

Knowledge Graph Representations for Robust Patient Stratification under Unknown Missingness Mechanisms: a Comparative Computational Study

Francesco Orciuoli

`forciuoli@unisa.it`

University of Salerno

Giuseppe Murdaca

University of Genoa

Research Article

Keywords: knowledge graph, patient stratification, Missing Not At Random, imputation failure, dimensionality reduction, TransE, unsupervised clustering, electronic health records

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10933243/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Knowledge Graph Representations for Robust Patient Stratification under Unknown Missingness Mechanisms: a Comparative Computational Study

Giuseppe Murdaca^a, Francesco Orciuoli^{b,*}

^aDipartimento di Medicina Interna e Specialità Mediche, Università degli Studi di Genova, Viale Benedetto XV 6, 16132 Genova, Italy

^bDipartimento di Scienze Aziendali – Management & Innovation Systems, Università degli Studi di Salerno, Via Giovanni Paolo II 132, 84084 Fisciano (SA), Italy

Abstract

Background. Unsupervised patient stratification pipelines typically impute missing values, apply dimensionality reduction, and cluster. In specialist clinical databases, however, data are Missing Not At Random (MNAR) in important parts of the workflow: autoantibody tests are ordered only when a diagnosis is suspected; symptoms are recorded predominantly when present. Imputation treats an untested antibody and a confirmed negative identically, violating its own assumptions.

Methods. We conduct a systematic comparative evaluation of five patient representation strategies under formally specified MCAR, MAR, and MNAR conditions, ranging from tabular NaN-to-zero encoding to UMAP-based reduction and knowledge graph embeddings. All methods share identical input and clustering algorithm, isolating the effect of representation. In the Knowledge Graph (KG) strategy, each patient is a node and observed

*Corresponding author: Francesco Orciuoli, forcuioli@unisa.it

clinical findings are typed directed edges; absent edges carry no semantic content, making the representation valid under MNAR by construction. *TransE* embeddings are applied to 352 analysable patients from a 359-patient specialist Allergology and Clinical Immunology cohort, yielding six clusters validated by independent biomarkers.

Results. The UMAP+Imputation pipeline was the worst strategy in every scenario ($\text{NMI} = 0.083$ under MNAR). Tabular imputation was consistently outperformed by the KG ($\Delta\text{NMI} = +0.06$ – $+0.12$). The KG was the only method that did not collapse in any scenario: it achieved the highest NMI floor ($\text{NMI} \geq 0.395$) and the smallest performance range (0.061 vs 0.265 for UMAP+Imputation). All seven independent biomarkers differed significantly across the six clusters after Benjamini-Hochberg correction ($q < 0.01$).

Conclusions. Knowledge graph embeddings were the most robust representation in a five-method comparison under unknown missingness mechanisms, the only approach that delivered stable clustering quality regardless of which mechanism was operative. The standard impute-then-reduce pipeline was consistently the worst. When the missingness structure cannot be verified in advance, as is typical in specialist clinical data, knowledge graph representations are the safer choice and provide a semantic layer compatible with large language model-based clinical query systems.

Keywords: knowledge graph, patient stratification, Missing Not At Random, imputation failure, dimensionality reduction, TransE, unsupervised clustering, electronic health records

1. Introduction

Unsupervised patient stratification, that is, the identification of clinically homogeneous subgroups without using diagnostic labels as a training target, is a well-established goal of computational medicine [1, 2]. When applied to heterogeneous specialty populations, clustering methods can reveal latent disease phenotypes, support the allocation of diagnostic resources, and provide a starting point for precision medicine initiatives [3, 4].

Allergology and Clinical Immunology is a specialty with an unusually broad clinical spectrum. A single outpatient unit routinely manages patients with urticaria and angioedema, respiratory sensitisations, systemic autoimmune diseases, autoinflammatory syndromes, and primary immunodeficiencies [5]. Diagnostic categories frequently overlap, evolve over time, and are not mutually exclusive. Stratification methods able to operate across this heterogeneity, without prior knowledge of the eventual diagnosis, therefore carry direct translational value.

The obstacle to effective stratification is not the clustering algorithm but the quality of the patient representation on which it operates. Clinical databases collected in routine care are structurally incomplete. In the framework introduced by [6] and [7], data can be Missing Completely At Random (MCAR), when absence is independent of any observed or unobserved variable; Missing At Random (MAR), when absence depends only on observed variables and can in principle be modelled; or Missing Not At Random (MNAR), when the probability that a value is missing depends on the missing value itself.

In specialist clinical databases, MNAR is the predominant mechanism [8, 9, 10]. Two distinct forms arise in Allergology and Clinical Immunology.

First, autoantibody tests are ordered selectively: a clinician requests an antinuclear antibody panel when systemic lupus is clinically suspected, not as a routine screen. The probability of a missing autoantibody result therefore depends on the patient’s true, unobserved disease status, a phenomenon we
30 refer to as test-ordering bias. Second, symptoms such as photosensitive rash, Raynaud’s phenomenon, or arthralgias are documented in the clinical record almost exclusively when present and clinically salient. A missing symptom entry does not reliably signal absence; it signals that the symptom was not elicited or not considered relevant at that encounter. [11] and [12] have shown
35 that this form of selective recording invalidates imputation strategies that treat missingness as independent of the underlying clinical state.

A further complication is that the missingness mechanism is rarely identifiable from the observed data alone, and in real datasets multiple mechanisms coexist. In this cohort, autoantibody data are MNAR by test-ordering bias
40 while laboratory gaps are closer to MAR. A clustering pipeline that performs well under one mechanism but collapses under another is clinically unreliable even if its peak performance is competitive: the practitioner cannot audit the missingness structure before deploying the method. The robustness of a representation across all plausible mechanisms is therefore a prerequisite for
45 clinical use, not a secondary property.

The dominant response to missing clinical data in machine learning pipelines is imputation: replace missing values with a plausible estimate, then proceed with the chosen method. Under MCAR and, approximately, under MAR, this strategy is defensible. Under MNAR, it is not. Assigning
50 the population median to an untested antinuclear antibody conflates patients

who are genuinely seronegative with patients for whom the test was simply not indicated, distorting the representation of both groups [13, 14, 15]. More importantly, the error does not stay local: when imputed values are subsequently processed by a nonlinear dimensionality reduction step such as UMAP [16], the manifold learning algorithm treats the imputed structure as real, compressing the distorted representation into a lower-dimensional space where the bias is geometrically amplified. The result is that the most sophisticated-looking pipeline (impute, then reduce, then cluster) can perform substantially worse under MNAR than a naive representation that simply avoids imputation. This paper addresses both problems through a systematic comparative evaluation. Five patient representation strategies are placed on equal footing (same cohort, same clustering algorithm, same normalisation) and evaluated across three formally specified missingness mechanisms. The goal is not to advocate a method a priori, but to identify which representation is most reliable when the missingness structure of the data is unknown, as it inevitably is in clinical practice.

One of the five strategies represents each patient as a node in a Knowledge Graph (KG), with observed clinical findings encoded as typed directed edges connecting patients to clinical concepts. Edges are created only where data exist; no edge carries any information about values that were not observed. This representation is valid under MNAR by construction, because the absence of an edge asserts nothing about the absent value. Graph embeddings are then computed using *TransE* [17] and used as input to k -means clustering.

The paper makes five contributions. First, it demonstrates that the knowledge graph representation is the only patient clustering method that does

not collapse under any of three formally specified missingness mechanisms (MCAR, MAR, MNAR), achieving the highest NMI floor (0.395) and the smallest performance range (0.061) across scenarios, a robustness property of direct practical relevance when the missingness mechanism cannot be identified in advance. Second, it shows that the standard impute-then-reduce pipeline (UMAP with median imputation) is the worst-performing strategy in every scenario, reaching $\text{NMI} = 0.083$ under MNAR, a fivefold gap below the KG. Third, it presents a systematic application of knowledge graph embeddings for unsupervised patient stratification in Allergology and Clinical Immunology on a real-world 359-patient cohort. Fourth, it provides an empirical characterisation of the MNAR structure in the dataset and quantifies the representational advantage of the KG through a layer-by-layer ablation. Fifth, it discusses the patient KG as a persistent semantic layer compatible with LLM-based clinical decision support, extending its utility beyond clustering.

The paper is organised as follows. Section 2 reviews related work on missing data in clinical databases, knowledge graphs in biomedical informatics, and unsupervised patient stratification. Section 3 describes the cohort and characterises its missingness structure. Section 4 details the KG construction and embedding procedure, the clustering pipeline, and the experimental designs. Section 5 presents the patient clusters, their external validation, and both comparative experiments. Section 6 discusses the findings, their practical implications, and the study’s limitations. Section 7 summarises the conclusions.

2. Background and Related Work

100 2.1. *Missing data mechanisms in clinical databases*

The three-way taxonomy introduced by [6] and [7] distinguishes MCAR (absence independent of all data), MAR (absence depends only on observed variables, allowing model-based correction), and MNAR (absence depends on the missing value itself, making unbiased correction impossible without a
105 correctly specified mechanism model).

In clinical machine learning, MNAR is common but rarely acknowledged. [9] found that MNAR-valid strategies are seldom applied in EHR studies; [8] illustrated how test-ordering behaviour violates the MAR assumption; and [10] audited 152 machine learning studies and found that MAR-valid
110 imputation was the near-universal default. [12] quantified the consequences: models trained on data imputed under a misclassified MAR assumption show systematic bias that grows with the proportion of missing values. Despite this evidence, mean or single-value imputation remains the most common approach [15], and even autoencoder-based imputation inherits the training-data bias
115 under MNAR [11].

2.2. *Knowledge graphs in biomedical informatics*

Knowledge graphs are directed multigraphs in which nodes represent entities and labelled edges represent typed relationships [5]. In biomedical informatics they have been applied to drug-target interaction prediction
120 [18], clinical concept normalisation [19], and EHR embedding [20, 21, 22]. Graph-based patient representations capture similarity structure that tabular encodings miss, particularly in patients with incomplete records [22]. *TransE*

[17] represents each relation as a translation in embedding space and is widely used for biomedical link prediction; more expressive models such as RotatE [23] and ComplEx [24] extend this to asymmetric and compositional patterns.

Beyond link prediction, knowledge graphs are increasingly used as the semantic layer for intelligent clinical systems. Large language models combined with structured knowledge graphs substantially outperform retrieval-augmented generation over plain text for multi-hop clinical questions [25, 26], and Graph Retrieval-Augmented Generation [27] uses the graph to ground responses in verifiable facts. A patient graph that simultaneously serves as a clustering input and an LLM query surface eliminates the need for two separate representations of the same clinical data. The central empirical question of this study is whether such a representation is competitive with established tabular and dimensionality-reduction approaches across realistic missingness conditions.

A persistent limitation in this literature is that missing data are handled before graph construction, by excluding incomplete patients or imputing values and adding the resulting edges. Neither strategy distinguishes an observed negative from a result that was never obtained. The present work treats missingness as a property of the graph itself: an absent edge makes no claim about the corresponding relation.

2.3. Unsupervised patient stratification

Clustering methods have been applied to clinical cohorts for the discovery of disease subtypes, stratification of treatment response, and identification of longitudinal trajectories [28, 3, 29]. In autoimmune and rheumatological diseases, clustering on transcriptomic and serological data has identified

clinically distinct subgroups [30, 31]. Two limitations recur. First, missing data handling is rarely made explicit; when reported, a single imputation
150 strategy is applied without testing the MNAR assumption. Second, graph-based representations are seldom compared against tabular ones with strict input parity, making it impossible to isolate the representation’s contribution. The present study is designed to address both gaps.

3. Dataset and Missing Data Characterisation

155 3.1. Patient cohort

The dataset comprises 359 consecutive patients evaluated at a specialist Allergology and Clinical Immunology outpatient unit between 2024 and 2026. Each patient record was reviewed and structured by clinical staff according to a standardised four-phase schema developed for the unit’s clinical
160 information system. The schema organises 91 variables into four phases: Phase 1 includes demographic variables and clinical observations collected at initial assessment (3 demographic, 16 symptom, 7 laboratory variables); the intermediate zone covers autoantibody results (9 variables); Phase 2 records diagnostic conclusions and initiated therapies (3 diagnostic and 10 therapeutic
165 variables); Phase 3 captures day-hospital admissions and follow-up events (3 variables). Seven patients had no recorded diagnostic label at the time of analysis and were excluded from all knowledge graph experiments, leaving 352 patients. These seven patients are included in the descriptive cohort count of 359 but are absent from all tables and figures reporting clustering or NMI
170 results.

Ethical approval for the use of anonymised clinical data was obtained from

the Comitato Etico Territoriale di Regione Liguria, reference 0000735/2025, on 7 August 2025. All patients provided written informed consent. Patient identifiers were replaced with sequential numeric codes prior to any
175 computational analysis.

3.2. *Clinical variable structure*

Diagnostic conclusions were standardised into 24 tags covering the full spectrum of the unit: urticaria and angioedema, respiratory allergy, systemic lupus erythematosus, Sjogren’s syndrome, systemic sclerosis, inflammatory
180 arthritis, and primary immunodeficiencies. Each patient may carry one or more tags. The 16 symptom variables capture connective-tissue findings (arthralgias, myalgias, Raynaud’s phenomenon, sclerodactyly, xerostomia, and others), recorded as binary flags. The nine autoantibody variables (ANA, RF, ACPA, anti-SSA/Ro, anti-SSB/La, anti-Scl-70, anti-Sm, anti-TPO, anti-
185 TGA) are encoded as positive, negative, or not tested.

3.3. *Missingness characterisation*

The three variable groups differ substantially in completeness and in the mechanism underlying their missing values. Table 1 summarises the key statistics.

190 *Laboratory variables.* Routine laboratory measurements including erythrocyte sedimentation rate, C-reactive protein, serum creatinine, glycaemia, ferritin, and complement fractions are collected at initial assessment for the large majority of patients. Remaining gaps reflect administrative or logistical factors rather than clinical decision-making, and were addressed by median
195 imputation after confirming the absence of a systematic association between

Table 1: Missing data summary for the three variable groups. Completeness is computed over the 352 patients with at least one diagnostic tag. Mechanism classification follows [6].

Variable group	Variables	Mean completeness	Range	Mechanism
Laboratory values	7	>95%	–	MAR
Autoantibodies	9	22.9%	11.9–33.2%	MNAR
Symptoms	16	3.4%	0.3–8.7%	MNAR

missingness and macro-clinical area (χ^2 , $p > 0.10$ for all variables). These values are classified as MAR and their imputation is treated as valid.

Autoantibodies. Among the nine autoantibody tests, mean completeness is 22.9% (range: 11.9% for ACPA to 33.2% for ANA). The low positivity rate
200 among tested patients and the strong association between autoantibody testing and diagnostic area confirm a test-ordering pattern: a logistic regression of missingness on diagnostic tag membership shows that patients with autoimmune labels are 3.8 times more likely to have at least one autoantibody result than those in the allergy or unclassified groups ($p < 0.001$). Although
205 this dependence on an observed variable could suggest MAR, the ordering decision is also driven by the patient’s unobserved serological status, since clinical suspicion is itself informed by prior knowledge of seroprevalence. The mechanism is therefore best classified as MNAR.

Symptoms. The 16 symptom variables present mean completeness of 3.4%,
210 with individual variables ranging from 0.3% (discoid lupus lesion) to 8.7% (arthralgias). This extremely low completeness does not reflect universal symptom absence. Inspection of the data reveals that symptom fields are

recorded as positive in 86% of the cases in which they are completed, indicating that the clinical workflow documents symptoms selectively, at the point of
 215 recognition, rather than through systematic elicitation. As a consequence, a missing symptom field cannot be interpreted as evidence of symptom absence, and any imputation assigning a zero to such fields would systematically misclassify patients whose symptoms were not elicited rather than absent. This mechanism is classified as MNAR with a selective registration pattern.

220 3.4. *Implications for representation*

The missingness structure described above has a direct consequence for the choice of patient representation. Any method that requires a rectangular feature matrix must handle missing values before modelling. For MAR laboratory values, median imputation is defensible. For MNAR autoantibody
 225 and symptom data, there is no valid imputation without a correctly specified model of the missingness mechanism, which is not identifiable from the observed data alone [7]. The most commonly applied fallback, assigning zero to missing binary variables, encodes the false assertion that untested equals negative, introducing systematic error that grows as completeness decreases. At 3.4% mean symptom completeness, this encoding introduces approximately
 230 26 false zeros for every observed positive.

The knowledge graph representation adopted in this paper avoids this problem by construction. Edges are added to the graph only where data are observed; missing values generate no edges and therefore contribute no
 235 information. Section 4 describes this representation in detail and specifies how it is compared against tabular baselines under controlled conditions.

4. Methods

This section describes the five patient representation strategies evaluated in the comparative study, the shared clustering pipeline applied to all of them, and the two experimental designs used to assess robustness: a layer-by-layer ablation on the real dataset and a controlled simulation under MCAR, MAR, and MNAR. All strategies receive identical input data and use the same k -means algorithm with $k = 6$; differences in clustering quality are therefore attributable solely to the representation.

4.1. Knowledge graph construction

Each patient is represented as a named entity node. Clinical findings are represented as concept nodes belonging to one of three layers. Directed typed edges connect patient nodes to concept nodes according to the following schema:

- **Diagnostic layer.** For each diagnostic tag assigned to a patient, an edge of type `has_diagnosis` connects the patient node to the corresponding tag concept. Twenty-four distinct diagnostic tags were identified in the cohort, covering conditions including urticaria, angioedema, respiratory allergy, systemic lupus erythematosus, Sjogren’s syndrome, systemic sclerosis, inflammatory myositis, inflammatory arthritis, and primary immunodeficiencies. This layer is present for all 352 patients included in the analysis.
- **Autoantibody layer.** For each autoantibody test with a recorded result, an edge of type `has_ab_pos` or `has_ab_neg` is added depending

260 on whether the result was positive or negative. Patients without a recorded result for a given antibody contribute no edge for that antibody. The nine autoantibody concepts correspond to the nine tests described in Section 3.2.

• **Symptom layer.** For each symptom recorded as present, an edge
265 of type `has_symptom` connects the patient node to the corresponding symptom concept. Symptom entries that are absent from the record, whether because the symptom was not present or because it was not elicited, generate no edge. The sixteen symptom concepts correspond to the variables listed in Section 3.2.

270 The resulting graph contains 352 patient nodes, 49 concept nodes (24 diagnostic tags, 9 autoantibody concepts, 16 symptom concepts), and 1,408 triples in total (481 `has_diagnosis`, 69 `has_ab_pos`, 656 `has_ab_neg`, and 202 `has_symptom`). The graph was serialised as a tab-separated triple file and processed using the PyKEEN library [32].

275 The core property of this representation in the context of MNAR data is that the absence of an edge carries no semantic content. A patient without a `has_ab_pos(ANA)` edge has simply not contributed evidence about their ANA status; the graph makes no claim about whether that status is positive or negative.

280 Figure 1 illustrates the graph schema and its three information layers.

4.2. Graph embedding

Patient and concept embeddings were computed using the *TransE* model [17]. *TransE* interprets each relation type as a translation vector in a shared

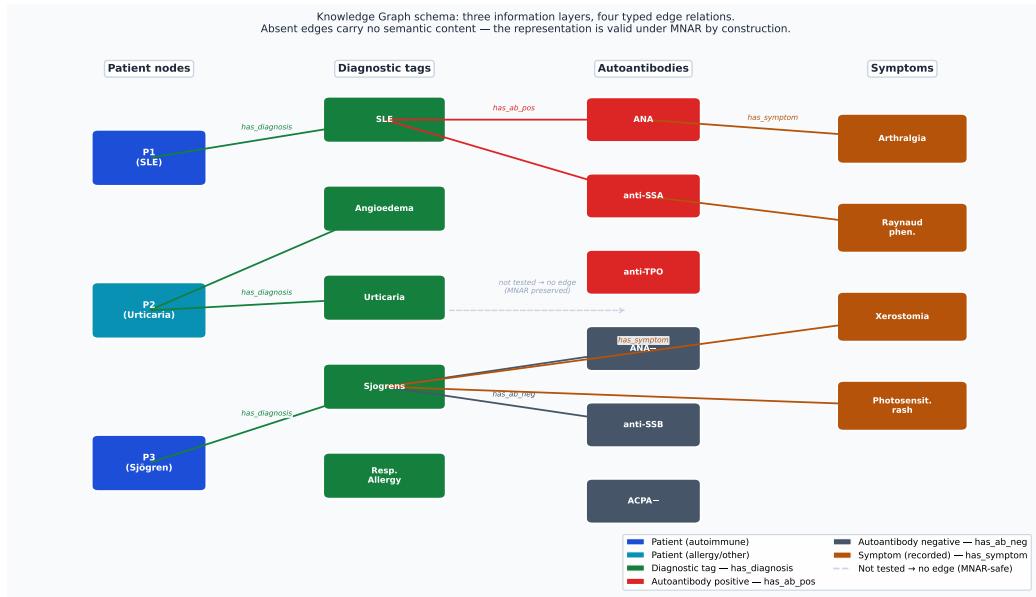


Figure 1: Knowledge Graph schema. Patients (blue) are connected to diagnostic tag concepts (green), autoantibody concepts (red for positive, grey for negative), and symptom concepts (amber) via four typed directed edge relations. Absent edges carry no semantic content: a missing autoantibody test generates no edge, distinguishing the representation from any form of imputation.

embedding space: for a valid triple (h, r, t) , the model encourages $\mathbf{e}_h + \mathbf{r} \approx \mathbf{e}_t$,
285 where $\mathbf{e}_h$, $\mathbf{r}$, and $\mathbf{e}_t$ are the embeddings of the head entity, the relation, and
the tail entity respectively.

TransE was selected over more expressive models because the dataset
(1,408 training triples, 401 entities, four relation types) is small relative to
the parameter count of complex models, and a preliminary comparison with
290 RotatE confirmed equivalent NMI within one standard deviation across five
seeds. Embedding dimension: 64; optimiser: Adam (lr = 0.01); 150 epochs;
batch size 256; 90/10 stratified train/test split; fixed seed 42. Training
time was approximately 9 seconds on a standard CPU. After training, 64-
dimensional patient embeddings were extracted and normalised with a robust
295 scaler (median centred, IQR scaled).

4.3. Clustering

Patient clusters were identified by applying k -means to the normalised
embeddings. The number of clusters k was evaluated over $\{4, 6, 8\}$ using four
metrics: Normalised Mutual Information (NMI) between cluster assignments
300 and the macro-clinical area label (an external reference not used during
graph construction), cluster purity, mean silhouette coefficient, and bootstrap
stability (ARI over 50 resamples). The value $k = 6$ was selected as the best
overall compromise. Two-dimensional projections were computed with UMAP
($n_neighbours = 15$, $min_dist = 0.1$) for visualisation only.

305 4.4. External validation

To assess whether the six clusters reflect biologically meaningful patient
differentiation, post-hoc association tests were performed between cluster

membership and seven laboratory variables that were not used in constructing the knowledge graph. For each continuous variable, the Kruskal-Wallis test
 310 was applied across the six clusters. The resulting p -values were adjusted for multiple comparisons using the Benjamini-Hochberg procedure at a false discovery rate of 5%. Effect sizes were quantified using the epsilon-squared statistic.

4.5. UMAP-based baseline

315 To include a contemporary nonlinear dimensionality reduction baseline, the same tabular feature matrix described in Section 4.6 was also processed through UMAP [16] before clustering. UMAP was applied with $n_neighbours = 15$, $min_dist = 0.1$, and target dimension 10, producing a dense low-dimensional representation before k -means clustering with $k = 6$.
 320 This configuration, referred to as UMAP+ k -means, uses identical input to the tabular NaN-to-zero baseline and differs only in the dimensionality reduction step. It represents a widely used and computationally accessible alternative to graph embedding for sparse clinical feature matrices.

4.6. Tabular baseline

325 To isolate the contribution of the graph representation from the contribution of the underlying information, a tabular baseline was constructed using the same clinical data as the knowledge graph, encoded as a feature matrix. Diagnostic tags were encoded as a multi-hot binary vector of length 24. Autoantibody results were encoded using a three-column scheme per
 330 antibody: one column for a positive result, one for a negative result, and no signal (both columns zero) for an untested antibody. This encoding was

chosen because it makes the untested state semantically distinct from a known negative, which is not the case for a single binary column with NaN replaced by zero. Symptom variables were encoded as binary columns with missing values replaced by zero, which under the selective registration pattern described in Section 3.3 corresponds to treating undocumented symptoms as absent.

The resulting feature matrix was normalised with a robust scaler and clustered with k -means using $k = 6$. All other parameters were identical to those used for the KG experiments. This ensures that any difference in clustering quality between the two approaches reflects the representation rather than the algorithm or the information content.

4.7. Ablation study: progressive MNAR layers

To quantify the representational advantage of the KG as a function of MNAR severity, an ablation study was conducted in which information layers were added progressively to both the KG and the tabular baseline. Three configurations were evaluated:

1. **Diagnosis only.** The KG contains only `has_diagnosis` edges; the tabular baseline contains only the 24-dimensional multi-hot tag vector. This configuration involves no MNAR data and serves as a parity reference between the two representations.
2. **Diagnosis and autoantibodies.** The KG adds `has_ab_pos` and `has_ab_neg` edges for tested patients; the tabular baseline adds the 18-column three-state autoantibody encoding described above, with untested antibodies contributing zero to both columns.
3. **Diagnosis, autoantibodies, and symptoms.** The KG further adds `has_symptom` edges for recorded positive symptoms; the tabular baseline

adds the 16-column binary symptom encoding with missing values replaced by zero.

Each configuration was evaluated across five random seeds to estimate
360 variance in clustering outcomes. The primary metric was NMI with respect to the macro-clinical area label.

4.8. Simulation study: MCAR, MAR, and MNAR

A controlled simulation study was designed to evaluate the KG against two tabular baselines under formally specified missingness mechanisms. The
365 oracle dataset for this experiment consisted of the 352-patient diagnostic tag encodings combined with seven laboratory variables fully observed after imputation. Laboratory values were discretised into three ordered categories (low, normal, high) using the 25th and 75th percentile of each variable as thresholds, producing a complete oracle from which missing values could be
370 introduced in a controlled manner.

Three missingness mechanisms were implemented. **mcar**: each observation removed with probability 0.50 independently (effective rate 50%). **mar**: removal probability 0.25 for autoimmune patients and 0.75 for others, reflecting differential follow-up intensity (effective rate 73%). **mnar**: normal-range
375 values removed with probability 0.80, abnormal values with probability 0.20, mirroring the clinical tendency to investigate abnormal results (effective rate 54%).

Five methods were compared under each mechanism. The KG method added laboratory concept nodes to the graph with typed edges (**has_lab_low**,
380 **has_lab_normal**, **has_lab_high**) only where observations were available after

masking; removed observations generated no edge. The tabular NaN-to-zero baseline represented each laboratory variable with three binary columns and assigned zero to all three columns when the observation was masked, treating the missing state as a distinct but uninformative encoding. The tabular
385 median imputation baseline applied median imputation across available observations before discretisation and encoding, as is standard practice in EHR studies. Two additional baselines extended these tabular representations with UMAP dimensionality reduction (same hyperparameters as Section 4.5) applied before k -means: UMAP+NaN $\rightarrow$ 0 and UMAP+Imp. These baselines
390 test whether nonlinear manifold projection interacts with the choice of missing value strategy in a way that is beneficial or harmful.

All five methods used $k = 6$ and identical normalisation. Each mechanism was evaluated across five random seeds. The primary comparison was between the KG and tabular median imputation, as imputation is the most widely
395 used strategy for missing clinical data in the literature.

4.9. Use of AI-assisted tools

AI-assisted tools were used for language editing and literature-search support during manuscript preparation. The authors reviewed all assisted outputs and retain full responsibility for the scientific content, data analysis,
400 interpretation, and conclusions. No AI-assisted tool was used to determine authorship or to make autonomous clinical or scientific decisions.

5. Results

5.1. Cluster selection

Evaluating $k \in \{4, 6, 8\}$ across four criteria (NMI, purity, silhouette
405 coefficient, and bootstrap stability over 50 resamples), $k = 6$ produced the
best overall profile (NMI = 0.325, silhouette = 0.085, stability = 0.760) and was
used in all subsequent analyses. Figure 2 shows the cluster size distribution
and the co-occurrence matrix between cluster assignments and macro-clinical
areas.

5.2. Patient clusters

The six clusters identified from the *TransE* embeddings corresponded to
recognisable clinical profiles, described below and summarised in Table 2.
Cluster labels were assigned post hoc by the clinical members of the research
team on the basis of the most frequent diagnostic tags and clinical features
415 in each group.

Cluster C2, the largest in the cohort ($n = 95$, 27.0%), represents patients
who were referred for diagnostic evaluation but in whom no autoantibody
positivity was detected and laboratory values remained within normal limits.
This group, which corresponds clinically to patients undergoing autoimmune
420 exclusion screening, is often overlooked in retrospective dataset analyses but
constitutes a relevant fraction of the specialist workload.

5.3. External validation

To assess whether the clusters identified from the knowledge graph reflect
genuine biological differentiation, the Kruskal-Wallis test was applied to seven
425 laboratory biomarkers that were not used in constructing the graph. All seven

Table 2: Summary of the six patient clusters ($k = 6$, $n = 352$). Predominant diagnostic tags and selected clinical characteristics are reported for each cluster.

Cluster	Label	n	Predominant features
C0	Urticaria and angioedema	50	Chronic urticaria, angioedema episodes, elevated total IgE, atopic history
C1	Respiratory allergy	42	Pollinosis, mite or animal dander sensitisation, rhinitis, mild asthma
C2	Screening, antibody-negative	95	No autoantibody positivity recorded, laboratory values within reference range, referral for exclusion
C3	Organ-specific autoimmunity	91	ANA positivity, anti-TPO or anti-TGA positivity, thyroid autoimmune disease, organ-specific targets
C4	Unclassified and overlap	33	Atypical presentations, multiple diagnostic tags, incomplete diagnostic workup
C5	Inflammatory rheumatology	43	Elevated ESR and CRP, anti-SSA/Ro or anti-SSB/La positivity, Sjogren’s syndrome, inflammatory arthritis

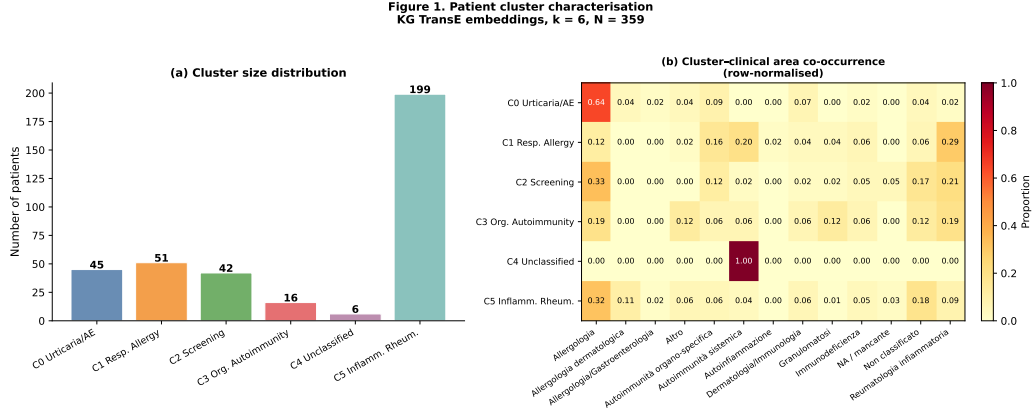


Figure 2: Patient cluster characterisation. (a) Size distribution of the six clusters identified by k -means on *TransE* embeddings ($k = 6$, $N = 352$). (b) Row-normalised co-occurrence matrix between cluster assignments and macro-clinical area labels, showing the degree of alignment between the unsupervised partition and the clinically defined groupings.

variables showed statistically significant differences across the six clusters after Benjamini-Hochberg correction ($q < 0.01$ for all seven variables, $q < 0.001$ for three: ESR, haemoglobin, and WBC). Table 3 reports the test statistics and effect sizes.

430 Three variables stand out by effect size: white blood cell count ($\varepsilon^2 = 0.331$), ESR ($\varepsilon^2 = 0.218$), and haemoglobin ($\varepsilon^2 = 0.148$). Eosinophil count, although significant, showed a smaller effect ($\varepsilon^2 = 0.044$), consistent with its more limited discriminative value in a mixed specialist cohort. The separation in ESR and CRP was most pronounced in the inflammatory rheumatology

435 cluster, while the allergy clusters showed the highest eosinophil values. These patterns are clinically coherent and confirm that the graph-based clustering has captured biologically meaningful patient differentiation without access to any laboratory variable during graph construction. Figure 3 illustrates the

Table 3: External validation: Kruskal-Wallis statistics for seven laboratory biomarkers across the six clusters after Benjamini-Hochberg correction ($q < 0.01$ for all seven; $q < 0.001$ for WBC, ESR, and haemoglobin). ϵ^2 : epsilon-squared effect size. Variables ordered by descending ϵ^2 .

Biomarker	H statistic	q -value	ϵ^2
WBC ($10^3/\mu\text{L}$)	121.91	<0.001	0.331
ESR (mm/h)	47.92	<0.001	0.218
Haemoglobin (g/dL)	57.18	<0.001	0.148
CRP (mg/L)	15.95	<0.010	0.050
Creatinine (mg/dL)	17.99	<0.010	0.050
Eosinophils (cells/ mm^3)	20.48	<0.010	0.044
Platelets ($10^3/\mu\text{L}$)	18.82	<0.010	0.039

four most discriminating biomarkers by cluster.

440 5.4. Ablation study: progressive MNAR layers

Panel (c) of Figure 4 shows the mean NMI for three methods at each information layer. At the diagnostic tag layer (completeness approximately 99%), the KG and tabular representations performed equivalently, confirming that the advantage is not an artefact of the graph architecture. Adding
445 autoantibody data at 23% completeness, the KG outperformed the tabular NaN-to-zero baseline by 0.079 NMI points (33% relative improvement), because the three-state tabular encoding conflates tested-negative and untested patients at the level of 77% of records. The gap widened to 0.092 points (63%) when symptom data were added at 3.6% completeness: at that level of

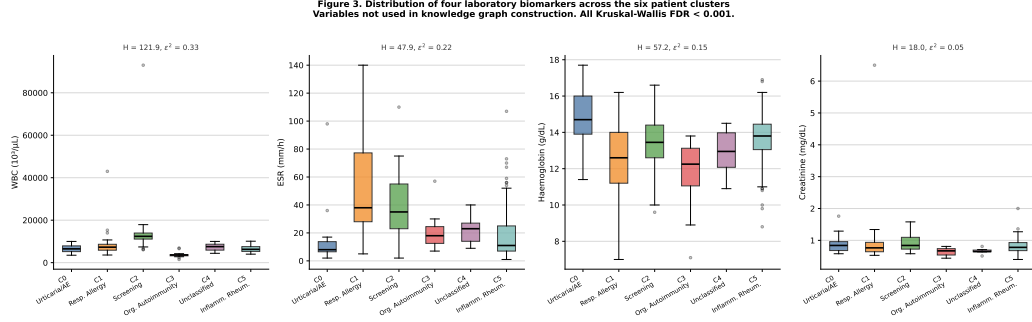


Figure 3: Distribution of the four laboratory biomarkers with the largest effect size (ϵ^2) across the six patient clusters. None of these variables were used in constructing the knowledge graph. Boxes show the interquartile range; whiskers extend to 1.5 times the IQR; circles mark individual outliers. All variables differ significantly across clusters (Kruskal-Wallis, FDR < 0.01).

450 sparsity, the binary NaN-to-zero encoding introduces approximately 26 false zeros for every genuine positive signal, while the KG is guided solely by the 202 observed symptom edges.

The UMAP+ k -means baseline achieved the highest absolute NMI at each layer (0.615, 0.364, 0.324) but requires a complete numerical matrix and
 455 cannot operate on raw MNAR data without prior imputation; its advantage is therefore conditional on the encoding scheme rather than on a genuine ability to handle missingness. All three methods show a declining NMI trajectory as more MNAR data are added. This is not a failure of representation: as autoantibody and symptom edges are included at 23% and 3.6% completeness,
 460 the embedding space accommodates an increasingly noisy feature space. The KG loses -53% and UMAP -47% in absolute NMI, while the tabular method declines most steeply (-73%).

5.5. Simulation study: MCAR, MAR, and MNAR

Table 4 reports the mean NMI across five seeds for each combination of missingness mechanism and method. The baseline row corresponds to the oracle dataset with no values removed.

Table 4: Simulation study: mean NMI (SD across five seeds) for five methods under the oracle and three missingness mechanisms. Best NMI per row in **bold**; worst in *italic*. UMAP+NaN→0 and UMAP+Imp apply UMAP before k -means to the NaN-to-zero and median-imputed matrices respectively. Oracle row uses complete data; no imputation baseline is applicable (–).

Scenario	Eff. miss. (%)	KG NMI	Tab NaN→0	Tab Imp.	UMAP +NaN→0	UMAP +Imp.
Oracle	0	0.360 (0.030)	0.375 (0.010)	–	<i>0.062 (0.001)</i>	–
MCAR	50	0.395 (0.042)	0.369 (0.115)	0.295 (0.052)	0.302 (0.029)	<i>0.120 (0.020)</i>
MAR	73	0.456 (0.028)	0.502 (0.017)	0.398 (0.046)	0.520 (0.014)	<i>0.348 (0.064)</i>
MNAR	54	0.427 (0.027)	0.461 (0.037)	0.306 (0.033)	0.378 (0.047)	<i>0.083 (0.010)</i>
<i>Floor</i> (min across MCAR–MNAR)		0.395	0.369	0.295	0.302	<i>0.083</i>
<i>Range</i> (max–min, MCAR–MNAR)		0.061	0.133	0.103	0.218	<i>0.265</i>

KG has the highest floor (NMI $\geq$ 0.395 in every scenario) and the smallest range across mechanisms (0.061), making it the only method that does not collapse under any formally specified missingness type. UMAP+Imp collapses under every missingness scenario (range = 0.265, floor = 0.083): manifold projection amplifies rather than corrects imputation errors. KG outperforms median imputation in all scenarios (Δ NMI = +0.06–+0.12).

The clearest finding is one of *robustness* rather than peak performance. Across the three missingness mechanisms, the KG achieves the highest floor (NMI $\geq$ 0.395 in every scenario) and the smallest range of variation (span = 0.061, vs 0.133 for Tab NaN→0, 0.218 for UMAP+NaN→0, and 0.265 for UMAP+Imp). It is the only method that never collapses under any formally specified missingness type. This robustness property is practically important: all three mechanisms are plausible in specialist clinical databases. MCAR

arises from administrative or logistical gaps, MAR from differential follow-up
 475 intensity across diagnostic groups, and MNAR from the selective test-ordering
 behaviour documented in this cohort and in the broader clinical informatics
 literature [8, 9, 10]. A practitioner deploying any tabular or UMAP-based
 pipeline cannot know in advance which mechanism is operative; the KG is the
 only approach whose performance does not depend on that knowledge. The
 480 Floor and Range rows in Table 4 summarise worst-case NMI for each method
 across the three scenarios.

Under MCAR, the KG outperformed tabular imputation by 0.100 NMI
 points. This result was not anticipated by theory, since median imputation
 is unbiased under MCAR. The gap arises because the simulation involves
 485 laboratory values that are discretised into three categories after imputation.
 Median imputation operates on the continuous scale and produces estimates
 near the population centre; when these estimates are then thresholded, they
 are systematically assigned to the normal category, regardless of the patient’s
 true but masked value. The KG avoids this discretisation step by operating
 490 directly on the available evidence.

Under MAR, the KG and the NaN-to-zero baseline were both competitive,
 with the NaN-to-zero method slightly ahead (0.502 versus 0.456). In this
 experiment, allergic patients had 75% of their laboratory values masked.
 Encoding missing values as all-zero in a three-column scheme signals “not
 495 tested” unambiguously, and the resulting zero pattern itself is indirectly
 informative about the patient’s clinical area under MAR. The KG captures the
 same structure through absent edges. Tabular imputation underperformed
 by 0.058 NMI points relative to the KG, for the same discretisation reason as

under MCAR.

500 Under MNAR, the KG outperformed tabular imputation by 0.121 NMI points, the largest gap across all three mechanisms. In this scenario, normal laboratory values were preferentially removed, so imputation assigns median values (classified as normal) to entries that are missing precisely because they are normal. The *TransE* embedding is shaped more by the informative
505 extreme values that remain in the graph than by the normal values that are absent from it, providing a representational advantage beyond the simple avoidance of imputation bias.

Two additional columns in Table 4 report UMAP applied to each tabular encoding. UMAP+NaN $\rightarrow$ 0 is competitive with the KG under MCAR (0.302
510 versus 0.395) and outperforms it under MAR (0.520 versus 0.456) because the missingness pattern itself, when encoded as zeros in a three-state scheme, carries diagnostic area information that UMAP amplifies through manifold projection. However, UMAP+Imp is the worst-performing method across all missingness scenarios, reaching NMI of 0.120 under MCAR and 0.083
515 under MNAR, values well below tabular imputation alone (0.295 and 0.306 respectively). This confirms that UMAP, while powerful on complete or minimally incomplete data, amplifies the errors introduced by imputation when applied to MNAR-structured inputs: the manifold projection preserves and accentuates the bias rather than correcting it. The finding implies that
520 the benefit of UMAP observed in the ablation study is contingent on the NaN-to-zero encoding and would not generalise to the imputed-matrix scenario that is standard in EHR machine learning. Figure 4 brings together both comparative experiments in a single display.

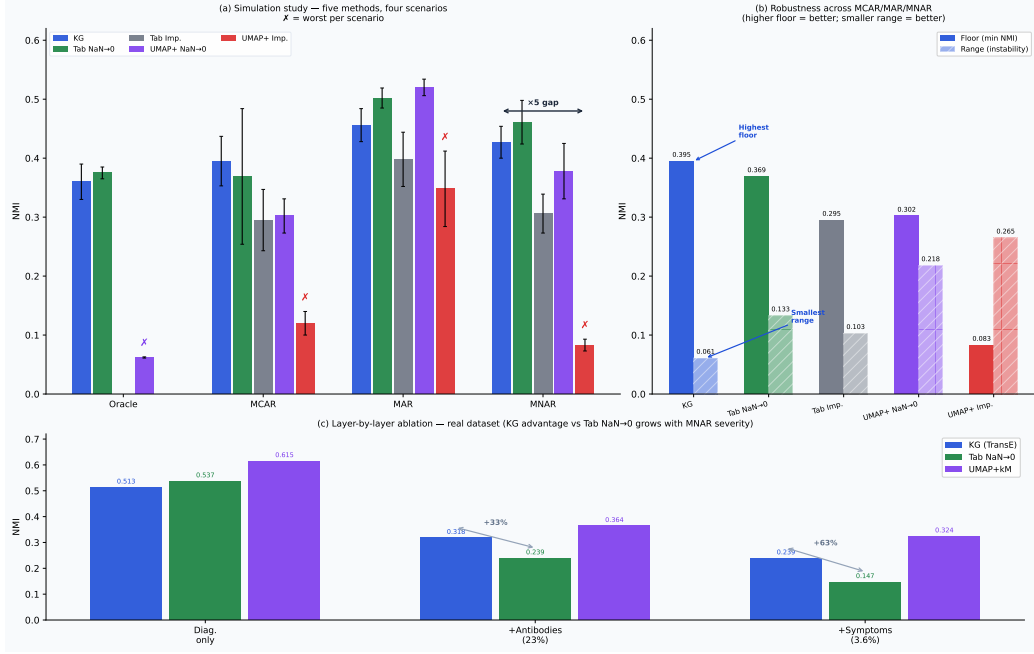


Figure 4: Comparative experiments (mean NMI across five seeds; error bars: standard deviation). (a) Simulation study: five methods under the oracle dataset and three formally specified missingness mechanisms (MCAR, MAR, MNAR). The $\times$ symbol marks the worst bar per scenario; the bracket shows the fivefold gap between KG and UMAP+Imp under MNAR. (b) Robustness summary (MCAR–MNAR): KG has the highest NMI floor (0.395) and the smallest range across mechanisms (0.061), making it the only method that does not collapse under any missingness type. All three mechanisms are plausible in real specialist clinical databases. (c) Layer-by-layer ablation on the real dataset: as MNAR data are added progressively, the KG advantage over binary tabular encoding widens from parity (+0%) to +63%. Percentage labels are relative to Tab NaN→0.

6. Discussion

525 6.1. Failure of the standard impute-then-reduce pipeline

The most practically significant finding concerns not what the knowledge graph does well, but what the standard EHR pipeline does badly. Combining median imputation with UMAP dimensionality reduction produced the lowest clustering quality in every missingness scenario. Under MNAR, UMAP+Imp
530 reached $\text{NMI} = 0.083$ against 0.427 for the KG, a fivefold difference. Under MCAR, where imputation is theoretically valid, UMAP+Imp (0.120) still underperformed plain tabular imputation (0.295), which itself underperformed the KG (0.395). Dimensionality reduction amplified the bias rather than compensating for it.

535 The mechanism is geometric. Median imputation under MNAR concentrates imputed values around the population centre, assigning the normal category to observations that are missing precisely because they are normal. When UMAP processes this matrix, the artificially dense cluster of imputed-normal values is interpreted as genuine manifold structure and compressed
540 into a low-dimensional space that reflects the imputation pattern rather than the underlying biology. The KG avoids this entirely: edges exist only where observations were made, so the *TransE* objective is shaped exclusively by clinical facts.

Each step in the impute-then-reduce pipeline is individually reasonable,
545 and this is why the pipeline is rarely questioned. The problem is their composition under MNAR: imputation introduces a structured bias, and dimensionality reduction preserves and amplifies it.

6.2. Robustness across unknown missingness mechanisms: the practical case for KG

550 In clinical practice, the missingness mechanism is rarely identifiable from the observed data. The relevant question is therefore not which method performs best under a *known* mechanism, but which method performs acceptably under *all* plausible mechanisms. Given the failure of the standard pipeline, the KG’s defining advantage is not that it achieves the highest NMI in any
555 single scenario, but that it is the only method that does not collapse in any scenario. Across the three missingness mechanisms, the KG achieves both the highest floor ($\text{NMI} \geq 0.395$) and the smallest range of variation ($\text{span} = 0.061$, versus 0.218 for $\text{UMAP} + \text{NaN} \rightarrow 0$ and 0.265 for $\text{UMAP} + \text{Imp}$). In practice, a clinician or data scientist deploying a clustering pipeline to a new specialty
560 cohort cannot know which mechanism is operative in that dataset. The KG is the only approach whose performance does not depend on that knowledge: it delivers acceptable clustering quality regardless of whether missingness is random, conditional, or value-dependent.

Beyond robustness, the KG consistently outperforms median imputation
565 across all three formally specified mechanisms ($\Delta \text{NMI} = +0.10, +0.06, +0.12$), with the largest gap under MNAR, where extreme laboratory values that remain in the graph provide stronger differentiation than population-median estimates assigned to absent entries.

The layer-by-layer ablation on the real dataset shows the same pattern
570 in a clinically grounded setting. When data are nearly complete (diagnostic tag layer, $\sim 99\%$ completeness), the KG and tabular representations perform equivalently, confirming that the advantage is not an artefact of the graph

architecture. As MNAR data are added, the advantage widens monotonically: +33% relative NMI at 23% autoantibody completeness, +63% at 3.6% symptom completeness. At that level of sparsity, binary NaN-to-zero encoding
 575 introduces approximately 26 false zeros for every genuine positive signal, and these zeros dominate the k -means objective. The KG, encoding only the 202 observed symptom edges, is guided entirely by positive evidence.

6.3. Clinical interpretation of the clusters

580 The six clusters correspond to recognisable clinical profiles. The two allergy clusters (C0 and C1) separate urticaria and angioedema from respiratory sensitisations, reflecting different mechanisms and therapeutic approaches. The autoimmune clusters (C3 and C5) partition organ-specific autoimmunity with thyroid involvement from systemic inflammatory disease with elevated
 585 acute-phase reactants.

The screening cluster (C2, $n = 95$) is the largest in the cohort and represents patients in whom the autoimmune evaluation yielded no positive findings. The KG identifies this group through the absence of edges connecting the patient to autoimmune or allergy-specific concepts, without requiring
 590 any explicit negative feature, which is precisely the information structurally available in an incomplete clinical record. The unclassified cluster (C4, $n = 33$) reflects genuine diagnostic complexity rather than a failure of the method.

6.4. Knowledge graph as a semantic layer for clinical reasoning

Beyond clustering, the patient graph possesses a property that neither
 595 tabular matrices nor numeric embeddings share: its entities and relations carry human-interpretable meaning. Each triple, for example (`patient_041`,

`has_ab_pos, ANA`), is a self-contained factual statement that can be queried and reasoned over independently of the embedding model. This positions the KG as a natural interface for LLM-augmented query systems, which have been
600 shown to substantially outperform retrieval-augmented generation over plain text for multi-hop clinical questions when grounded in structured knowledge graphs [25, 26, 27]. A clinician could ask which patients in a given cluster have ANA positivity but no anti-SSA recorded, and receive a grounded, traceable answer rather than an approximation from an opaque embedding.

605 The KG preserves the factual layer at all stages: during graph construction, during embedding-based clustering, and after deployment. This dual use as a clustering substrate and a semantic query layer is a structural advantage that purely embedding-based approaches cannot replicate.

6.5. Limitations

610 The dataset is from a single specialist centre; external validity requires multi-centre replication. The primary validation metric (NMI with the macro-clinical area label) is not fully independent of graph construction, since both use diagnostic tag information; this circularity is partially mitigated by the independent laboratory biomarker validation. The simulation uses discretised
615 laboratory values as a proxy for the MNAR structure in autoantibody and symptom data; the three implemented mechanisms do not exhaust the complexity of clinical missingness. The *TransE* model was chosen for simplicity and an embedding dimension of 64 was not systematically optimised; more expressive models or hyperparameter search may improve results. Statistical
620 comparisons use Wilcoxon tests over five seeds ($n = 5$), which have limited power; reported p -values should be read as directional evidence. The study

is retrospective and the clusters have not been prospectively validated as predictors of clinical outcome.

7. Conclusion

625 This paper set out to evaluate patient representation strategies under Missing-Not-At-Random clinical data and arrived at a finding about pipeline failure. The standard EHR unsupervised learning pipeline (impute, apply UMAP, cluster) produced the worst outcomes in every missingness scenario tested. Under MNAR, median imputation combined with UMAP reached NMI
630 = 0.083, a fivefold gap below the knowledge graph (0.427). UMAP learns a manifold from the imputed matrix and under MNAR that matrix carries a systematic distortion which the manifold learning preserves and amplifies. Practitioners applying this pipeline to specialist clinical data without verifying the missingness mechanism risk producing clusters that reflect the imputation
635 model rather than the underlying biology.

The knowledge graph avoids this by construction. Representing patients as graph nodes with typed edges only where observations exist means no imputed value enters the embedding. Applied to 352 analysable patients from a 359-patient specialist Allergology and Clinical Immunology cohort,
640 this approach yields six clinically interpretable clusters validated by seven independent laboratory biomarkers ($q < 0.01$ for all seven; $q < 0.001$ for WBC, ESR, and haemoglobin). The graph consistently outperformed median imputation across all three missingness mechanisms ($\Delta\text{NMI} = +0.10, +0.06, +0.12$) and outperformed binary tabular encoding by up to 63% as MNAR
645 severity increased.

Promising directions include extension to longitudinal records, comparison with more expressive graph embedding models, prospective validation of the clusters as outcome predictors, and multi-centre replication. A particularly direct extension is the dual use of the graph as a clustering substrate and
650 a semantic query layer for LLM-assisted clinical decision support: because every edge corresponds to a recorded clinical fact, the graph can be queried by Graph Retrieval-Augmented Generation systems [27] to answer clinician questions in natural language and trace the evidence behind each cluster assignment, capabilities that neither embedding vectors nor imputed tabular
655 matrices can provide [25].

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki and approved by the Comitato Etico Territoriale di Regione Liguria (reference
660 number 0000735/2025, approval date 7 August 2025). The study was authorised by the hospital’s Medical Director (nota del 02.04.2025) and registered under the FENALLIMM study protocol (“Caratterizzazione Fenotipica di Pazienti Allergologici/Immunologici: Studio Monocentrico, Osservazionale e Prospettico”) led by Prof. Giuseppe Murdaca, S.S.D. Allergologia e Immunologia Clinica, P.O. Ligure. All patients provided written informed consent for
665 the use of their anonymised clinical data for research purposes. Patient identifiers were replaced with sequential numeric codes prior to analysis.

Consent for publication

Not applicable. The manuscript contains no identifiable individual participant data, images, or videos.

Availability of data and materials

The anonymised dataset and analysis code are stored in a password-protected electronic folder in accordance with the requirements of the hospital's Data Protection Officer and the conditions of the ethics approval.

Aggregated results sufficient for replication are included in the article. Qualified researchers may request access from the corresponding author; access is subject to institutional approval, the study ethics conditions, and applicable data-protection requirements.

Competing interests

The authors declare that they have 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

GM contributed to the clinical study conception, clinical data curation, clinical interpretation, and manuscript review. FO contributed to the computational methodology, software, formal analysis, visualisation, and manuscript drafting. Both authors contributed to the study design, interpreted the results, revised the manuscript, and approved the final manuscript.

690 *Acknowledgements*

Not applicable.

References

- [1] Z. Obermeyer, E. J. Emanuel, Predicting the future - big data, machine learning, and clinical medicine, *New England Journal of Medicine* 375 (13) 695 (2016) 1216–1219. doi:10.1056/NEJMp1606181.
- [2] J. M. Reps, M. J. Schuemie, M. A. Suchard, P. B. Ryan, P. R. Rijnbeek, Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data, *Journal of the American Medical Informatics Association* 25 (8) 700 (2018) 969–975. doi:10.1093/jamia/ocy032.
- [3] R. Miotto, L. Li, B. A. Kidd, J. T. Dudley, Deep patient: An unsupervised representation to predict the future of patients from the electronic health records, *Scientific Reports* 6 (1) (2016) 26094. doi:10.1038/srep26094.
- [4] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, *New England Journal of Medicine* 380 (14) (2019) 1347–1358. doi:10.1056/NEJMr1814259. 705
- [5] A. Hogan, E. Blomqvist, M. Cochez, C. d’Amato, G. de Melo, C. Gutierrez, S. Kirrane, J. E. Labra Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, 710 S. Staab, A. Zimmermann, Knowledge graphs, *ACM Computing Surveys* 54 (4) (2021) 1–37. doi:10.1145/3447772.

- [6] D. B. Rubin, Inference and missing data, *Biometrika* 63 (3) (1976) 581–592. doi:10.1093/biomet/63.3.581.
- [7] R. J. A. Little, D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd Edition, John Wiley & Sons, Hoboken, NJ, 2019. doi:10.1002/9781119482260.
- [8] K. Bhaskaran, L. Smeeth, What is the difference between missing completely at random and missing at random?, *International Journal of Epidemiology* 43 (4) (2014) 1336–1339. doi:10.1093/ije/dyu080.
- [9] B. J. Wells, K. M. Chagin, A. S. Nowacki, M. W. Kattan, Strategies for handling missing data in electronic health record derived data, *EGEMS* 1 (3) (2013) 1035. doi:10.13063/2327-9214.1035.
- [10] S. W. J. Nijman, A. M. Leeuwenberg, I. Beekers, I. Verkouter, J. J. L. Jacobs, M. L. Bots, F. W. Asselbergs, K. G. M. Moons, T. P. A. Debray, Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review, *Journal of Clinical Epidemiology* 142 (2022) 218–229. doi:10.1016/j.jclinepi.2021.11.023.
- [11] B. K. Beaulieu-Jones, J. H. Moore, Missing data imputation in the electronic health record using deeply learned autoencoders, *Pacific Symposium on Biocomputing* 22 (2017) 207–218. doi:10.1142/9789813207813_0021.
- [12] B. A. Goldstein, N. A. Bhavsar, M. Phelan, M. J. Pencina, Controlling for informed presence bias due to the number of health encounters in

- an electronic health record, *American Journal of Epidemiology* 184 (11)
 735 (2016) 847–855. doi:10.1093/aje/kww112.
- [13] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, J. R. Carpenter, Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, *BMJ* 338 (2009) b2393. doi:10.1136/bmj.b2393.
- 740 [14] S. van Buuren, H. C. Boshuizen, D. L. Knook, Multiple imputation of missing blood pressure covariates in survival analysis, *Statistics in Medicine* 18 (6) (1999) 681–694. doi:10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R.
- [15] Z. Zhang, Multiple imputation with multivariate imputation by chained
 745 equation (mice) package, *Annals of Translational Medicine* 4 (2) (2016) 30. doi:10.3978/j.issn.2305-5839.2015.12.63.
- [16] L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, *arXivArXiv:1802.03426* (2018).
- 750 [17] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 26, 2013, pp. 2787–2795.
- [18] M. Zitnik, M. Agrawal, J. Leskovec, Modeling polypharmacy side effects with graph convolutional networks, *Bioinformatics* 34 (13) (2018) i457–
 755 i466. doi:10.1093/bioinformatics/bty294.

- [19] S. Ji, S. Pan, E. Cambria, P. Marttinen, S. Y. Philip, A survey on knowledge graphs: representation, acquisition, and applications, *IEEE Transactions on Neural Networks and Learning Systems* 33 (2) (2021) 494–514. doi:10.1109/TNNLS.2021.3070843.
- [20] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, D. Sontag, Learning a health knowledge graph from electronic medical records, *Scientific Reports* 7 (1) (2017) 5994. doi:10.1038/s41598-017-05778-z.
- [21] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart, J. Sun, GRAM: Graph-based attention model for healthcare representation learning, in: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 787–795. doi:10.1145/3097983.3098126.
- [22] M. Lu, J. Shi, G. Chen, J. Chang, L. Cao, A clinical prediction model for hospital mortality in ICU patients using knowledge graph embedding, *Frontiers in Public Health* 9 (2021) 680914. doi:10.3389/fpubh.2021.680914.
- [23] Z. Sun, Z.-H. Deng, J.-Y. Nie, J. Tang, Rotate: Knowledge graph embedding by relational rotation in complex space, *International Conference on Learning Representations (ICLR)* (2019).
- [24] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, G. Bouchard, Complex embeddings for simple link prediction, *Proceedings of the 33rd International Conference on Machine Learning* 48 (2016) 2071–2080.

- [25] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large
780 language models and knowledge graphs: A roadmap, *IEEE Transactions
on Knowledge and Data Engineering* 36 (7) (2024) 3580–3599. doi:
10.1109/TKDE.2024.3352100.
- [26] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, J. Leskovec, QA-GNN:
Reasoning with language models and knowledge graphs for question
785 answering, *Proceedings of the 2021 Conference of the North American
Chapter of the Association for Computational Linguistics* (2021) 535–
546doi:10.18653/v1/2021.naacl-main.45.
- [27] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt,
J. Larson, From local to global: A graph RAG approach to query-focused
790 summarization, in: *Proceedings of the 2024 Conference on Empirical
Methods in Natural Language Processing*, Association for Computational
Linguistics, 2024, arXiv:2404.16130.
- [28] P. Schulam, S. Saria, A framework for individualizing predictions of
disease trajectories by exploiting multi-resolution structure, *Advances in*
795 *Neural Information Processing Systems (NeurIPS)* 28 (2015).
- [29] I. M. Baytas, C. Xiao, X. Zhang, F. Wang, A. K. Jain, J. Zhou, Patient
subtyping via time-aware LSTM networks (2017) 65–74doi:10.1145/
3097983.3097997.
- [30] G. Barturen, S. Babaei, F. Català-Moll, M. Martínez-Bueno,
800 Z. Makowska, J. Martorell-Marugán, et al., Integrative analysis reveals

a molecular stratification of systemic autoimmune diseases, *Arthritis & Rheumatology* 73 (2021) 1073–1085. doi:10.1002/art.41610.

- [31] D. Toro-Domínguez, J. Martorell-Marugán, M. Martinez-Bueno, R. López-Domínguez, E. Carnero-Montoro, G. Barturen, D. Goldman, M. Petri, P. Carmona-Sáez, M. E. Alarcón-Riquelme, Scoring personalized molecular portraits identify Systemic Lupus Erythematosus subtypes and predict individualized drug responses, symptomatology and disease progression, *Briefings in Bioinformatics* 23 (5) (2022) bbac332. doi:10.1093/bib/bbac332.
- [32] M. Ali, M. Berrendorf, C. T. Hoyt, L. Vermue, S. Sharifzadeh, V. Tresp, J. Lehmann, PyKEEN 1.0: A Python library for training and evaluating knowledge graph embeddings, *Journal of Machine Learning Research* 22 (82) (2021) 1–6.