higher missingness levels particularly for heart rate, respiratory rate, and systolic blood pressure indicating greater sensitivity to incomplete data.

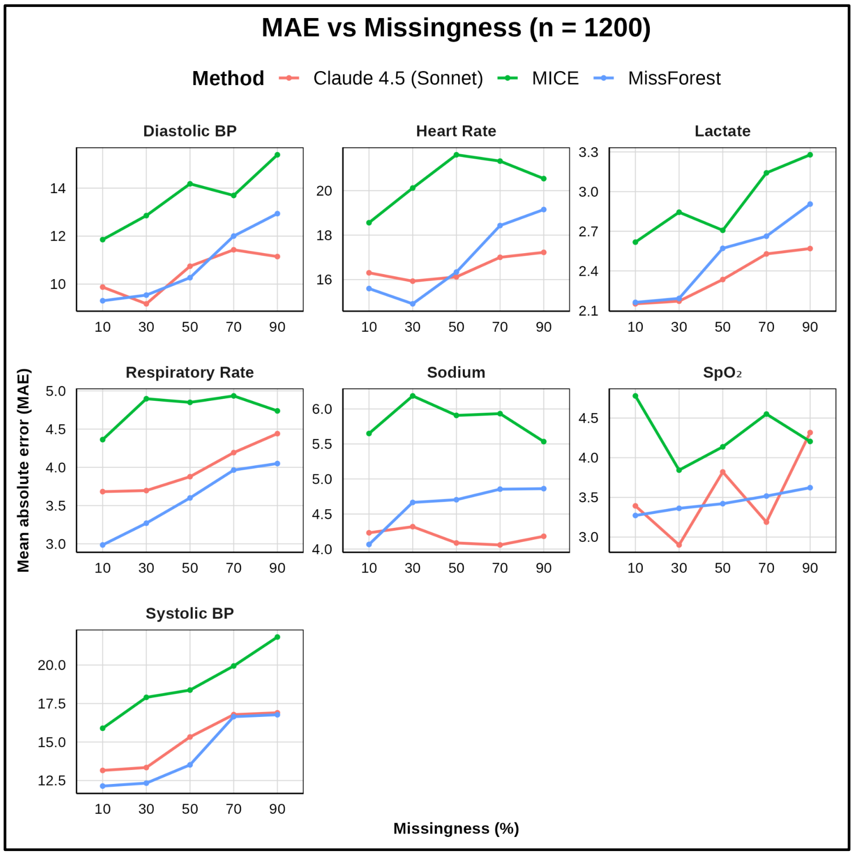


Figure 1. Mean absolute error (MAE) for three imputation methods—Claude 4.5 (Sonnet), MICE, and MissForest for continuous variables across increasing levels of missingness from 10% to 90% in a dataset with 1200 records.

This pattern is consistent with the NRMSE boxplots, where MICE generally exhibits higher error distributions, while Claude 4.5 and MissForest cluster more closely with lower and more stable median values. The closeness of the Claude and MissForest curves, along with their tighter NRMSE spreads, suggests that these approaches may better accommodate nonlinear relationships in the data compared with MICE, which relies on more restrictive and linear modeling assumptions. For variables with lower variability, such as lactate and sodium, all three methods behave more similarly showing smaller overall differences in performance. Taken together, the results indicate that the LLM-based method and MissForest maintain more stable imputation accuracy across a range of missingness levels, whereas MICE shows a more pronounced decline as missingness increases.

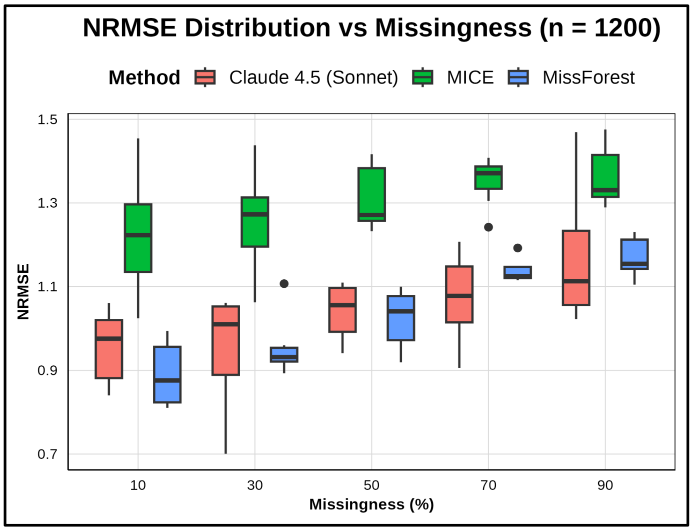


Figure 2. Normalized root mean square error (NRMSE) for three imputation methods - Claude 4.5 (Sonnet), MICE, and MissForest across increasing levels of missingness from 10% to 90% in a dataset with 1200 records.

Categorical imputation accuracy varies by method and missingness level

Figure 3 shows that categorical imputation accuracy declines with increasing missingness, but the pattern varies across methods and variables. For both congestive heart failure and hypertension, each method performs better at some missingness levels and worse at others. Claude, MICE, and MissForest all show fluctuating accuracy, with no method consistently outperforming the others. Overall, the results indicate that categorical imputation performance is variable and method-dependent, with no clear method outperforming the other.

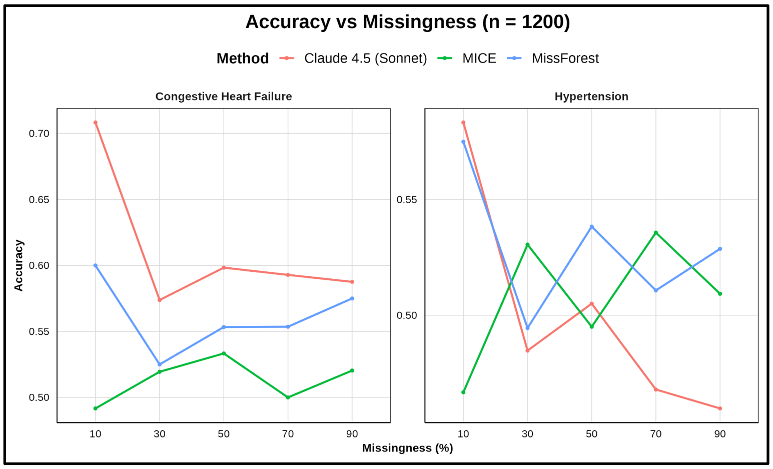


Figure 3. Accuracy for three imputation methods - Claude 4.5 (Sonnet), MICE, and MissForest across increasing levels of missingness from 10% to 90% in a dataset with 1200 records.

All methods show low median bias, but the spread of bias grows substantially as missingness increases.

Bias reflects the average difference between imputed values and the true values, indicating whether a method systematically over or under estimates missing data.

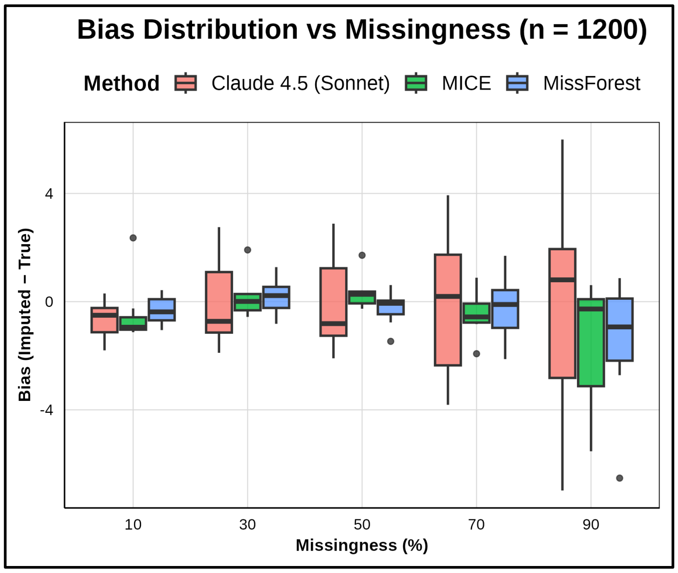


Figure 4. Bias distribution three imputation methods - Claude 4.5 (Sonnet), MICE, and MissForest across increasing levels of missingness from 10% to 90% in a dataset with 1200 records.

The distribution of imputation bias for Claude 4.5 (Sonnet), MICE, and MissForest across increasing levels of missingness. For all methods, median bias remained close to zero across missingness levels, indicating that none of the approaches consistently over- or under-estimated values. However, the spread of bias increased as missingness rose, particularly at 70% and 90%, suggesting greater variability in estimates when more data are absent. Claude 4.5 exhibited wider bias ranges at higher missingness levels, while MICE and MissForest generally show tighter distributions except for occasional outliers. No method demonstrated consistently lower bias across all levels of missingness, and variation appeared to depend on both method and missingness levels. Overall, the figure suggests that while central bias is generally small, imputation uncertainty increases substantially as missingness grows.

Sample size has minimal impact on NRMSE and MICE consistently produces higher errors than other methods.

To assess whether larger datasets meaningfully reduce imputation error or alter relative method performance across missingness levels, we assessed the imputation performance across datasets containing 60, 100, 300, 600, and 1200 patient records.

Across all missingness percentages (Figure 5), NRMSE remained relatively stable as sample size increases, indicating that within this range, data size had limited impact on overall imputation accuracy. MICE consistently produced higher NRMSE than the other methods at all sample sizes and missingness levels, reflecting greater error regardless of cohort size. Claude 4.5 (Sonnet) and MissForest showed similar NRMSE patterns, with lower and more stable error across sample sizes. At higher missingness levels (70% and 90%), MissForest showed modest variability, but overall trends remained consistent.

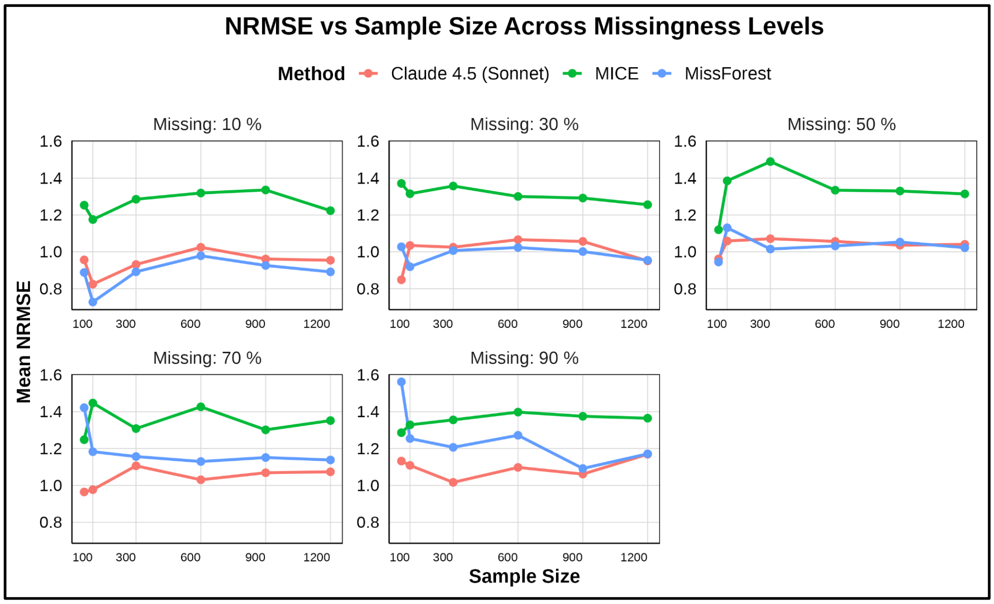


Figure 5. Mean Normalized root mean square (NRMSE) for continuous variables with varying missingness (10%, 30%, 50%, 70%, 90 %) across sample size with 60, 100, 300, 600, 1200 patient records

Categorical imputation accuracy is largely unaffected by sample size, with no consistent best-performing method.

Accuracy remained relatively stable from sample sizes of 60 to 1200 for all three methods, suggesting that sample size had minimal impact within this range. Claude 4.5 (Sonnet), MICE, and MissForest each performed better in some conditions and worse in others, with no method consistently outperforming the others. At higher missingness levels, accuracy values converged further, confirming the absence of clear method-specific advantages. Overall, sample size did not appear to meaningfully influence categorical imputation accuracy across the scenarios tested.

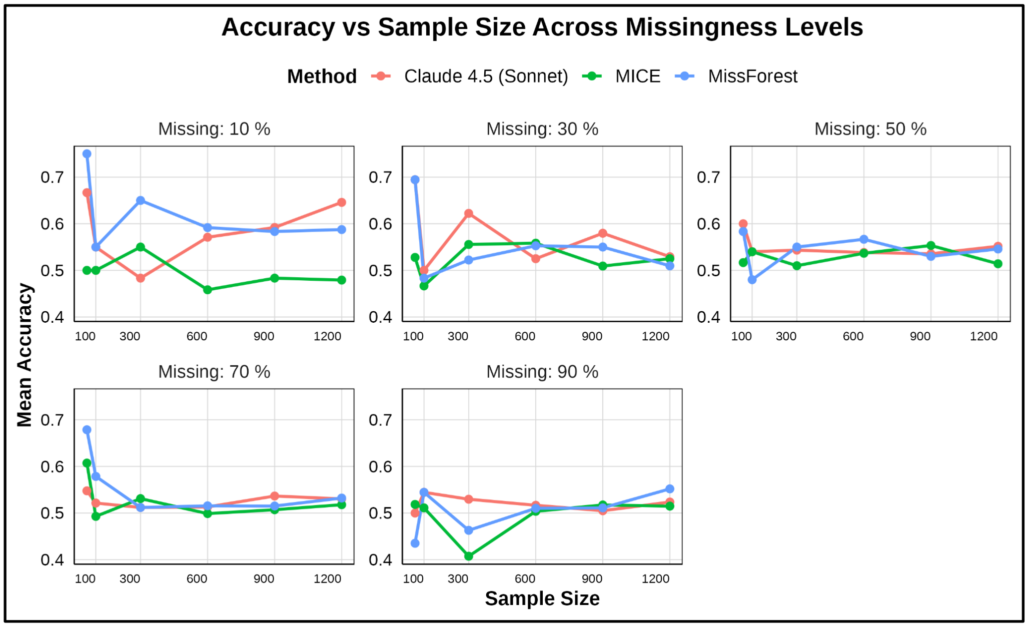


Figure 6. Mean accuracy for categorical variables with varying missingness (10%, 30%, 50%, 70%, 90 %) across sample size with 60, 100, 300, 600, 1200 patient records.

Few-shot and zero-shot prompting yield similar imputation performance across missingness levels

Zero-shot prompting refers to asking the LLM to perform imputation without providing any example responses, whereas few-shot prompting supplies one or more example imputations to guide the model's behavior. This comparison was conducted to assess whether providing examples improves the LLM's accuracy or stability when imputing clinical variables under moderate (30%) and high (70%) missingness.

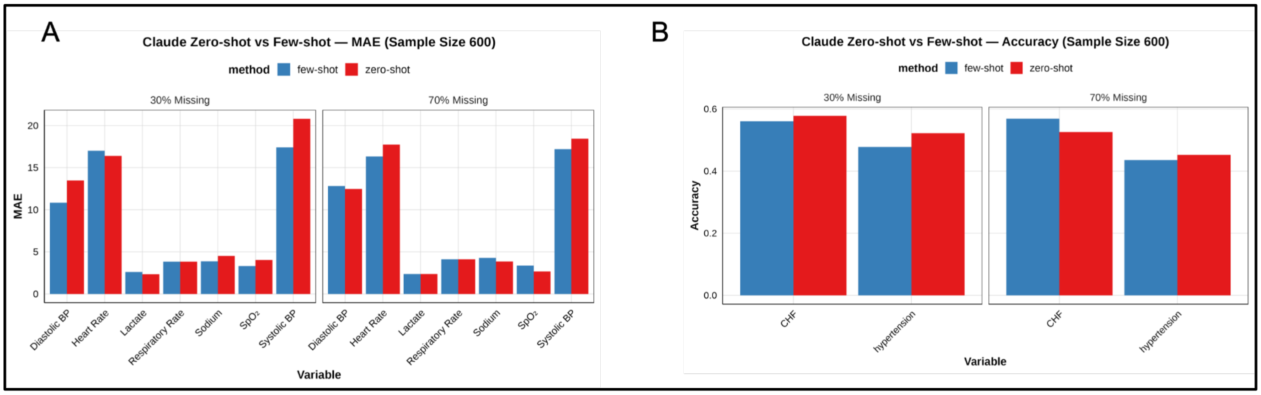


Figure 7. A. Mean absolute error (MAE) for continuous variables and (B) Accuracy for categorical variables are shown for few and zero-shot prompting in a dataset with 600 patient records with a missingness of 30% and 70% for each variable. 5 complete patient records were used for few-shot prompting.

Across continuous variables, few-shot prompting generally resulted in slightly lower MAE than zero-shot prompting, although differences were small and varied by variable. For variables such as diastolic BP, heart rate, and systolic BP, few-shot prompting showed modest improvements, while for others (e.g., lactate, respiratory rate, sodium), both approaches produced nearly identical errors. For categorical variables, accuracy differences between few-shot and zero-shot prompting were also minor, with each approach performing slightly better in different scenarios. At both 30% and 70% missingness, accuracy remained similar across prompting strategies, and neither method consistently outperformed the other.

We also investigated if few-shot prompting had an impact on the imputation bias. Figure 8A shows that both zero-shot and few-shot prompting produced similar patterns of bias across clinical variables at 30% and 70% missingness, with no consistent advantage for either method. At moderate missingness, small reductions in bias for some variables under few-shot prompting were offset by increases in others, such as systolic blood pressure. As missingness increased to 70%, bias grew for both approaches, indicating reduced stability under sparse data conditions.

Further, to validate the reproducibility of imputation, we conducted 5 independent runs of LLM based imputations. 600 Patient records were randomly sampled and LLM based imputation was performed at 30% missingness. Figure 8B shows that mean absolute error at 30% missingness was closely aligned between zero-shot and few-shot prompting across all variables. Any small MAE differences fell within the variability observed across repeated LLM runs and were not systematic. Few-shot prompting did not provide a substantial or consistent improvement over zero-shot prompting for numerical imputation in this setting.

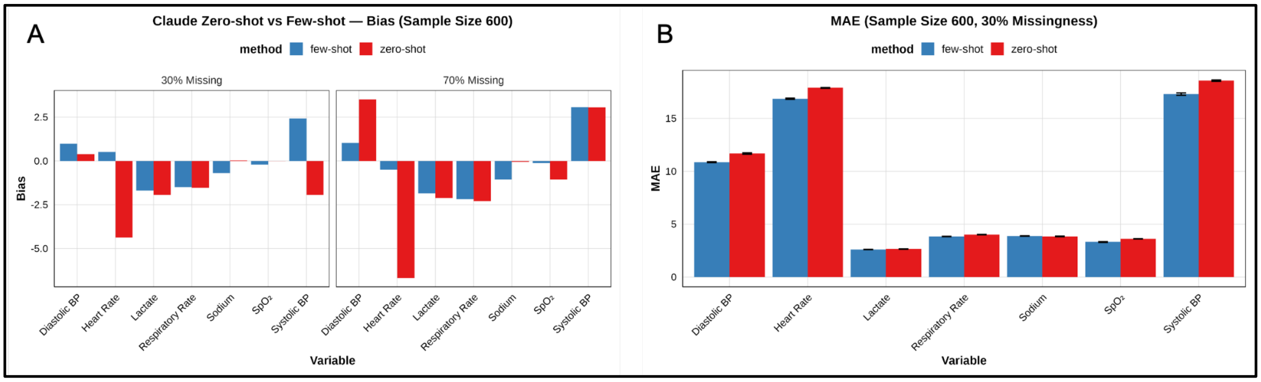


Figure 8. Bias and accuracy of LLM-based numerical imputation under zero-shot and few-shot prompting.(A) shows the imputation bias (imputed minus true values) for key clinical variables at 30% and 70% missingness for zero-shot and few-shot Claude prompting. (B) shows the mean absolute error (MAE) at 30% missingness, estimated across multiple repeated LLM imputations to capture run-to-run variability.

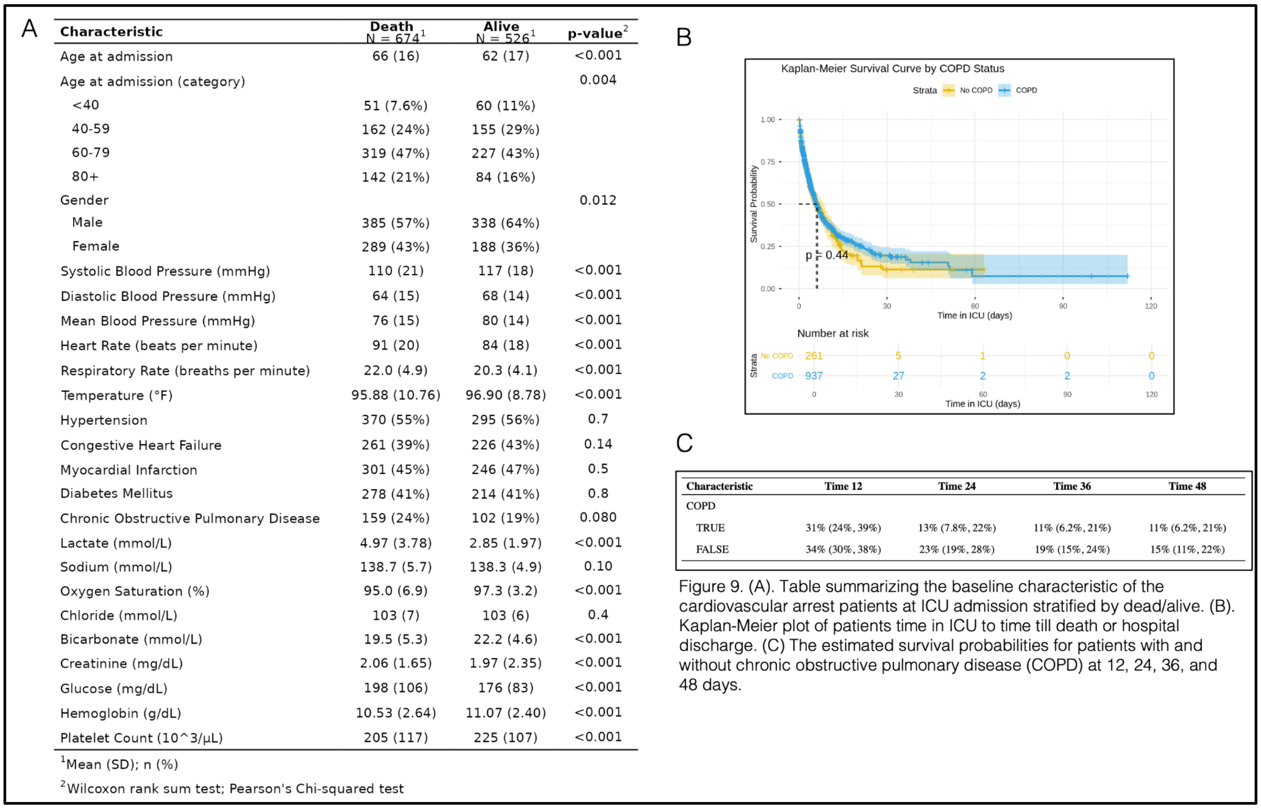
Overall, the analysis suggests that providing example imputations (few-shot) offers only limited and variable benefit over zero-shot prompting, and both strategies deliver comparable performance under the tested conditions.

Generating clinical summaries from RWE outputs

To evaluate the capabilities of LLMs in interpreting RWE outputs and generating accurate clinical summaries, we prompted Claude Sonnet 4.5 to summarise a baseline characteristic table (Figure 9A) and interpret a survival curve plot (Figure 9B and 9C). Two separate prompts were constructed to interpret and generate clinical summaries - one for baseline characteristic table (9A) and another one, where the KM plot and table containing survival probabilities of cardiovascular arrest patients were included.

The prompt instructed Claude Sonnet 4.5 to act as a clinical data analyst and to generate a structured, clinical interpretation of a demographic and clinical summary table. Specifically, the model was asked to summarize patient characteristics, laboratory values, and comorbidities; describe distributions, central tendency, and variability for continuous variables; and interpret counts and percentages for categorical variables. It was also directed to identify notable imbalances or skewed distributions and to comment on their potential clinical relevance, including reference to the existing literature where appropriate. Importantly, the prompt constrained the model to rely only on variables present in the provided table and to avoid introducing unsupported information. The output was required to be written in clear, formal prose suitable for inclusion in a medical or statistical report.

The summaries were assessed by 5 distinct categories and graded from 1 (lowest) - 5 (highest) for each category as described in the methods section. The assessment is for baseline characteristic table summary is shown in Table 1 and for survival curve plot in Table 2. The clinical summary for baseline characteristics table scored 4 on coverage and completeness, interpretation quality, clarity of communication, and absence of hallucinations, but 3 on factual accuracy. The factual accuracy was below other criteria as the summary contained references to literature, which were loosely consistent with the results. The summary for KM and survival probabilities table scored 3 for factual accuracy, interpretation quality, clarity and communication as multiple objects (plot and table) and its interconnectedness may have reduced its overall quality of interpretation.



Characteristic	Death N = 674[1]	Alive N = 526[1]	p-value[2]
Age at admission	66 (16)	62 (17)	<0.001
Age at admission (category)			0.004
<40	51 (7.6%)	60 (11%)	
40-59	162 (24%)	155 (29%)	
60-79	319 (47%)	227 (43%)	
80+	142 (21%)	84 (16%)	
Gender			0.012
Male	385 (57%)	338 (64%)	
Female	289 (43%)	188 (36%)	
Systolic Blood Pressure (mmHg)	110 (21)	117 (18)	<0.001
Diastolic Blood Pressure (mmHg)	64 (15)	68 (14)	<0.001
Mean Blood Pressure (mmHg)	76 (15)	80 (14)	<0.001
Heart Rate (beats per minute)	91 (20)	84 (18)	<0.001
Respiratory Rate (breaths per minute)	22.0 (4.9)	20.3 (4.1)	<0.001
Temperature (°F)	95.88 (10.76)	96.90 (8.78)	<0.001
Hypertension	370 (55%)	295 (56%)	0.7
Congestive Heart Failure	261 (39%)	226 (43%)	0.14
Myocardial Infarction	301 (45%)	246 (47%)	0.5
Diabetes Mellitus	278 (41%)	214 (41%)	0.8
Chronic Obstructive Pulmonary Disease	159 (24%)	102 (19%)	0.080
Lactate (mmol/L)	4.97 (3.78)	2.85 (1.97)	<0.001
Sodium (mmol/L)	138.7 (5.7)	138.3 (4.9)	0.10
Oxygen Saturation (%)	95.0 (6.9)	97.3 (3.2)	<0.001
Chloride (mmol/L)	103 (7)	103 (6)	0.4
Bicarbonate (mmol/L)	19.5 (5.3)	22.2 (4.6)	<0.001
Creatinine (mg/dL)	2.06 (1.65)	1.97 (2.35)	<0.001
Glucose (mg/dL)	198 (106)	176 (83)	<0.001
Hemoglobin (g/dL)	10.53 (2.64)	11.07 (2.40)	<0.001
Platelet Count (10^3/µL)	205 (117)	225 (107)	<0.001

[1] Mean (SD); n (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test

Characteristic	Time 12	Time 24	Time 36	Time 48
COPD				
TRUE	31% (24%, 39%)	13% (7.8%, 22%)	11% (6.2%, 21%)	11% (6.2%, 21%)
FALSE	34% (30%, 38%)	23% (19%, 28%)	19% (15%, 24%)	15% (11%, 22%)

Figure 9. (A). Table summarizing the baseline characteristic of the cardiovascular arrest patients at ICU admission stratified by dead/alive. (B). Kaplan-Meier plot of patients time in ICU to time till death or hospital discharge. (C) The estimated survival probabilities for patients with and without chronic obstructive pulmonary disease (COPD) at 12, 24, 36, and 48 days.

Figure 9. The baseline characteristics table (A), KM plot (B), and survival probabilities (C) for cardiovascular arrest patients.

Metric	Score	Score
	Baseline characteristics table	**KM plot and probabilities table**
Factual accuracy	3	3
Coverage and completeness	4	4
Interpretation quality	4	3
Clarity and communication	4	3
Clinical Appropriateness	3	4
Absence of Hallucinations	4	4

Table 1. Evaluation of baseline characteristic table summary and survival curve plot.

Conclusions

This study demonstrates that Large Language Models (LLMs) can serve as effective tools for both missing data imputation and interpretation of real-world evidence outputs. Across continuous variables, LLM-based imputation showed performance comparable to tree-based methods such as MissForest and consistently outperformed MICE, particularly at higher levels of missingness. While categorical imputation accuracy varied across methods, no single approach demonstrated persistent superiority, and performance remained relatively stable across sample sizes. Although median bias remained low for all methods, increased variability at extreme missingness highlights the inherent uncertainty of imputation under sparse data conditions. Few-shot prompting provided minimal and inconsistent benefit over zero-shot prompting, suggesting that LLM performance is largely driven by learned clinical context rather than handcrafted examples. Importantly, LLM-generated clinical summaries of baseline tables and survival analyses were judged to be mostly accurate and clinically appropriate, and free of hallucinations, supporting their potential role as interpretive aids. Overall, these findings suggest that LLMs offer a promising complementary approach for enhancing both data completeness and interpretability in RWD-based clinical research, while emphasizing the need for careful validation in high-missingness settings.

References

1. Gurupur V, Hooshmand S, Prabhu DF, Trader E, Salvi S. Incompleteness of Electronic Health Records: An Impending Process Problem Within Healthcare. Healthcare. 2025; 13(22):2900
2. Nazir, A., Cheeema, M.N., Wang, Z.: ChatGPT-based biological and psychological data imputation. Meta-radiology 1(3), 100034 (2023)
3. Masood, S., Al Bashrawi, M.A., Khan, M.A. *et al.* Exploring ChatGPT applications in healthcare: a comprehensive overview. *Vis Comput* 41, 4893–4914 (2025)
4. Johnson, A.E.W., Bulgarelli, L., Shen, L. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 10, 1 (2023). https://doi.org/10.1038/s41597-022-01899-x
5. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03
6. Stekhoven DJ, Bühlmann P (2012). “MissForest: nonparametric missing value imputation for mixed-type data.” *Bioinformatics*, **28**(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597.
7. OpenAI. (2025). ChatGPT (GPT-4o, 2 July version) [Large language model]. https://chat.openai.com

Supplementary Information-
The code used to extract MIMIC IV data and conduct analysis is upload on github. The link to access the code - https://github.com/rahusomavanshi/phuse_apac_connect2026