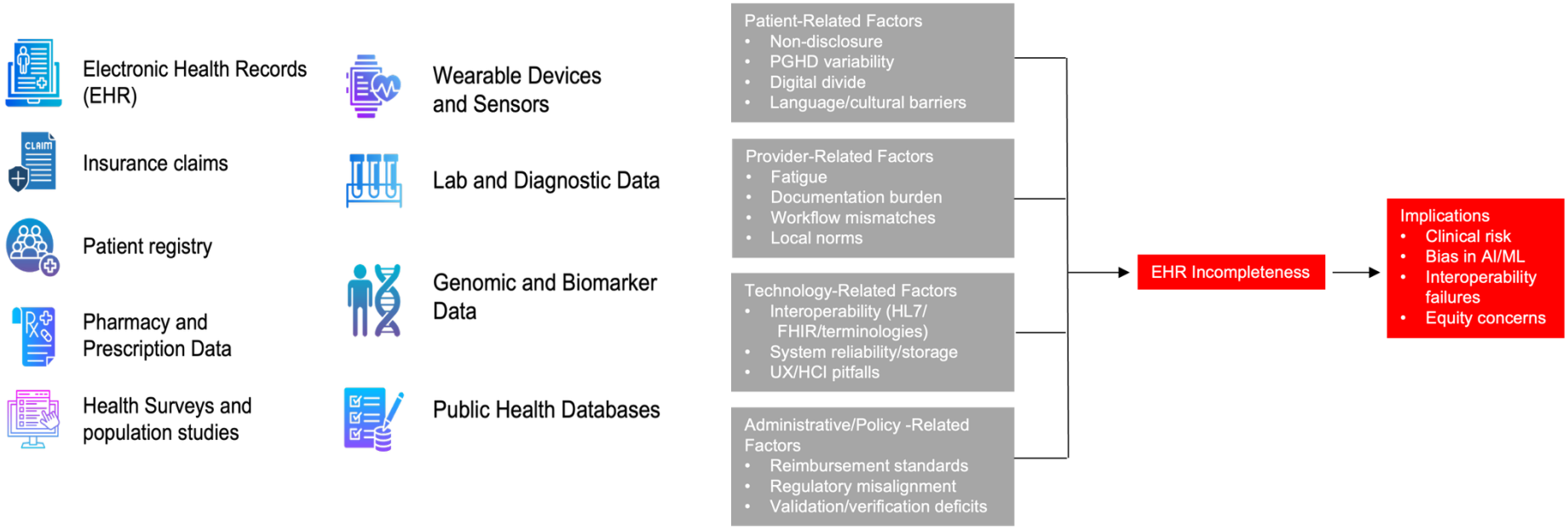
Leveraging Large Language Model (LLM) for Missing Data Imputation and Interpretation of Real-World Evidence Outputs.

Phuse APAC connect – Hyderabad, India
21 February 2026

Rahul Somavanshi &
Spencer Langerman

Incomplete RWD, incomplete care: data gaps fuel risk, errors, and unequal outcomes.

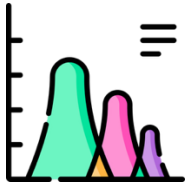


(Adapted from Gurupur et al 2025)

From statistical to AI-based imputation: evolving ways to recover missing RWD.

Simple deterministic methods

- Mean/Median
- Mode
- Last Observation Carried Forward (LOCF)



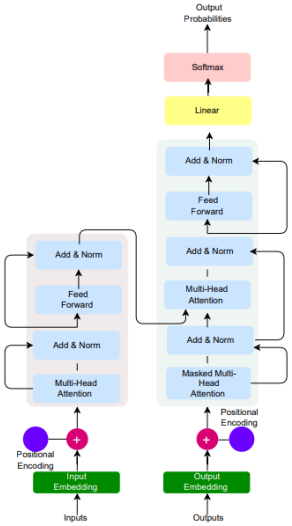
Statistical model–based

- Multiple imputation (MICE)
- Machine learning –
  - KNN
  - MissForest
  - XGBoost

Large Language models (LLMs)

GPT – Generative Pre-trained Transformer

Encoder | Decoder



Translation (Encoder – Decoder)

ప్యూజ్ ఏపీఏసి కనెక్ట్ అద్భుతం!

↓

Phuse APAC Connect is amazing!

Next word prediction (Decoder only)

Cat sat on the mat.

Cat sat on the mat.

Cat sat on the mat.

Cat sat on the mat.

Transformers: The Architecture Powering Modern LLMs

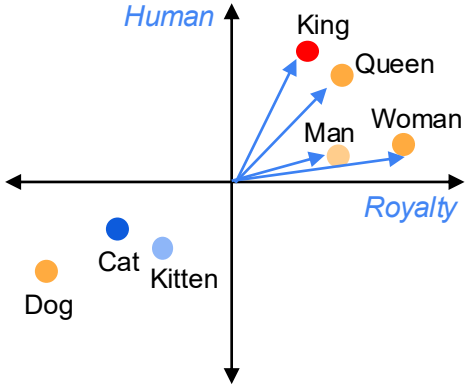
Tokenization

The cat sat on the mat.

The	cat	sat	on	the	mat
976	9059	10139	402	290	2450

Word embeddings

	Cat	Dog	Kitten	King	Queen	Man	Woman
Living being	0.6	0.7	0.5	0.7	0.7	0.6	0.7
Feline	0.6	-0.1	0.8	0.4	0.4	-0.2	0.4
Human	0.1	0.4	-0.1	0.7	0.7	0.8	0.9
Gender	0.4	0.3	0.2	0.8	0.8	0.9	-0.7
Royalty	-0.7	-0.4	-0.6	0.9	0.9	-0.1	0.1
Plural	-0.2	-0.3	-0.1	-0.6	-0.6	-0.7	-0.4



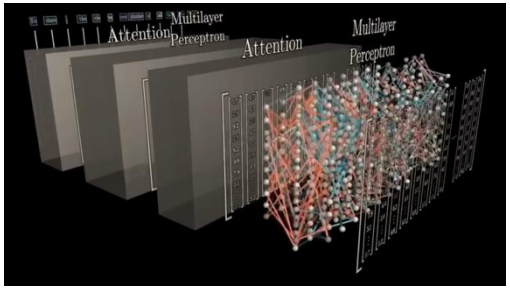
The cat sat on the mat.
The mat sat on the cat.

Positional embeddings – order matters !

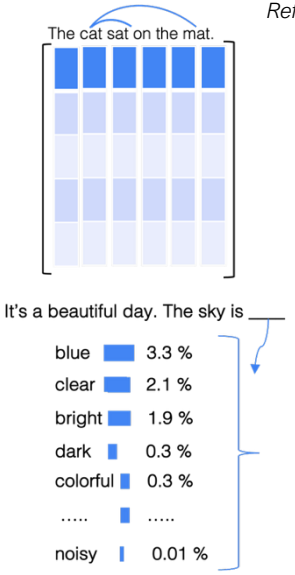
+

Word embeddings

Attention layer



Ref:3Blue1brown

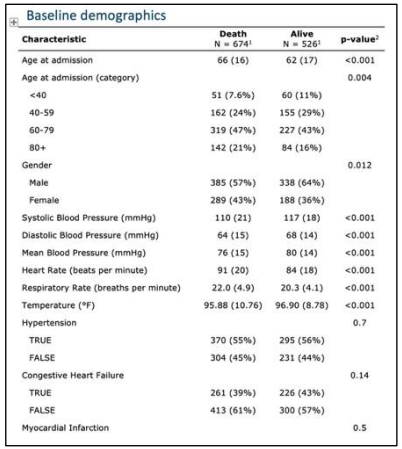


Using LLMs to impute missing clinical data and generate RWE summaries

Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable N
10	8	?	4.4	32	?
15	7	8	4.6	43	45
?	?	5	?	28	68
22	15	?	4.0	?	53
16	?	7.9	3.9	67.5	?
?	6	6.3	?	?	37

LLM

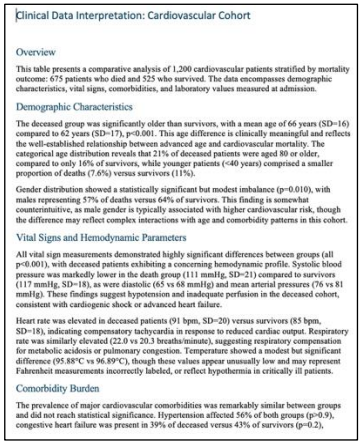
Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable N
10	8	6	4.4	32	?
15	7	8	4.6	43	45
?	10	5	?	28	68
22	15	?	4.0	32	53
16	7	7.9	3.9	67.5	?
?	6	6.3	4.2	45	37



Baseline demographics

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			0.7
TRUE	370 (55%)	295 (56%)	
FALSE	304 (45%)	231 (44%)	
Congestive Heart Failure			0.14
TRUE	261 (39%)	226 (43%)	
FALSE	413 (61%)	300 (57%)	
Myocardial Infarction			0.5

LLM



Clinical Data Interpretation: Cardiovascular Cohort

Overview

This table presents a comparative analysis of 1,200 cardiovascular patients stratified by mortality outcome: 675 patients who died and 525 who survived. The data encompasses demographic characteristics, vital signs, comorbidities, and laboratory values measured at admission.

Demographic Characteristics

The deceased group was significantly older than survivors, with a mean age of 66 years (SD=16) compared to 62 years (SD=17), p<0.001. This age difference is clinically meaningful and reflects the well-established relationship between advanced age and cardiovascular mortality. The categorical age distribution reveals that 21% of deceased patients were aged 80 or older, compared to only 16% of survivors, while younger patients (<40 years) comprised a smaller proportion of deaths (7.6%) versus survivors (11%).

Gender distribution showed a statistically significant but modest imbalance (p=0.010), with males representing 57% of deaths versus 64% of survivors. This finding is somewhat counterintuitive, as male gender is typically associated with higher cardiovascular risk, though the difference may reflect complex interactions with age and comorbidity patterns in this cohort.

Vital Signs and Hemodynamic Parameters

All vital sign measurements demonstrated highly significant differences between groups (all p<0.001), with deceased patients exhibiting a concerning hemodynamic profile. Systolic blood pressure was markedly lower in the death group (111 mmHg, SD=21) compared to survivors (117 mmHg, SD=18), as were diastolic (65 vs 68 mmHg) and mean arterial pressures (76 vs 81 mmHg). These findings suggest hypotension and inadequate perfusion in the deceased cohort, consistent with cardiogenic shock or advanced heart failure.

Heart rate was elevated in deceased patients (91 bpm, SD=20) versus survivors (85 bpm, SD=18), indicating compensatory tachycardia in response to reduced cardiac output. Respiratory rate was similarly elevated (22.0 vs 20.3 breaths/minute), suggesting respiratory compensation for metabolic acidosis or pulmonary congestion. Temperature showed a modest but significant difference (95.88°C vs 96.89°C), though these values appear unusually low and may represent Fahrenheit measurements incorrectly labeled, or reflect hypothermia in critically ill patients.

Comorbidity Burden

The prevalence of major cardiovascular comorbidities was remarkably similar between groups and did not reach statistical significance. Hypertension affected 56% of both groups (p>0.9), congestive heart failure was present in 39% of deceased versus 43% of survivors (p=0.2),

Real-World Dataset and LLM-Enabled R Analytics Environment

MIMIC-IV – Cardiovascular arrest (CA)
1200 patients dataset

Variables	Missingness
Systolic BP	10
Diastolic BP	30
Respiratory rate	50
Heart rate	70
SpO2	90
Lactate	
Sodium	
CHF	
Hypertension	

mice

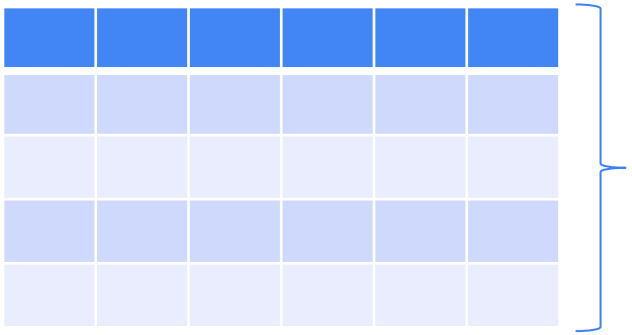
missForest

Claude Sonnet 4.5

Amazon Bedrock

Paws

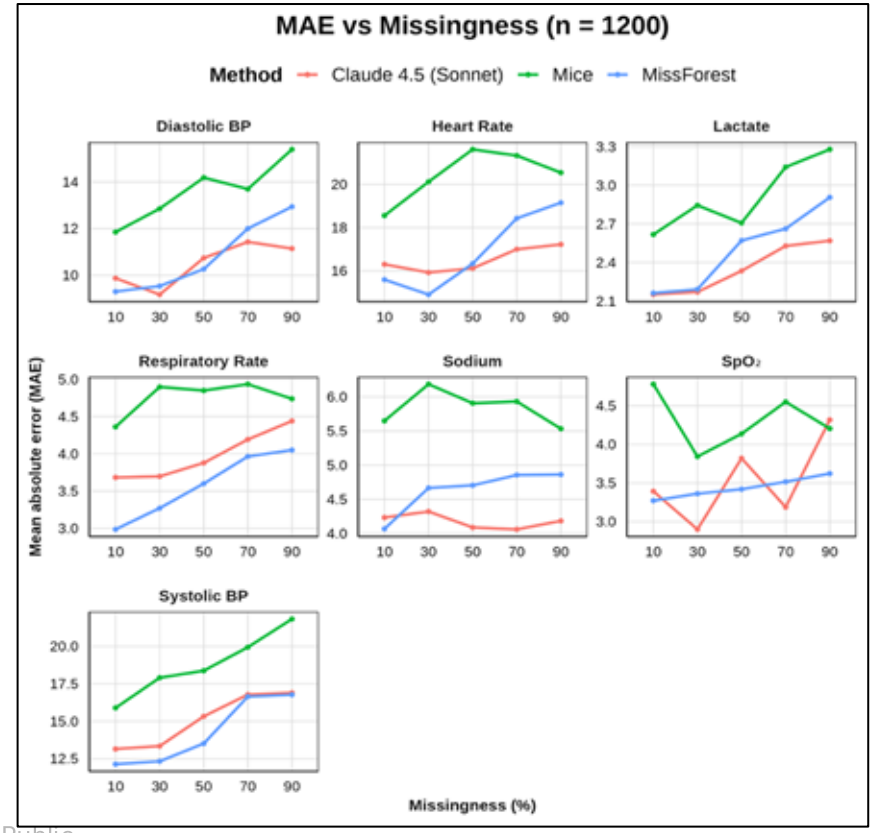
LLM prompt – Batches of 25 records JSON

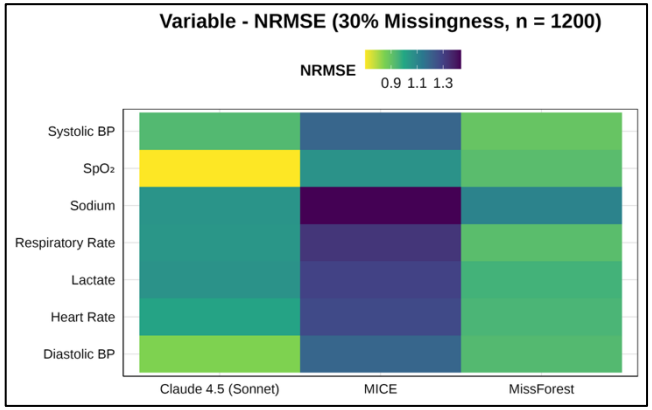


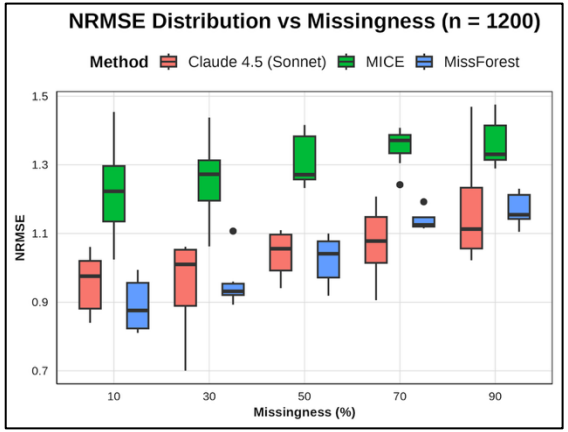
Prompt –

- *Clinical data analyst specializing in cardiovascular and critical care.*
- *Use standard clinical reasoning, relationships among other patient features, and population norms for cardiovascular or critical care patients.*
- *Use imputed values of one variable to inform others when appropriate.*

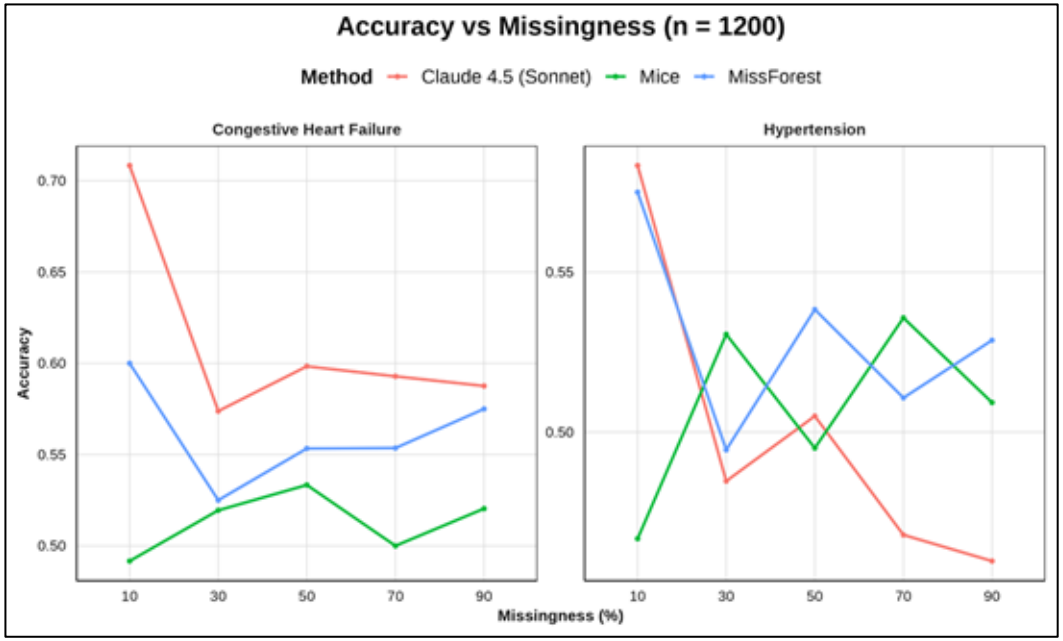
LLM and MissForest show similar imputation performance for continuous variables, with MICE more sensitive to missingness.

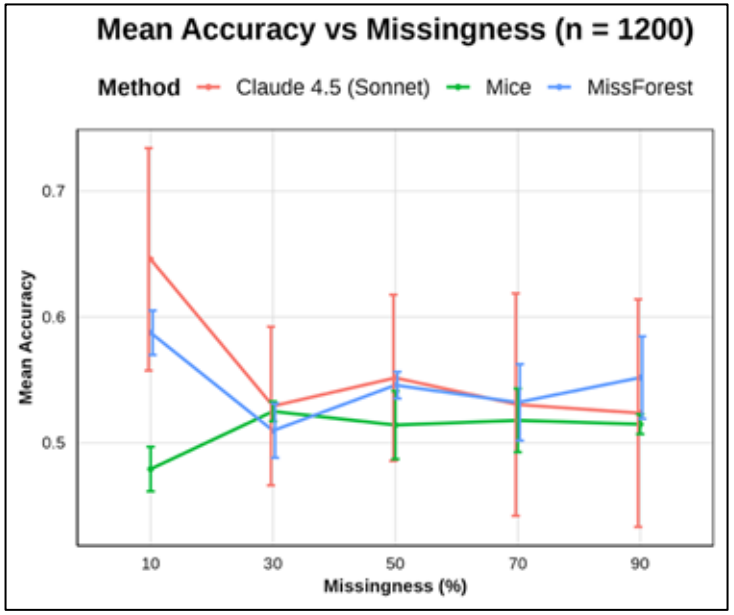




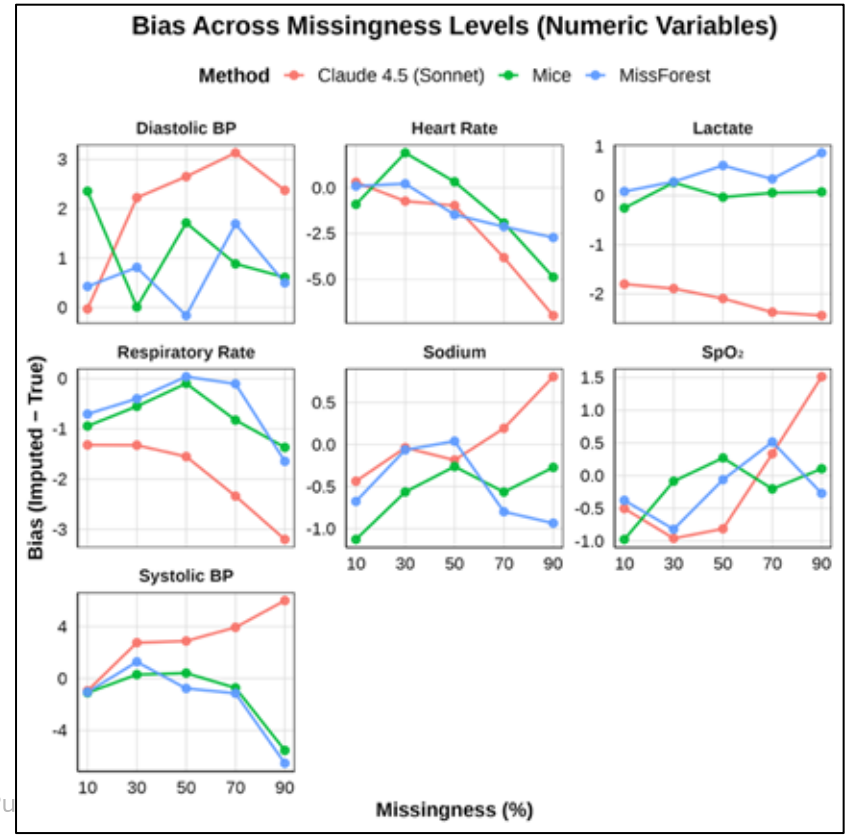


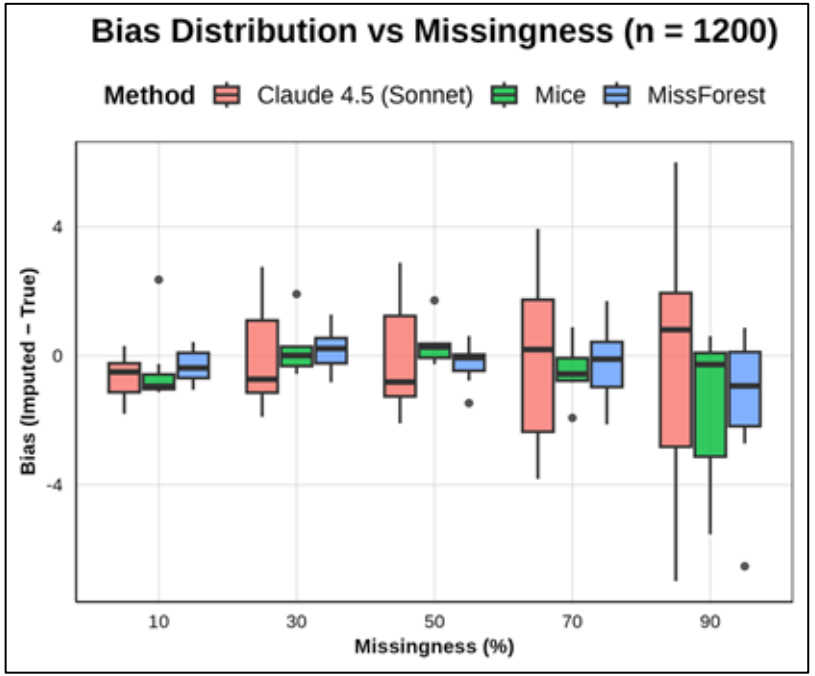
Categorical imputation accuracy varies by method and missingness level



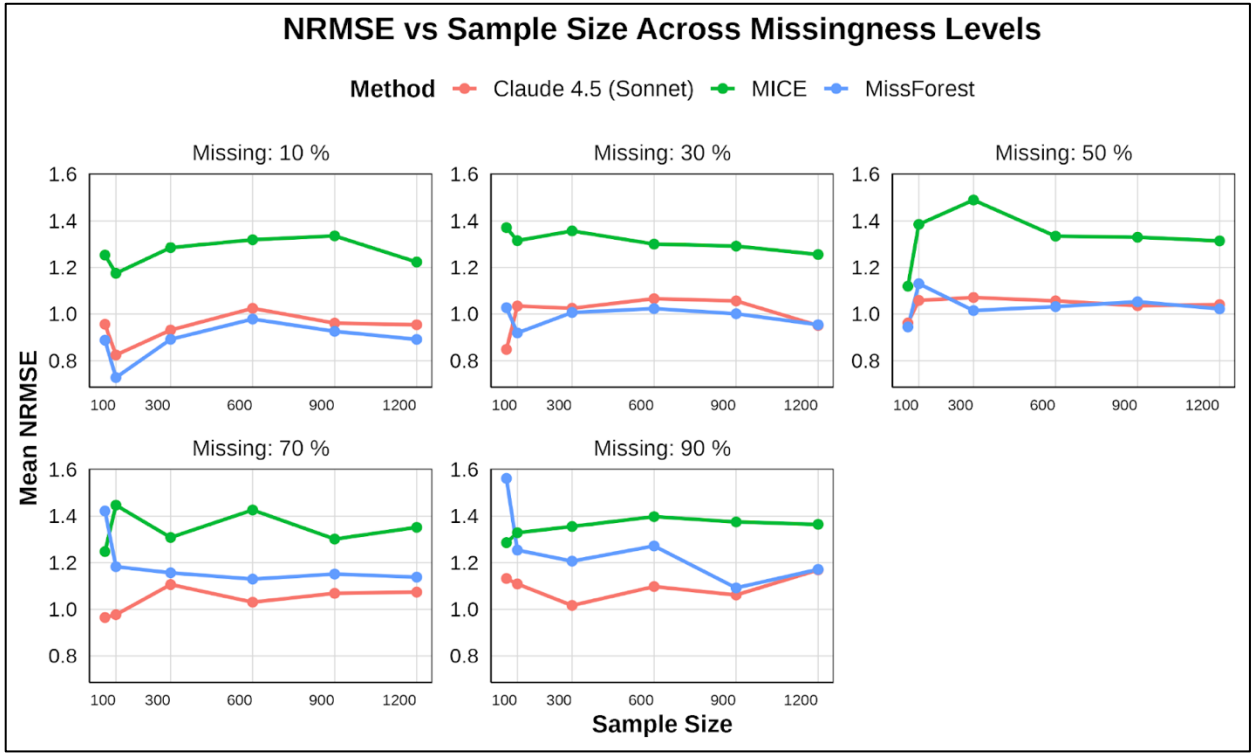


All methods show low median bias, but the spread of bias grows substantially as missingness increases.

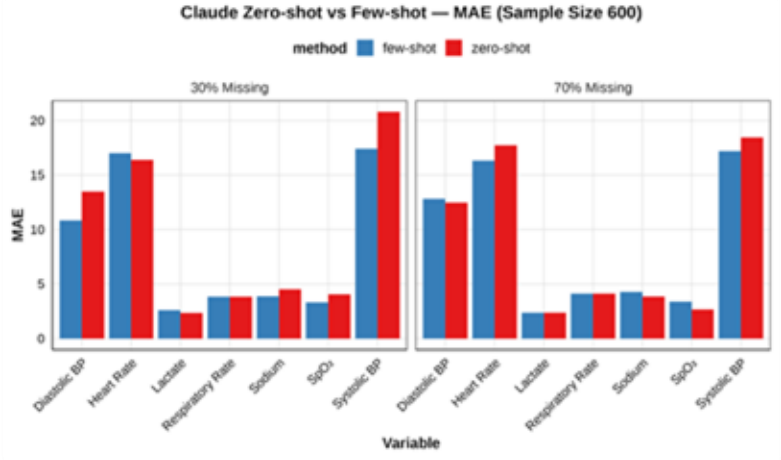


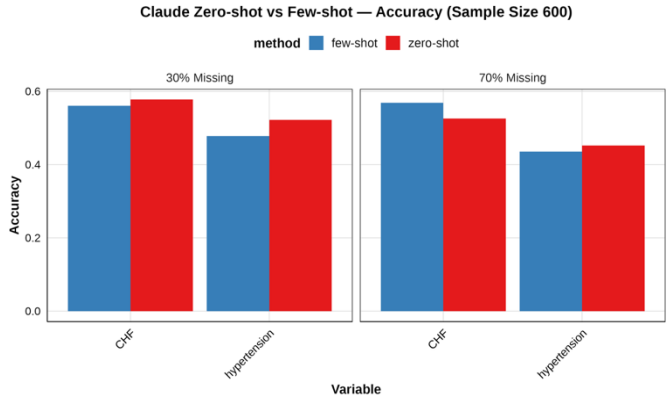


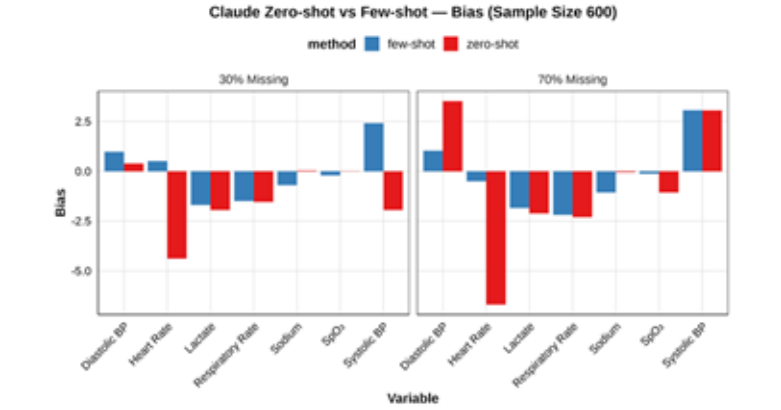
Sample size has minimal impact on NRMSE and MICE consistently produces higher errors than other methods.

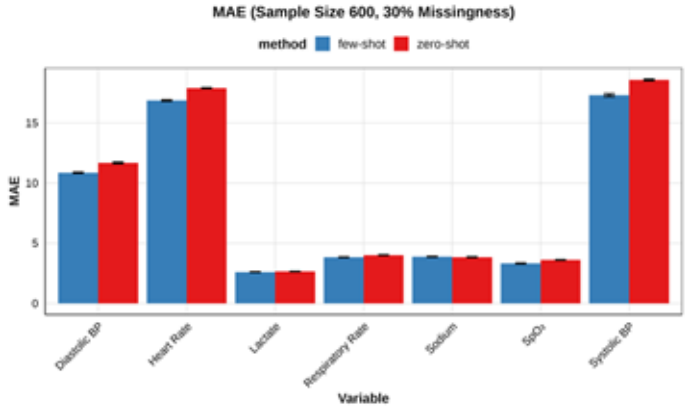


Few-shot and zero-shot prompting yield similar imputation performance across missingness levels







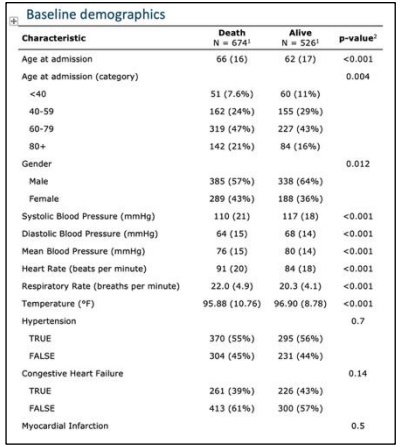


Using LLMs to impute missing clinical data and generate RWE summaries

Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable N
10	8	?	4.4	32	?
15	7	8	4.6	43	45
?	?	5	?	28	68
22	15	?	4.0	?	53
16	?	7.9	3.9	67.5	?
?	6	6.3	?	?	37

LLM

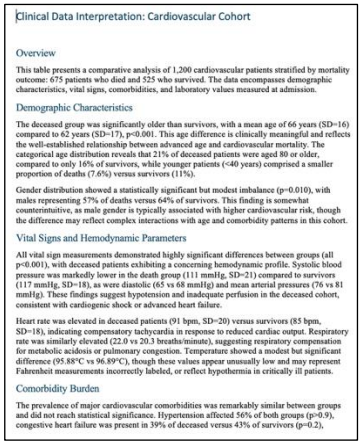
Variable 1	Variable 2	Variable 3	Variable 4	Variable 5	Variable N
10	8	6	4.4	32	?
15	7	8	4.6	43	45
?	10	5	?	28	68
22	15	?	4.0	32	53
16	7	7.9	3.9	67.5	?
?	6	6.3	4.2	45	37



Baseline demographics

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			0.7
TRUE	370 (55%)	295 (56%)	
FALSE	304 (45%)	231 (44%)	
Congestive Heart Failure			0.14
TRUE	261 (39%)	226 (43%)	
FALSE	413 (61%)	300 (57%)	
Myocardial Infarction			0.5

LLM



Clinical Data Interpretation: Cardiovascular Cohort

Overview

This table presents a comparative analysis of 1,200 cardiovascular patients stratified by mortality outcome: 675 patients who died and 525 who survived. The data encompasses demographic characteristics, vital signs, comorbidities, and laboratory values measured at admission.

Demographic Characteristics

The deceased group was significantly older than survivors, with a mean age of 66 years (SD=16) compared to 62 years (SD=17), p<0.001. This age difference is clinically meaningful and reflects the well-established relationship between advanced age and cardiovascular mortality. The categorical age distribution reveals that 21% of deceased patients were aged 80 or older, compared to only 16% of survivors, while younger patients (<40 years) comprised a smaller proportion of deaths (7.6%) versus survivors (11%).

Gender distribution showed a statistically significant but modest imbalance (p=0.010), with males representing 57% of deaths versus 64% of survivors. This finding is somewhat counterintuitive, as male gender is typically associated with higher cardiovascular risk, though the difference may reflect complex interactions with age and comorbidity patterns in this cohort.

Vital Signs and Hemodynamic Parameters

All vital sign measurements demonstrated highly significant differences between groups (all p<0.001), with deceased patients exhibiting a concerning hemodynamic profile. Systolic blood pressure was markedly lower in the death group (111 mmHg, SD=21) compared to survivors (117 mmHg, SD=18), as were diastolic (65 vs 68 mmHg) and mean arterial pressures (76 vs 81 mmHg). These findings suggest hypotension and inadequate perfusion in the deceased cohort, consistent with cardiogenic shock or advanced heart failure.

Heart rate was elevated in deceased patients (91 bpm, SD=20) versus survivors (85 bpm, SD=18), indicating compensatory tachycardia in response to reduced cardiac output. Respiratory rate was similarly elevated (22.0 vs 20.3 breaths/minute), suggesting respiratory compensation for metabolic acidosis or pulmonary congestion. Temperature showed a modest but significant difference (95.88°C vs 96.89°C), though these values appear unusually low and may represent Fahrenheit measurements incorrectly labeled, or reflect hypothermia in critically ill patients.

Comorbidity Burden

The prevalence of major cardiovascular comorbidities was remarkably similar between groups and did not reach statistical significance. Hypertension affected 56% of both groups (p=0.9), congestive heart failure was present in 39% of deceased versus 43% of survivors (p=0.2),

Automated clinical summary generation and evidence contextualization

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			0.7
TRUE	370 (55%)	295 (56%)	
FALSE	304 (45%)	231 (44%)	
Congestive Heart Failure			0.14
TRUE	261 (39%)	226 (43%)	
FALSE	413 (61%)	300 (57%)	
Myocardial Infarction			0.5
TRUE	301 (45%)	246 (47%)	
FALSE	373 (55%)	280 (53%)	
Diabetes Mellitus			0.8
TRUE	278 (41%)	214 (41%)	
FALSE	396 (59%)	312 (59%)	
Chronic Obstructive Pulmonary Disease			0.080
TRUE	159 (24%)	102 (19%)	
FALSE	515 (76%)	424 (81%)	
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

```
You are an expert clinical data analyst providing summary for a clinical domain expert. Interpret the following table for the cohort:
**{cohort_name}**.
The table is provided below in HTML form generated from R's flextable.
Please:

1. Summarize key patient characteristics, lab values, and comorbidities.
2. Highlight distributions, central tendency, and variation for continuous variables.
3. Interpret categorical counts and percentages.
4. Identify notable imbalances or skewed distributions.
5. Describe any clinical relevance or expected vs unexpected patterns by providing literature references.
6. Avoid hallucinating variables not present in the table.
7. Respond in clear paragraphs suitable for a medical/statistics report.

Below is the table (HTML):

------------------------
{table_html}
------------------------

Provide your interpretation now.
```

Metric	Score
Factual accuracy	3
Coverage and completeness	4
Interpretation quality	4
Clarity and communication	4
Clinical appropriateness	3
Absence of hallucinations	4

Clinical Summary: Cardiovascular Cohort

Overview

This analysis compares 674 patients who died with 526 survivors in a cardiovascular cohort. The data reveal significant differences in demographics, vital signs, laboratory values, and physiological parameters between the two groups, with several findings consistent with established prognostic indicators in cardiovascular disease.

Demographic Characteristics

The deceased group was significantly older (mean age 66 ± 16 years) compared to survivors (62 ± 17 years, p<0.001). The age distribution shows a notable pattern: patients aged 80+ comprised 21% of deaths but only 16% of survivors, while younger patients (<40 years) represented 11% of survivors versus 7.6% of deaths. This age-mortality relationship aligns with established cardiovascular literature demonstrating age as an independent predictor of adverse outcomes (Odden et al., *JAMA*, 2012). Gender distribution showed a statistically significant difference (p=0.012), with males representing 64% of survivors but only 57% of deaths—an unexpected finding given that male sex is typically associated with higher cardiovascular mortality risk (Mosca et al., *Circulation*, 2011).

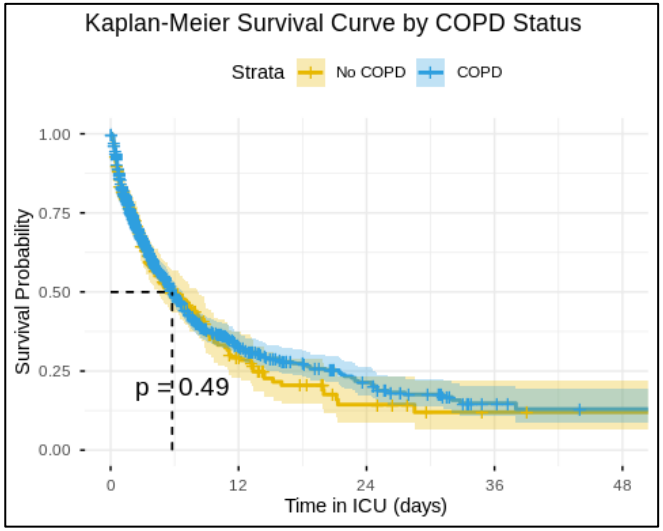
Vital Signs and Hemodynamic Parameters

The deceased group exhibited markedly compromised hemodynamic profiles. Systolic blood pressure was significantly lower in the death group (110 ± 21 mmHg) compared to survivors (117 ± 18 mmHg, p<0.001), as were diastolic (64 vs. 68 mmHg) and mean arterial pressures (76 vs. 80 mmHg). These findings are consistent with hypotension as a marker of cardiogenic shock or severe hemodynamic compromise (Reynolds & Hochman, *JAMA*, 2008). Heart rate was elevated in the deceased group (91 ± 20 vs. 84 ± 18 bpm, p<0.001), suggesting compensatory tachycardia or underlying cardiac dysfunction. Respiratory rate was also significantly higher in non-survivors (22.0 ± 4.9 vs. 20.3 ± 4.1 breaths/minute, p<0.001), indicating respiratory distress or metabolic compensation. Temperature showed a modest but significant difference (95.88°F vs. 96.90°F, p<0.001), though both groups exhibited hypothermia, which may reflect measurement artifacts or critical illness.

Laboratory Values

Several laboratory parameters demonstrated clinically meaningful differences. Lactate levels were dramatically elevated in the death group (4.97 ± 3.78 mmol/L) compared to survivors (2.85 ± 1.97 mmol/L, p<0.001), indicating tissue hypoperfusion and anaerobic metabolism—a well-established predictor of mortality in critically ill cardiovascular patients (Nichol et al., *Critical Care*, 2011). Bicarbonate was significantly lower in non-survivors (19.5 ± 5.3 vs. 22.2 ± 4.6 mmol/L, p<0.001), consistent with metabolic acidosis from poor tissue perfusion. Oxygen saturation was markedly reduced in the death group (95.0 ± 6.9% vs. 97.3 ± 3.2%, p<0.001), with notably greater variability suggesting more severe and heterogeneous respiratory compromise.

Single Clinical Summaries Generated from Multiple RWE Outputs Using LLMs



Metric	Score
Factual accuracy	3
Coverage and completeness	4
Interpretation quality	3
Clarity and communication	3
Clinical Appropriateness	4
Absence of Hallucinations	4

Characteristic	Time 12	Time 24	Time 36	Time 48
COPD				
TRUE	29% (23%, 37%)	14% (8.8%, 23%)	12% (6.5%, 22%)	12% (6.5%, 22%)
FALSE	33% (30%, 38%)	21% (17%, 26%)	15% (11%, 20%)	13% (8.6%, 20%)

Conclusions

- For continuous clinical variables, LLMs maintained accuracy similar to MissForest and exceeded MICE performance as data became sparse.
- Few-shot prompting offered little added benefit over zero-shot, indicating LLM imputation performance here is largely context-driven than example-based learning.
- LLM-generated summaries were fairly accurate and clinically appropriate, free of hallucinations, and capable of contextualizing findings against existing literature.
- LLMs can complement traditional methods by improving both imputation and interpretability of RWD, though rigorous validation remains essential.

Thank you !

Rahul Somavanshi
Navitas Life Sciences, Bangalore
rahulmanohar.s@navitaslifesciences.com

Spencer Langerman
Flatiron Health, NY, USA
spencer.langerman@flatiron.com